SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 1–16

ASYMPTOTICS FOR SELFDUAL VORTICES ON THE TORUS AND ON THE PLANE: A GLUING TECHNIQUE*

MARTA MACRÌ†, MARGHERITA NOLASCO‡, AND TONIA RICCIARDI†

Abstract. We consider multivortex solutions for the selfdual Abelian Higgs model, as the ratio of the vortex core size to the separation distance between vortex points (the scaling parameter) tends to zero. To this end, we use a gluing technique (a shadowing lemma) for solutions to the corresponding semilinear elliptic equation on the plane, allowing any number (finite or countable) of arbitrarily prescribed singular sources. Our approach is particularly convenient and natural for the study of the asymptotics. In particular, in the physically relevant cases where the vortex points are either finite or periodically arranged in the plane, we prove that a frequently used factorization ansatz for multivortex solutions is rigorously satisfied, up to an error which is exponentially small.

Key words. selfdual Abelian Higgs model, elliptic equation, shadowing lemma

AMS subject classifications. 35J60, 58E15, 81T13

DOI. 10.1137/040619843

1. Introduction. We consider the energy density for the static, two-dimensional selfdual Abelian Higgs model in the following form:

$$\mathcal{E}_\delta(A,\phi) = \delta^2|\mathrm{d}A|^2 + |D\phi|^2 + \frac{1}{4\delta^2}\left(|\phi|^2-1\right)^2,$$

where $A = A_1\mathrm{d}x_1 + A_2\mathrm{d}x_2$, $A_1(x), A_2(x) \in \mathbb{R}$ is a gauge potential (a connection over a principal $U(1)$ bundle), ϕ, $\phi(x) \in \mathbb{C}$ is a Higgs matter field (a section over an associated complex line bundle), $D = \mathrm{d} - iA$ is the covariant derivative, and $\delta > 0$ is the scaling parameter. $\mathcal{E}_\delta$ is a rescaling of $\mathcal{E}_1 = \mathcal{E}_\delta|_{\delta=1}$, which coincides with the two-dimensional Ginzburg–Landau energy density in the so-called Bogomol'nyi limit. Such a limit describes the borderline between type I and type II superconductors; see, e.g., Jaffe and Taubes [12]. By the selfdual structure, solutions to the Euler–Lagrange equations of $\mathcal{E}_\delta$ may be obtained from solutions to the first order system:

$$(D_1 \pm iD_2)\phi = 0, \tag{1.1}$$

$$F_{12} = \partial_1 A_2 - \partial_2 A_1 = \pm\frac{1}{2\delta^2}(1-|\phi|^2). \tag{1.2}$$

The vortex-type critical points for the energy associated with $\mathcal{E}_\delta$, namely, the solutions of (1.1)–(1.2), have received considerable attention in recent years, in view of both their physical and geometrical interest; see, e.g., García-Prada [8], Hong, Jost, and Struwe [10], Stuart [17], Taubes [19], Wang and Yang [20], and the references therein. In particular, Hong, Jost, and Struwe [10] consider (1.1)–(1.2) on a compact

*Received by the editors November 30, 2004; accepted for publication (in revised form) April 13, 2005; published electronically August 17, 2005. This research was supported in part by the M.I.U.R. Project "Metodi Variazionali ed Equazioni Differenziali Non Lineari."
http://www.siam.org/journals/sima/37-1/61984.html

†Dipartimento di Matematica e Applicazioni, Università di Napoli Federico II, Via Cintia, 80126 Napoli, Italy (macri@unina.it, tonia.ricciardi@unina.it). The last author's research was supported in part by Regione Campania L.R. 5/02.

‡Dipartimento di Matematica Pura ed Applicata, Università di L'Aquila, Via Vetoio, Coppito, 67010 L'Aquila, Italy (nolasco@univaq.it).

Riemannian surface and perform a detailed analysis of the asymptotics as $\delta \to 0^+$. Indeed, the small $\delta > 0$ regime, corresponding to the limit of small vortex core size with respect to the separation distance between vortices, is an appropriate approximation for the analysis near the vortex points of solutions of (1.1)–(1.2) with $\delta = 1$. This type of asymptotics is also relevant in the context of Ginzburg–Landau vortices; see, e.g., Aftalion, Sandier, and Serfaty [2], Alama and Bronsard [3], André, Bauman, and Phillips [4], Bethuel, Brezis, and Hélein [6], Lin [13], Rubinstein and Sternberg [16], just to mention a few. It has also been widely investigated in the context of other selfdual gauge theories; see the monographs of Tarantello [18] and Yang [22].

The fundamental results concerning finite energy solutions of (1.1)–(1.2) on $\mathbb{R}^2$ were obtained by Taubes [12, 19]. In particular, Taubes showed that such solutions are completely determined by solutions to the singular elliptic problem

$$\begin{cases} -\Delta u = \delta^{-2}(1 - \mathrm{e}^u) - 4\pi \sum_{j=1}^s m_j \delta_{p_j} & \text{on } \mathbb{R}^2, \\ u(x) \to 0 & \text{as } |x| \to +\infty. \end{cases} \tag{1.3}$$

Here $s \in \mathbb{N}$, and for $j = 1, 2, \dots, s$, $p_j \in \mathbb{R}^2$ are the vortex points, $m_j \in \mathbb{N}$ is the multiplicity of p_j, and δ_{p_j} is the Dirac measure at p_j. By variational methods, Taubes proved that, for any $\delta > 0$, there exists a unique solution of (1.3) such that the configuration (A, ϕ) defined in complex notation by

$$\begin{cases} \phi(z) = \exp\{\frac{1}{2}u(z) \pm i \sum_{j=1}^s m_j \arg(z - p_j)\}, \\ A_1 \mp iA_2 = -i(\partial_1 \pm i\partial_2) \ln \phi \end{cases} \tag{1.4}$$

is a smooth, finite energy solution of (1.1)–(1.2) on $\mathbb{R}^2$, satisfying $\phi(p_j) = 0$ (with the corresponding multiplicity $m_j \in \mathbb{N}$) and $E = \int_{\mathbb{R}^2} \mathcal{E}_\delta(A, \phi) = \pm \int_{\mathbb{R}^2} F_{12} = 2\pi \sum_{j=1}^s m_j = 2\pi N$, where $F_{12} = \partial_1 A_2 - \partial_2 A_1$ is the magnetic field (the curvature of A).

In connection with Abrikosov's mixed states in superconductivity [1], it is also of physical interest to analyze (1.1)–(1.2) on the flat torus $\mathbb{T}^2 \equiv \mathbb{R}^2/(a\mathbb{Z} \times b\mathbb{Z})$, where $a, b > 0$. Such a case has been considered by Wang and Yang in [20]. It is shown in [20] that solutions of (1.1)–(1.2) with $\delta = 1$ exist for any given set of vortex points $p_j \in \mathbb{T}^2$, $j = 1, 2, \dots, s$, with multiplicity $m_j \in \mathbb{N}$, if and only if $N = \sum_{j=1}^s m_j < |\mathbb{T}^2|/(4\pi)$. In particular, on $\mathbb{T}^2$ the total number of vortices N cannot be arbitrarily large. Similarly as on $\mathbb{R}^2$, denoting $\Omega = (0, a) \times (0, b)$, solutions of (1.1)–(1.2) on $\mathbb{T}^2$ correspond to solutions for the singular elliptic problem

$$\begin{cases} -\Delta u = \delta^{-2}(1 - \mathrm{e}^u) - 4\pi \sum_{j=1}^s m_j \delta_{p_j} & \text{in } \Omega, \\ u \text{ doubly periodic} & \text{on } \partial\Omega. \end{cases} \tag{1.5}$$

The periodic boundary conditions are justified by certain more general gauge invariant conditions on the configuration (A, ϕ) introduced by 't Hooft [11]. Such conditions force the magnetic flux through a lattice cell to be a "quantized" value proportional to the number of vortices confined. Namely, the 't Hooft boundary conditions imply the topological constraint $\pm\int_\Omega F_{12} = 2\pi \sum_{j=1}^s m_j = 2\pi N$ on the solutions of (1.5), exactly as for finite energy solutions on $\mathbb{R}^2$. Integrating (1.5) on the periodic cell Ω, we obtain that a necessary condition to the solvability of (1.5) is $\delta^2 < |\Omega|/(4\pi N)$. This is obviously satisfied for any finite vortex number N provided $\delta > 0$ is sufficiently small.

Our aim in this note is to show that a "shadowing-type lemma" as introduced in the context of elliptic PDEs by Angenent [5] (see also Nolasco [15]) may be adapted to elliptic equations with singular sources in order to construct solutions for the following more general equation:

$$-\Delta u = \delta^{-2}(1 - \mathrm{e}^u) - 4\pi \sum_{j\in\mathcal{P}} m_j \delta_{p_j} \qquad \text{in } \mathbb{R}^2, \tag{1.6}$$

where the set of indices $\mathcal{P}$ may be either finite or countable, and the vortex points p_j, $j \in \mathbb{N}$, are *arbitrarily* distributed in the plane with the only constraint that

$$d := \inf_{k\neq j} |p_j - p_k| > 0 \qquad \text{and} \qquad m := \sup_{j\in\mathcal{P}} m_j < +\infty. \tag{1.7}$$

The solution we obtain for (1.6) coincides with the solution obtained by Taubes for problem (1.3) when $\mathcal{P}$ is finite and with the solution obtained by Wang and Yang for problem (1.5) when $\mathcal{P}$ is infinite and the vortex points are periodically arranged in $\mathbb{R}^2$. In fact, unlike the previous approaches, our method provides a unified analysis of (1.3) and (1.5). It should be mentioned that suitable modifications to the method described in [5] are necessary due to the singular sources appearing in (1.6). The case where $\mathcal{P}$ is countable and the vortex points are arbitrarily arranged in $\mathbb{R}^2$ does not seem to have been considered before. Of course, if $\mathcal{P}$ is countable, the energy of such a solution is infinite and only locally bounded. On the other hand, our "gluing" technique is, particularly, convenient and natural to analyze the asymptotics as $\delta \to 0^+$. In particular, as a by-product of our construction, we derive a rigorous proof of the following *approximate product formula*:

$$\phi(x) = \prod_{j\in\mathcal{P}} \Phi_{m_j}\left(\frac{x - p_j}{\delta}\right) + \eta_\delta, \tag{1.8}$$

where $\|\eta_\delta\|_{L^\infty(\mathbb{R}^2)} \le C\mathrm{e}^{-c/\delta}$ with $C, c > 0$ independent of δ. Here (A_{m_j}, Φ_{m_j}) is the unique, up to gauge transformation, single vortex (or antivortex) solution with multiplicity m_j to (1.1)–(1.2) with $\delta = 1$ on $\mathbb{R}^2$. We note that in the small $\delta > 0$ regime, a product formula of the form (1.8) is a widely used ansatz in the physics literature, in particular in the study of the dynamics of vortices in the Ginzburg–Landau model; see, e.g., E [7], Neu [14], and Weinstein and Xin [21]. However, we have found a rigorous proof of (1.8) only for the case $N = 2$ on $\mathbb{R}^2$ in Stuart [17]. The asymptotic behavior of solutions of (1.1)–(1.2) as $\delta \to 0^+$ is readily derived from formula (1.8) as well as the convergence rates. In fact, in the case of $\mathbb{T}^2$, our asymptotic description improves the previous result obtained (for general compact Riemann surfaces) by Hong, Jost, and Struwe [10] (see Corollary 2.1).

Although we have chosen to consider the Abelian Higgs model for the sake of simplicity, we will show in a forthcoming note that our method may be adapted to other selfdual gauge theories as considered, e.g., in the monographs [18, 22].

2. Main results and outline of the proof. In order to state precisely our results, we denote by U_N the unique radial solution for the problem (see [12])

$$\begin{cases} -\Delta U_N = 1 - \mathrm{e}^{U_N} - 4\pi N \delta_0 & \text{in } \mathbb{R}^2, \\ U_N(x) \to 0 & \text{as } |x| \to +\infty. \end{cases} \tag{2.1}$$

Our main result is the following theorem.

THEOREM 2.1. *Let $p_j \in \mathbb{R}^2$, $m_j \in \mathbb{N}$, $j \in \mathcal{P} \subseteq \mathbb{N}$, and assume that conditions* (1.7) *hold. Then there exists a constant $\delta_1 > 0$ (depending on* d *and* m *only) such* *that for every $\delta \in (0, \delta_1)$ there exists a solution u_δ for* (1.6)*. Furthermore, u_δ satisfies the approximate superposition rule*

$$u_\delta(x) = \sum_{j \in \mathcal{P}} U_{m_j}\left(\frac{|x - p_j|}{\delta}\right) + \omega_\delta, \tag{2.2}$$

where the error term ω_δ satisfies $\|\omega_\delta\|_\infty \le C\mathrm{e}^{-c/\delta}$ for some $C, c > 0$ independent of δ. In particular, u_δ satisfies the following properties:

(i) $0 \le \mathrm{e}^{u_\delta} < 1$, *$\mathrm{e}^{u_\delta}$ vanishes exactly at p_j, $j \in \mathcal{P}$;*

(ii) *for every compact subset K of $\mathbb{R}^2 \setminus \cup_{j\in\mathcal{P}}\{p_j\}$, there exist $C, c > 0$ such that $\sup_K(1 - \mathrm{e}^{u_\delta}) \le C\mathrm{e}^{-c/\delta}$ as $\delta \to 0^+$;*

(iii) $\pm F_{12} = \frac{1}{2\delta^2}(1 - \mathrm{e}^{u_\delta}) \to 2\pi \sum_{j\in\mathcal{P}} m_j \delta_{p_j}$ *in the sense of distributions as $\delta \to 0^+$.*

In the case that $\mathcal{P}$ is countable, we say that the vortex points p_j, $j \in \mathcal{P}$, are doubly periodically arranged in $\mathbb{R}^2$ if there exists $s \in \mathbb{N}$ such that

$$\{p_k\}_{k\in\mathcal{P}} = \{p_j + m\underline{e}_1 + n\underline{e}_2 \ : \ j = 1, \dots, s;\ m, n \in \mathbb{Z}\}, \tag{2.3}$$

where $\underline{e}_1$ and $\underline{e}_2$ are the unit vectors in $\mathbb{R}^2$ defining the periodic cell domain Ω (for simplicity, we assume $a = b = 1$). Under this condition, solving (1.6) is equivalent to solving (1.5). Namely, we deal with the physically relevant case of a finite number of vortex points $p_1, \dots, p_s \in \Omega$, with the corresponding multiplicity m_j, $j = 1, \dots, s$, such that $\sum_{j=1}^s m_j = N$, where N is the vortex number and Ω is the periodic cell domain. As a consequence of Theorem 2.1, and proving in addition that if (2.3) is satisfied, then the solution u_δ for (1.6) is in fact doubly periodic with periodic cell domain Ω, we derive the following result.

Corollary 2.1. *If the p_j's are doubly periodically arranged in $\mathbb{R}^2$, there exists a constant $\delta_1 > 0$ (depending on N only) such that for every $\delta \in (0, \delta_1)$ the solution u_δ, given in Theorem* 2.1*, is a solution for* (1.5)*. Furthermore, the corresponding vortex configurations (A_δ, ϕ_δ) satisfy the approximate factorization rule*

$$\phi_\delta(x) = \prod_{j=1}^{s} \Phi_{m_j}\left(\frac{x - p_j}{\delta}\right) + \eta_\delta, \qquad x \in \Omega,$$

where the error term η_δ satisfies $\|\eta_\delta\|_\infty \le C\mathrm{e}^{-c/\delta}$ for some $C, c > 0$ independent of δ, and (A_{m_j}, Φ_{m_j}) is the unique, up to gauge transformation, single vortex (or antivortex) solution with multiplicity m_j, to (1.1)–(1.2) *with $\delta = 1$ on $\mathbb{R}^2$. In particular, we have the following:*

(i) $0 \le |\phi_\delta|^2 < 1$, *$\phi_\delta$ vanishes exactly at p_j, $j = 1, \dots, s$;*

(ii) *for every compact subset K of $\Omega \setminus \{p_1, \dots, p_s\}$, there exist $C, c > 0$ such that $0 \le \sup_K(1 - |\phi_\delta|^2) \le C\mathrm{e}^{-c/\delta}$ as $\delta \to 0^+$;*

(iii) $\pm F_{12}(A_\delta, \phi_\delta) = \frac{1}{2\delta^2}(1 - |\phi_\delta|^2) \to 2\pi \sum_{j=1}^s m_j \delta_{p_j}$ *in the sense of distributions (on Ω) as $\delta \to 0^+$;*

(iv) $\int_\Omega \mathcal{E}_\delta(A_\delta, \phi_\delta) = \pm \int_\Omega F_{12}(A_\delta, \phi_\delta) = 2\pi N$.

An outline of the proof is as follows. Our starting point in proving Theorem 2.1 is to consider $\delta > 0$ as a scaling parameter. Setting $\hat{u}(x) = u(\delta x)$, we have that $\hat{u}$ satisfies

$$-\Delta \hat{u} = 1 - \mathrm{e}^{\hat{u}} - 4\pi \sum_{j\in\mathcal{P}} m_j \delta_{\hat{p}_j} \qquad \text{in } \mathbb{R}^2, \tag{2.4}$$

where $\hat{p}_j = p_j/\delta$. Note that the vortex points $\hat{p}_j$ "separate" as $\delta \to 0^+$. Section 3 contains some properties of the radial solutions U_N to (2.1). We rely on the results of Taubes [19] for the existence and uniqueness of U_N as well as for the exponential decay properties at infinity. We also prove a nondegeneracy property of U_N. The exponential decay of solutions justifies the following approximate superposition picture for small values of δ, i.e., for vortex points $\hat{p}_j$ which are "far apart":

$$\hat{u}(x) \approx \sum_{j\in\mathcal{P}} U_{m_j}\left(|x - \hat{p}_j|\right).$$

In fact, we take the following preliminary form of the superposition rule:

$$\hat{u} = \sum_{j\in\mathcal{P}} \hat{\varphi}_j U_{m_j}(x - \hat{p}_j) + z \tag{2.5}$$

as an *ansatz* for $\hat{u}_\delta$. Here, radial solutions centered at $\hat{p}_j$ are "glued" together by the functions $\hat{\varphi}_j$, which belong to a suitable locally finite partition of unity. Section 4 contains the definition and the main properties of the partition as well as of the appropriate functional spaces $\hat{X}_\delta, \hat{Y}_\delta$, which are also obtained by "gluing" $H^2(\mathbb{R}^2)$ and $L^2(\mathbb{R}^2)$, respectively. Hence, we are reduced to show that for small values of $\delta > 0$, there exists an exponentially small "error" z such that $\hat{u}$ defined by (2.5) is a solution of (2.4). The existence of such a $z \in \hat{X}_\delta$ is the aim of section 5 (see Proposition 5.1). To this end, we use the shadowing lemma. We characterize z by the property $F_\delta(z) = 0$, where $F_\delta : \hat{X}_\delta \to \hat{Y}_\delta$ is suitably defined. The nondegeneracy property of U_N is essential in order to prove that the operator $DF_\delta(0)$ is invertible, and that its inverse is bounded independently of $\delta > 0$ (Lemma 5.3). At this point, the Banach fixed point argument applied to $\mathbb{I} - (DF_\delta(0))^{-1} F_\delta$ yields the existence of the desired error term z. In section 6 we show that (2.5) implies (2.2) and we derive the asymptotic behavior of solutions, thus concluding the proof of Theorem 2.1. Finally, we derive Corollary 2.1 by showing that periodically arranged vortex points lead to periodic solutions.

Henceforth, unless otherwise stated, we denote by $C, c > 0$ general constants independent of $\delta > 0$ and of $j \in \mathcal{P}$.

3. Single vortex point solutions. In this section, we consider the solution U_N to the radially symmetric equation (2.1). For every $r > 0$, we denote $B_r = \{x \in \mathbb{R}^2 : |x| < r\}$. The following lemma contains some properties of U_N that will be needed in the following. The proof is a consequence of the results of Taubes [12, 19] on the existence, uniqueness, and the exponential decay of U_N together with standard elliptic theory as in, e.g., [9]. Therefore, it is omitted.

LEMMA 3.1. *The following properties hold:*

(i) $e^{U_N(x)} < 1$ *for any* $x \in \mathbb{R}^2$*;*

(ii) *for every* $r > 0$ *there exist constants* $C_N > 0$ *and* $\alpha_N > 0$*, depending on* r *and* N *only, such that*

$$|1 - e^{U_N(x)}| + |\nabla U_N(x)| + |U_N(x)| \le C_N e^{-\alpha_N |x|}$$

for all $x \in \mathbb{R}^2 \setminus B_r$.

We consider the bounded linear operator

$$L_N = -\Delta + e^{U_N} : H^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2).$$

In order to apply the shadowing lemma, we also need the following nondegeneracy property of U_N.

LEMMA 3.2. *The operator L_N is invertible and for every $N > 0$ there exists $C_N > 0$ such that $\|L_N^{-1}\| \le C_N$.*

Proof. It is readily seen that L_N is injective. Indeed, suppose $L_N u = 0$ for some $u \in H^2(\mathbb{R}^2)$. Multiplying by u and integrating on $\mathbb{R}^2$, we have

$$\int |\nabla u|^2 + \int \mathrm{e}^{U_N} u^2 = 0.$$

Therefore, $u = 0$. Now, we claim that L_N is a Fredholm operator. Indeed, we write

$$L_N = (-\Delta + 1)(\mathbb{I} - T)$$

with $T = (-\Delta + 1)^{-1}(1 - \mathrm{e}^{U_N}) : H^2(\mathbb{R}^2) \to H^2(\mathbb{R}^2)$. Clearly, T is continuous. Let us check that T is compact. To this end, let $u_n \in H^2(\mathbb{R}^2)$, $\|u_n\|_{H^2} = 1$. We have to show that $T u_n$ has a convergent subsequence. Note that by the Sobolev embedding

$$\|u\|_{L^\infty(\mathbb{R}^2)} \le C_S \|u\|_{H^2(\mathbb{R}^2)}, \tag{3.1}$$

for all $u \in H^2(\mathbb{R}^2)$, we have $\|u_n\|_\infty \le C'$, for some $C' > 0$ independent of n, and there exists u_∞, $\|u_\infty\|_{H^2} \le 1$, such that $u_{n_k} \to u_\infty$ strongly in L^2_{loc} for a subsequence u_{n_k}. Now, by Lemma 3.1, for any fixed $\varepsilon > 0$, there exists $R > 0$ such that $\|1 - \mathrm{e}^{U_N}\|_{L^2(\mathbb{R}^2 \setminus B_R)} \le \varepsilon$. Consequently, $\|(1 - \mathrm{e}^{U_N})(u_{n_k} - u_\infty)\|_{L^2(\mathbb{R}^2 \setminus B_R)} \le 2C'\varepsilon$. On the other hand, $\|(1 - \mathrm{e}^{U_N})(u_{n_k} - u_\infty)\|_{L^2(B_R)} \to 0$. We conclude that $(1 - \mathrm{e}^{U_N})(u_{n_k} - u_\infty) \to 0$ in L^2. In turn, we have $T(u_{n_k} - u_\infty) = (-\Delta + 1)^{-1}(1 - \mathrm{e}^{U_N})(u_{n_k} - u_\infty) \to 0$ in H^2, which implies that T is compact. It follows that L_N is a Fredholm operator. Consequently, L_N is also surjective. At this point, the open mapping theorem concludes the proof. □

4. A partition of unity. In this section, we introduce a partition of unity and we prove some technical results which will be needed in the following. Let $p_j \in \mathbb{R}^2$ ($j \in \mathcal{P} \subseteq \mathbb{N}$) be the vortex points. By assumption (1.7), $r_0 = d/8 = \inf_{j \ne k} |p_j - p_k|/8 > 0$. We consider the set $K = (-\frac{3}{4}r_0, \frac{3}{4}r_0) \times (-\frac{3}{4}r_0, \frac{3}{4}r_0)$. Then, for any $\underline{n} \in \mathbb{Z}^2$, we introduce $K_{\underline{n}} = K + \underline{n} r_0$. The collection of sets $\{K_{\underline{n}}\}_{\underline{n} \in \mathbb{Z}^2}$ is a locally finite covering of $\mathbb{R}^2$. We consider an associated partition of unity defined as follows: let $0 \le \zeta \in C_c^\infty(K)$ be such that $\sum_{\underline{n} \in \mathbb{Z}^2} \zeta_{\underline{n}} = 1$ pointwise on $\mathbb{R}^2$, where $\zeta_{\underline{n}}(x) = \zeta(x - \underline{n} r_0)$ for all $x \in \mathbb{R}^2$. Then, for any $j \in \mathcal{P}$, we introduce the set

$$\mathcal{P}_j = \left\{ \underline{n} \in \mathbb{Z}^2 \,:\, d(p_j, K_{\underline{n}}) < \frac{r_0}{4} \right\}.$$

Note that the cardinality of $\mathcal{P}_j$ is uniformly bounded, namely, $|\mathcal{P}_j| \le 4$ for any $j \in \mathcal{P}$. We set

$$P_j = \bigcup_{\underline{n} \in \mathcal{P}_j} K_{\underline{n}}, \qquad \varphi_j = \sum_{\underline{n} \in \mathcal{P}_j} \zeta_{\underline{n}}$$

for any $j \in \mathcal{P}$. Since, by choice of r_0, $\mathbb{Z}^2 \setminus \bigcup_{j \in \mathcal{P}} \mathcal{P}_j$ is a countable set (even in the case that $\mathcal{P}$ is countable), there exists a bijection $\mathcal{I} : \mathbb{N} \to \mathbb{Z}^2 \setminus \bigcup_{j \in \mathcal{P}} \mathcal{P}_j$. For convenience of notation, we denote by $\mathcal{Q}$ the countable set of indices defined by

$$\mathcal{Q} = \mathcal{I}^{-1}\Big(\mathbb{Z}^2 \setminus \bigcup_{j \in \mathcal{P}} \mathcal{P}_j \Big).$$

We set

$$Q_k = K_{\mathcal{I}(k)}, \quad \psi_k = \zeta_{\mathcal{I}(k)} \quad \forall k \in \mathcal{Q}.$$

Then, $\{P_j, Q_k\}_{(j,k)\in\mathcal{P}\times\mathcal{Q}}$ is a locally finite open covering of $\mathbb{R}^2$ with the property that $P_j \cap P_{j'} = \emptyset$ for every $j' \neq j$. Moreover, $\{\varphi_j, \psi_k\}_{(j,k)\in\mathcal{P}\times\mathcal{Q}}$ is a partition of unity associated with $\{P_j, Q_k\}_{(j,k)\in\mathcal{P}\times\mathcal{Q}}$ such that

$$\operatorname{supp}\varphi_j \subset P_j, \qquad \operatorname{supp}\psi_k \subset Q_k$$

and such that

$$\sup_{j\in\mathcal{P}}\{\|\nabla\varphi_j\|_\infty,\ \|D^2\varphi_j\|_\infty\} < +\infty, \qquad \sup_{k\in\mathcal{Q}}\{\|\nabla\psi_k\|_\infty,\ \|D^2\psi_k\|_\infty\} < +\infty.$$

In particular,

$$0 \le \varphi_j, \psi_k \le 1 \qquad \text{and} \qquad \sum_{j\in\mathcal{P}}\varphi_j + \sum_{k\in\mathcal{Q}}\psi_k = \sum_{\underline{n}\in\mathbb{Z}^2}\zeta_{\underline{n}} = 1.$$

We define a rescaled covering

$$\hat{P}_j = P_j/\delta, \qquad \hat{Q}_k = Q_k/\delta.$$

Then, $\{\hat\varphi_j, \hat\psi_k\}_{(j,k)\in\mathcal{P}\times\mathcal{Q}}$ defined by

$$\hat\varphi_j(x) = \varphi_j(\delta x), \qquad \hat\psi_k(x) = \psi_k(\delta x)$$

is a partition of unity associated with $\{\hat{P}_j, \hat{Q}_k\}_{(j,k)\in\mathcal{P}\times\mathcal{Q}}$. It will also be convenient to define the sets

$$\hat{C}_j = \{x \in \hat{P}_j \,:\, \hat\varphi_j(x) = 1\}, \qquad j \in \mathcal{P}.$$

Note that

$$\operatorname{supp}\{\nabla\hat\varphi_j, D^2\hat\varphi_j\} \subset \hat{P}_j \setminus \hat{C}_j$$

and

$$\sup_{(j,k)\in\mathcal{P}\times\mathcal{Q}}\{\|\nabla\hat\varphi_j\|_\infty + \|\nabla\hat\psi_k\|_\infty\} \le C\delta, \qquad \sup_{(j,k)\in\mathcal{P}\times\mathcal{Q}}\{\|D^2\hat\varphi_j\|_\infty + \|D^2\hat\psi_k\|_\infty\} \le C\delta^2. \tag{4.1}$$

For every fixed $x \in \mathbb{R}^2$, we define the following subsets of indices:

$$J(x) = \{j \in \mathcal{P} \,:\, \hat\varphi_j(x) \neq 0\}, \qquad K(x) = \{k \in \mathcal{Q} \,:\, \hat\psi_k(x) \neq 0\}.$$

Note that, for every $x \in \mathbb{R}^2$,

$$|J(x)| \le 1, \qquad |K(x)| \le 4, \tag{4.2}$$

where $|J(x)|$ and $|K(x)|$ denote the cardinality of $J(x)$ and $K(x)$, respectively. We shall use the following Banach spaces:

$$\hat{X}_\delta = \left\{u \in H^2_{\mathrm{loc}}(\mathbb{R}^2) \,:\, \sup_{(j,k)\in\mathcal{P}\times\mathcal{Q}}\left\{\|\hat\varphi_j u\|_{H^2(\mathbb{R}^2)}, \|\hat\psi_k u\|_{H^2(\mathbb{R}^2)}\right\} < +\infty\right\},$$

$$\hat{Y}_\delta = \left\{f \in L^2_{\mathrm{loc}}(\mathbb{R}^2) \,:\, \sup_{(j,k)\in\mathcal{P}\times\mathcal{Q}}\left\{\|\hat\varphi_j f\|_{L^2(\mathbb{R}^2)}, \|\hat\psi_k f\|_{L^2(\mathbb{R}^2)}\right\} < +\infty\right\}.$$

We collect in the following lemma some estimates that will be used in what follows.

LEMMA 4.1. *There exists a constant $C>0$ such that for any $u\in\hat X_\delta$ and $j\in\mathcal{P}$, we have*

(i) $\|u\|_{H^2(\hat P_j)}\le C\|u\|_{\hat X_\delta}$,

(ii) $\|u\|_{L^\infty(\mathbb{R}^2)}\le C\|u\|_{\hat X_\delta}$.

Proof. (i) For every fixed $j\in\mathcal{P}$, let $\mathcal{J}(j)=\{k\in\mathcal{Q}\,:\,\operatorname{supp}\hat\varphi_j\cap\operatorname{supp}\hat\psi_k\neq\emptyset\}$. Then, $\sup_{j\in\mathcal{P}}|\mathcal{J}(j)|<+\infty$, and we estimate

$$\begin{aligned}\|u\|_{H^2(\hat P_j)}=&\Big\|\hat\varphi_j u+\sum_{k\in\mathcal{J}(j)}\hat\psi_k u\Big\|_{H^2(\hat P_j)}\le\|\hat\varphi_j u\|_{H^2(\hat P_j)}+\sum_{k\in\mathcal{J}(j)}\|\hat\psi_k u\|_{H^2(\hat P_j)}\\ &\le(1+|\mathcal{J}(j)|)\,\|u\|_{\hat X_\delta}\le C\|u\|_{\hat X_\delta}.\end{aligned}$$

(ii) For any fixed $x\in\mathbb{R}^2$, we have in view of (3.1) and (4.2)

$$\begin{aligned}|u(x)|&=\sum_{j\in\mathcal{P}}\hat\varphi_j(x)|u(x)|+\sum_{k\in\mathcal{Q}}\hat\psi_k(x)|u(x)|\\ &=\sum_{j\in J(x)}\hat\varphi_j(x)|u(x)|+\sum_{k\in K(x)}\hat\psi_k(x)|u(x)|\\ &\le\sum_{j\in J(x)}C_S\|\hat\varphi_j u\|_{H^2(\mathbb{R}^2)}+\sum_{k\in K(x)}C_S\|\hat\psi_k u\|_{H^2(\mathbb{R}^2)}\\ &\le\sup_{x\in\mathbb{R}^2}(|J(x)|+|K(x)|)C_S\|u\|_{\hat X_\delta}=C\|u\|_{\hat X_\delta}.\end{aligned}$$

Hence, (ii) is established. ∎

We shall also need the following family of functions:

$$\hat g_j(x)=\frac{\hat\varphi_j(x)}{\Big(\sum_{k\in\mathcal{P}}\hat\varphi_k^2+\sum_{k\in\mathcal{Q}}\hat\psi_k^2\Big)^{1/2}},\qquad \hat h_k(x)=\frac{\hat\psi_k(x)}{\Big(\sum_{j\in\mathcal{P}}\hat\varphi_j^2+\sum_{j\in\mathcal{Q}}\hat\psi_j^2\Big)^{1/2}}.$$

In view of (4.1), it is readily checked that the following lemma follows.

LEMMA 4.2. *The family $\{\hat g_j,\hat h_k\}_{(j,k)\in\mathcal{P}\times\mathcal{Q}}$ satisfies $\operatorname{supp}\hat g_j\subset\hat P_j$, $\operatorname{supp}\hat h_k\subset\hat Q_k$ and furthermore,*

(i) $\sum_{j\in\mathcal{P}}\hat g_j^2(x)+\sum_{k\in\mathcal{Q}}\hat h_k^2(x)=1\ \forall x\in\mathbb{R}^2$;

(ii) $C^{-1}\hat\varphi_j(x)\le\hat g_j(x)\le C\hat\varphi_j(x)$ *and* $C^{-1}\hat\psi_k(x)\le\hat h_k(x)\le C\hat\psi_k(x)\ \forall x\in\mathbb{R}^2$;

(iii) $\sup_{(j,k)\in\mathcal{P}\times\mathcal{Q}}\{\|\nabla\hat g_j\|_\infty+\|\nabla\hat h_k\|_\infty\}\le C\delta$ *and* $\sup_{(j,k)\in\mathcal{P}\times\mathcal{Q}}\{\|D^2\hat g_j\|_\infty+\|D^2\hat h_k\|_\infty\}\le C\delta^2$.

5. The shadowing lemma. Recall from the introduction that $\hat p_j=p_j/\delta$, $j\in\mathcal{P}$. For every $j\in\mathcal{P}$ we define

$$\hat U_j(x)=U_{m_j}(x-\hat p_j).$$

We make the following ansatz for solutions $\hat u$ to (2.4):

$$\hat u=\sum_{j\in\mathcal{P}}\hat\varphi_j\hat U_j+z. \tag{5.1}$$

Our aim in this section is to prove the following.

Proposition 5.1. *There exists $\delta_1>0$ such that for all $\delta\in(0,\delta_1)$ there exists $z_\delta\in\hat X_\delta$ such that $\hat u_\delta$ defined by $\hat u_\delta=\sum_{j\in\mathcal P}\hat\varphi_j\hat U_j+z_\delta$ is a solution of* (2.4). *Moreover, $\|z_\delta\|_{\hat X_\delta}\le C\mathrm{e}^{-c/\delta}$.*

We note that the functional $F_\delta:\hat X_\delta\to\hat Y_\delta$ given by

$$F_\delta(z)=-\Delta z+\sum_{j\in\mathcal P}\hat\varphi_j(1-\mathrm{e}^{\hat U_j})-(1-\mathrm{e}^{\sum_{j\in\mathcal P}\hat\varphi_j\hat U_j+z})-\sum_{j\in\mathcal P}[\hat\varphi_j,\Delta]\hat U_j$$

is well defined as well as C^1. Here $[\Delta,\hat\varphi_j]=\Delta\hat\varphi_j+2\nabla\hat\varphi_j\nabla$. Moreover, if $z\in\hat X_\delta$ satisfies $F_\delta(z)=0$, then $\hat u$ defined by (5.1) is a solution of (2.4).

Lemma 5.1. *For $\delta>0$ sufficiently small, we have*

$$\|F_\delta(0)\|_{\hat Y_\delta}\le C\mathrm{e}^{-c/\delta}\qquad \textit{as } \delta\to0^+ \tag{5.2}$$

for some constants $C,c>0$ independent of δ.

Proof. Let

$$\begin{aligned}\mathcal R&=\sum_{j\in\mathcal P}\hat\varphi_j(1-\mathrm{e}^{\hat U_j})-(1-\mathrm{e}^{\sum_{j\in\mathcal P}\hat\varphi_j\hat U_j}),\\ \mathcal C&=\sum_{j\in\mathcal P}[\hat\varphi_j,\Delta]\hat U_j.\end{aligned}$$

Note that $\{\operatorname{supp}\mathcal R,\operatorname{supp}\mathcal C\}\subset\cup_{j\in\mathcal P}\hat P_j\setminus\hat C_j$. We fix $x\in\cup_{j\in\mathcal P}\hat P_j$. We estimate

$$\begin{aligned}|\mathcal R(x)|\le&\sup_{j\in\mathcal P}\|\hat\varphi_j(1-\mathrm{e}^{\hat U_j})\|_{L^\infty(\hat P_j\setminus\hat C_j)}+\sup_{j\in\mathcal P}\|1-\mathrm{e}^{\hat\varphi_j\hat U_j}\|_{L^\infty(\hat P_j\setminus\hat C_j)}\\ \le& C\sup_{j\in\mathcal P}\|\hat U_j\|_{L^\infty(\hat P_j\setminus\hat C_j)}\le C_1\mathrm{e}^{-c_1/\delta}.\end{aligned}$$

On the other hand, in view of (4.1) and Lemma 3.1, for $x\in\cup_{j\in\mathcal P}\hat P_j$, we have

$$\begin{aligned}|\mathcal C(x)|\le&\sup_{j\in\mathcal P}\|\,[\Delta,\hat\varphi_j]\hat U_j\|_{L^\infty(\hat P_j\setminus\hat C_j)}\\ \le& C\left(\sup_{j\in\mathcal P}\|\hat U_j\,\Delta\hat\varphi_j\|_{L^\infty(\hat P_j\setminus\hat C_j)}+\sup_{j\in\mathcal P}\|\,|\nabla\hat U_j|\,|\nabla\hat\varphi_j|\,\|_{L^\infty(\hat P_j\setminus\hat C_j)}\right)\le C_2\mathrm{e}^{-c_2/\delta}.\end{aligned}$$

Here and above, $c_1,C_1,c_2,C_2>0$ are positive constants independent of $\delta>0$. Hence, we conclude that, as $\delta\to0^+$,

$$\|F_\delta(0)\|_{\hat Y_\delta}\le C\sup_{j\in\mathcal P}\left(\|\mathcal R\|_{L^2(\hat P_j)}+\|\mathcal C\|_{L^2(\hat P_j)}\right)\le C\mathrm{e}^{-c/\delta}$$

for some constants $C,c>0$ independent of $\delta>0$. ∎

Now, we consider the operator $L_\delta\equiv DF_\delta(0):\hat X_\delta\to\hat Y_\delta$ given by

$$L_\delta=-\Delta+\mathrm{e}^{\sum_{j\in\mathcal P}\hat\varphi_j\hat U_j}.$$

For every $j\in\mathcal P$, we define the operators

$$\hat L_j=-\Delta+\mathrm{e}^{\hat U_j}.$$

It will also be convenient to define

$$\hat{L}_0 = -\Delta + 1.$$

The following lemma holds.

Lemma 5.2. *There exist $C, c > 0$ such that for any $u \in \hat{X}_\delta$, we have*

$$\|(L_\delta - \hat{L}_j)\hat{\varphi}_j u\|_{L^2} \le C\mathrm{e}^{-c/\delta}\|\hat{\varphi}_j u\|_{L^2}, \qquad j \in \mathcal{P},$$
$$\|(L_\delta - \hat{L}_0)\hat{\psi}_k u\|_{L^2} \le C\mathrm{e}^{-c/\delta}\|\hat{\psi}_k u\|_{L^2}, \qquad k \in \mathcal{Q}.$$

Proof. For any $j \in \mathcal{P}$, by Lemma 3.1, we have, as $\delta \to 0^+$,

$$\begin{aligned}\|(L_\delta - \hat{L}_j)\hat{\varphi}_j u\|_{L^2} &\le \Big(\|1 - \mathrm{e}^{\hat{U}_j}\|_{L^\infty(\hat{P}_j \setminus \hat{C}_j)} + \|1 - \mathrm{e}^{\hat{\varphi}_j \hat{U}_j}\|_{L^\infty(\hat{P}_j \setminus \hat{C}_j)}\Big)\|\hat{\varphi}_j u\|_{L^2}\\ &\le C\|1 - \mathrm{e}^{\hat{U}_j}\|_{L^\infty(\hat{P}_j \setminus \hat{C}_j)}\|\hat{\varphi}_j u\|_{L^2} \le C\mathrm{e}^{-c/\delta}\|\hat{\varphi}_j u\|_{L^2}.\end{aligned}$$

Similarly, as $\delta \to 0^+$,

$$\begin{aligned}\|(L_\delta - \hat{L}_0)\hat{\psi}_k u\|_{L^2} \le& \|(1 - \mathrm{e}^{\sum_{j\in\mathcal{P}} \hat{\varphi}_j \hat{U}_j})\hat{\psi}_k u\|_{L^2}\\ &\le \sup_{j\in\mathcal{P}} \|1 - \mathrm{e}^{\hat{\varphi}_j \hat{U}_j}\|_{L^\infty(\hat{P}_j \setminus \hat{C}_j)}\|\hat{\psi}_k u\|_{L^2} \le C\mathrm{e}^{-c/\delta}\|\hat{\psi}_k u\|_{L^2}.\end{aligned}$$ ∎

Now, we prove an essential nondegeneracy property of L_δ.

Lemma 5.3. *There exists $\delta_0 > 0$ such that for any $\delta \in (0, \delta_0)$, the operator L_δ is invertible. Moreover, $L_\delta^{-1} : \hat{Y}_\delta \to \hat{X}_\delta$ is uniformly bounded with respect to $\delta \in (0, \delta_0)$.*

Proof. Following a gluing technique introduced in [5], we construct an "approximate inverse" $S_\delta : \hat{Y}_\delta \to \hat{X}_\delta$ for L_δ^{-1} as follows:

$$S_\delta = \sum_{j\in\mathcal{P}} \hat{g}_j \hat{L}_j^{-1} \hat{g}_j + \sum_{k\in\mathcal{Q}} \hat{h}_k \hat{L}_0^{-1} \hat{h}_k,$$

where $\hat{g}_j$ and $\hat{h}_k$ are the functions introduced in section 4. We claim that the operator S_δ is well defined and uniformly bounded with respect to δ. That is, we claim that

$$\|S_\delta f\|_{\hat{X}_\delta} \le C\|f\|_{\hat{Y}_\delta} \tag{5.3}$$

for some $C > 0$ independent of $f \in \hat{X}_\delta$ and of $\delta > 0$.

Indeed, for any $f \in \hat{Y}_\delta$, we have

$$\|S_\delta f\|_{\hat{X}_\delta} = \sup_{(j,k)\in\mathcal{P}\times\mathcal{Q}} \{\, \|\hat{\varphi}_j S_\delta f\|_{H^2}, \|\hat{\psi}_k S_\delta f\|_{H^2}\}$$

and

$$\|\hat{\varphi}_j S_\delta f\|_{H^2} \le \|\hat{\varphi}_j \hat{g}_j \hat{L}_j^{-1} \hat{g}_j f\|_{H^2} + \Big\|\hat{\varphi}_j \sum_{k\in\mathcal{Q}} \hat{h}_k \hat{L}_0^{-1} \hat{h}_k f\Big\|_{H^2},$$
$$\|\hat{\psi}_k S_\delta f\|_{H^2} \le \Big\|\hat{\psi}_k \sum_{j\in\mathcal{P}} \hat{g}_j \hat{L}_j^{-1} \hat{g}_j f\Big\|_{H^2} + \Big\|\hat{\psi}_k \sum_{j\in\mathcal{Q}} \hat{h}_j \hat{L}_0^{-1} \hat{h}_j f\Big\|_{H^2}.$$

We estimate, recalling the properties of $\hat{\varphi}_j$ and $\hat{g}_j$ as in Lemma 4.2, and in view of Lemma 3.2

$$\|\hat{\varphi}_j \hat{g}_j \hat{L}_j^{-1} \hat{g}_j f\|_{H^2} \le C\|\hat{L}_j^{-1} \hat{g}_j f\|_{H^2} \le C\|\hat{g}_j f\|_{L^2} \le C\|\hat{\varphi}_j f\|_{L^2} \le C\|f\|_{\hat{Y}_\delta}.$$

We have

$$\left\|\hat{\varphi}_j \sum_{k\in\mathcal{Q}} \hat{h}_k \hat{L}_0^{-1} \hat{h}_k f\right\|_{H^2} \le \left\|\hat{\varphi}_j \sum_{k\in\mathcal{J}(j)} \hat{h}_k \hat{L}_0^{-1} \hat{h}_k f\right\|_{H^2} \le \sum_{k\in\mathcal{J}(j)} \|\hat{\varphi}_j \hat{h}_k \hat{L}_0^{-1} \hat{h}_k f\|_{H^2},$$

where $\mathcal{J}(j) = \{k \in \mathcal{Q} \,:\, \operatorname{supp}\hat{\psi}_k \cap \operatorname{supp}\hat{\varphi}_j \ne \emptyset\}$ satisfies $\sup_{j\in\mathcal{P}} |\mathcal{J}(j)| < +\infty$. In view of Lemmas 4.2 and 3.2, we estimate

$$\begin{aligned}\sum_{k\in\mathcal{J}(j)} \|\hat{\varphi}_j \hat{h}_k \hat{L}_0^{-1} \hat{h}_k f\|_{H^2} &\le C \sum_{k\in\mathcal{J}(j)} \|\hat{L}_0^{-1} \hat{h}_k f\|_{H^2} \le C \sum_{k\in\mathcal{J}(j)} \|\hat{h}_k f\|_{L^2} \\ &\le C \sum_{k\in\mathcal{J}(j)} \|\hat{\psi}_k f\|_{L^2} \le C|\mathcal{J}(j)| \sup_{k\in\mathcal{Q}} \|\hat{\psi}_k f\|_{L^2} \le C\|f\|_{\hat{Y}_\delta}.\end{aligned}$$

Therefore,

$$\sup_{j\in\mathcal{P}} \left\|\hat{\varphi}_j \sum_{k\in\mathcal{Q}} \hat{h}_k \hat{L}_0^{-1} \hat{h}_k f\right\|_{H^2} \le C\|f\|_{\hat{Y}_\delta}.$$

Similarly, we obtain that

$$\sup_{k\in\mathcal{Q}} \left\|\hat{\psi}_k \sum_{j\in\mathcal{P}} \hat{g}_j \hat{L}_j^{-1} \hat{g}_j f\right\|_{H^2} \le C\|f\|_{\hat{Y}_\delta}, \qquad \sup_{k\in\mathcal{Q}} \left\|\hat{\psi}_k \sum_{j\in\mathcal{Q}} \hat{h}_j \hat{L}_0^{-1} \hat{h}_j f\right\|_{H^2} \le C\|f\|_{\hat{Y}_\delta},$$

and (5.3) follows.

Now, we claim that there exists δ_0 such that for any $\delta \in (0, \delta_0)$, the operator $S_\delta L_\delta : \hat{X}_\delta \to \hat{X}_\delta$ is invertible, and furthermore, $\|S_\delta L_\delta\| \le C$ for some $C > 0$ independent of $\delta > 0$. We note that $(L_\delta - \hat{L}_j)\hat{g}_j : \hat{X}_\delta \to \hat{Y}_\delta$ and $(L_\delta - \hat{L}_0)\hat{h}_k : \hat{X}_\delta \to \hat{Y}_\delta$ are well-defined bounded linear operators. Thus, we decompose

$$\begin{aligned} S_\delta L_\delta =& \mathbb{I}_{\hat{X}_\delta} + \sum_{j\in\mathcal{P}} \hat{g}_j \hat{L}_j^{-1}(\hat{g}_j L_\delta - \hat{L}_j \hat{g}_j) + \sum_{k\in\mathcal{Q}} \hat{h}_k \hat{L}_0^{-1}(\hat{h}_k L_\delta - \hat{L}_0 \hat{h}_k) \\ =& \mathbb{I}_{\hat{X}_\delta} + \sum_{j\in\mathcal{P}} \hat{g}_j \hat{L}_j^{-1}(L_\delta - \hat{L}_j)\hat{g}_j + \sum_{k\in\mathcal{Q}} \hat{h}_k \hat{L}_0^{-1}(L_\delta - \hat{L}_0)\hat{h}_k \\ &+ \sum_{j\in\mathcal{P}} \hat{g}_j \hat{L}_j^{-1}[\Delta, \hat{g}_j] + \sum_{k\in\mathcal{Q}} \hat{h}_k \hat{L}_0^{-1}[\Delta, \hat{h}_k]. \end{aligned} \tag{5.4}$$

Hence, it suffices to prove that the last four terms in (5.4) are sufficiently small, in the operator norm, provided $\delta > 0$ is sufficiently small. By Lemmas 5.2 and 4.2, we have, for any $u \in \hat{X}_\delta$,

$$\begin{aligned}\left\|\sum_{j\in\mathcal{P}} \hat{g}_j \hat{L}_j^{-1}(L_\delta - \hat{L}_j)\hat{g}_j u\right\|_{\hat{X}_\delta} &\le C \sup_{j\in\mathcal{P}} \|\hat{L}_j^{-1}(L_\delta - \hat{L}_j)\hat{g}_j u\|_{H^2} \\ &\le C \sup_{j\in\mathcal{P}} \|(L_\delta - \hat{L}_j)\hat{g}_j u\|_{L^2} \\ &\le C\mathrm{e}^{-c/\delta} \sup_{j\in\mathcal{P}} \|\hat{\varphi}_j u\|_{L^2} \le C\mathrm{e}^{-c/\delta}\|u\|_{\hat{X}_\delta}.\end{aligned}$$

Similarly, for $u \in \hat{X}_\delta$, we have

$$\Big\| \sum_{j\in\mathcal{P}} \hat{g}_j \hat{L}_j^{-1}[\Delta, \hat{g}_j]u \Big\|_{\hat{X}_\delta} \le C \sup_{j\in\mathcal{P}} \|\hat{L}_j^{-1}[\Delta, \hat{g}_j]u\|_{H^2} \le C \sup_{j\in\mathcal{P}} \|[\Delta, \hat{g}_j]u\|_{L^2}.$$

Recalling that $[\Delta, \hat{g}_j]u = 2\nabla u \nabla \hat{g}_j + u\Delta \hat{g}_j$, by Lemmas 4.2 and 4.1(i) we derive that

$$\|[\Delta, \hat{g}_j]u\|_{L^2} \le C\delta \|u\|_{H^1(\hat{P}_j)} \le C\delta \|u\|_{\hat{X}_\delta}.$$

The remaining terms are estimated similarly. Hence, $\|S_\delta L_\delta - \mathbb{I}_{\hat{X}_\delta}\| \to 0$ as $\delta \to 0^+$. Now, we observe that $L_\delta^{-1} = (S_\delta L_\delta)^{-1} S_\delta$. It follows that for any $f \in \hat{Y}_\delta$, we have

$$\|L_\delta^{-1} f\|_{\hat{X}_\delta} = \|(S_\delta L_\delta)^{-1} S_\delta f\|_{\hat{X}_\delta} \le C\|S_\delta f\|_{\hat{Y}_\delta} \le C\|f\|_{\hat{Y}_\delta}$$

with $C > 0$ independent of δ. Hence, L_δ is invertible and its inverse is bounded independently of δ, as asserted. □

Now we can provide the following proof.

Proof of Proposition 5.1. We use the Banach fixed point argument. For any $\delta \in (0, \delta_0)$, with $\delta_0 > 0$ given by Lemma 5.3, we introduce the nonlinear map $G_\delta \in C^1(\hat{X}_\delta, \hat{X}_\delta)$ defined by

$$G_\delta(z) = z - L_\delta^{-1} F_\delta(z).$$

Then, fixed points of G_δ correspond to solutions of the functional equation $F_\delta(z) = 0$. First, note that $DG_\delta(0) = 0$ and that

$$DF(z) = -\Delta + \mathrm{e}^{\sum_{j\in\mathcal{P}} \hat{\varphi}_j \hat{U}_j + z}.$$

By Lemma 5.3, for any $z \in \hat{X}_\delta$ and $u \in \hat{X}_\delta$, we have

$$\begin{aligned}\|DG_\delta(z)u\|_{\hat{X}_\delta} &= \|(DG_\delta(z) - DG_\delta(0))u\|_{\hat{X}_\delta} = \|L_\delta^{-1}(DF_\delta(z) - L_\delta)u\|_{\hat{X}_\delta} \\ &\le C\|(DF_\delta(z) - L_\delta)u\|_{\hat{Y}_\delta} = C\|\mathrm{e}^{\sum_{j\in\mathcal{P}} \hat{\varphi}_j \hat{U}_j}(\mathrm{e}^z - 1)u\|_{\hat{Y}_\delta} \le C\|(\mathrm{e}^z - 1)u\|_{\hat{Y}_\delta}.\end{aligned}$$

By the elementary inequality $\mathrm{e}^t - 1 \le Ct\mathrm{e}^t$, for all $t > 0$, where $C > 0$ does not depend on t, and in view of Lemma 4.1, we have

$$\|\mathrm{e}^z - 1\|_\infty \le \mathrm{e}^{\|z\|_\infty} - 1 \le C\|z\|_\infty \mathrm{e}^{\|z\|_\infty} \le C\|z\|_{\hat{X}_\delta} \mathrm{e}^{\|z\|_{\hat{X}_\delta}}.$$

Hence,

$$\|DG_\delta(z)u\|_{\hat{X}_\delta} \le C\|(\mathrm{e}^z - 1)u\|_{\hat{Y}_\delta} \le C\|z\|_{\hat{X}_\delta} \mathrm{e}^{\|z\|_{\hat{X}_\delta}} \|u\|_{\hat{Y}_\delta} \le C\|z\|_{\hat{X}_\delta} \mathrm{e}^{\|z\|_{\hat{X}_\delta}} \|u\|_{\hat{X}_\delta}.$$

Consequently, there exists $R_0 > 0$ such that for every $R \in (0, R_0)$, we have

$$\|DG_\delta(z)\| < \frac{1}{2} \qquad \forall z \in \mathcal{B}_R$$

for all $\delta > 0$, where

$$\mathcal{B}_R = \{u \in \hat{X}_\delta \,:\, \|u\|_{\hat{X}_\delta} < R\}.$$

Now, for every $R \in (0, R_0)$,

$$\begin{aligned}\|G_\delta(z)\|_{\hat{X}_\delta} &\le \|G_\delta(z) - G_\delta(0)\|_{\hat{X}_\delta} + \|G_\delta(0)\|_{\hat{X}_\delta} \\ &\le \frac{1}{2}\|z\|_{\hat{X}_\delta} + \|L_\delta^{-1}F_\delta(0)\|_{\hat{X}_\delta}.\end{aligned}$$

By Lemmas 5.3 and 5.1, there exist $C_0, c_0 > 0$ independent of $\delta > 0$ such that

$$\|L_\delta^{-1}F_\delta(0)\|_{\hat{X}_\delta} \le C\|F_\delta(0)\|_{\hat{Y}_\delta} \le C_0 \mathrm{e}^{-c_0/\delta}.$$

Choosing $R = R_\delta = 2C_0\mathrm{e}^{-c_0/\delta}$, we obtain that $G_\delta(B_{R_\delta}) \subset B_{R_\delta}$. Hence, G_δ is a strict contraction in B_{R_δ}, for any $\delta \in (0, \delta_1)$, with $\delta_1 = c_0/(\ln(2C_0/R_0))$. By the Banach fixed-point theorem, for any $\delta \in (0, \delta_1)$, there exists a unique $z_\delta \in B_{R_\delta}$ such that $F_\delta(z_\delta) = 0$. ◻

6. Proof of the main results. In this section, we finally provide the proof of Theorem 2.1 and derive Corollary 2.1. In view of Proposition 5.1, the function $\hat{u}_\delta$ defined by

$$\hat{u}_\delta = \sum_{j\in\mathcal{P}} \hat{\varphi}_j \hat{U}_j + z_\delta$$

is a solution of (2.4). Consequently, u_δ defined by

$$u_\delta(x) = \hat{u}_\delta\left(\frac{x}{\delta}\right) = \sum_{j\in\mathcal{P}} \varphi_j(x) U_{m_j}\left(\frac{x-p_j}{\delta}\right) + z_\delta\left(\frac{x}{\delta}\right) \tag{6.1}$$

is a solution of (1.6).

Lemma 6.1. *The solution u_δ defined in* (6.1) *satisfies the approximate superposition rule*

$$u_\delta(x) = \sum_{j\in\mathcal{P}} U_{m_j}\left(\frac{x-p_j}{\delta}\right) + \omega_\delta(x)$$

with $\|\omega_\delta\|_\infty \le C\mathrm{e}^{-c/\delta}$.

Proof. In view of (6.1) and of the definition of $J(x)$ in section 4, we have

$$u_\delta(x) = \sum_{j\in J(x)} U_{m_j}\left(\frac{x-p_j}{\delta}\right) + \tilde{\omega}_\delta(x),$$

where

$$\tilde{\omega}_\delta(x) = -\sum_{j\in J(x)} (1-\varphi_j(x)) U_{m_j}\left(\frac{x-p_j}{\delta}\right) + z_\delta\left(\frac{x}{\delta}\right).$$

In view of Lemma 3.1, we estimate

$$\left\| \sum_{j\in J(x)} (1-\varphi_j(x)) U_{m_j}\left(\frac{x-p_j}{\delta}\right)\right\|_\infty \le \sum_{j\in J(x)} \sup_{\mathbb{R}^2\setminus C_j} \left| U_{m_j}\left(\frac{x-p_j}{\delta}\right)\right| \le C\mathrm{e}^{-c/\delta}.$$

On the other hand, by Proposition 5.1 and Lemma 4.1(ii), we have

$$\left\| z_\delta\left(\frac{\cdot}{\delta}\right)\right\|_\infty = \|z_\delta\|_\infty \le C\mathrm{e}^{-c/\delta}.$$

Therefore, $\|\tilde\omega_\delta\|_\infty \le C\mathrm{e}^{-c/\delta}$. We have to show that

$$\left\| \sum_{j\in\mathcal{P}\setminus J(x)} U_{m_j}\left(\frac{x-p_j}{\delta}\right)\right\|_\infty \le C\mathrm{e}^{-c/\delta}.$$

To this end, we fix $x\in\mathbb{R}^2$ and for every $M\in\mathbb{N}$ we define $B_M=\{y\in\mathbb{R}^2\,:\,|y-x|<dM\}$. Then,

$$\sum_{j\in\mathcal{P}\setminus J(x)} U_{m_j}\left(\frac{x-p_j}{\delta}\right) = \sum_{M\in\mathbb{N}}\sum_{p_j\in B_{M+1}\setminus B_M} U_{m_j}\left(\frac{x-p_j}{\delta}\right).$$

Since $\inf_{j\neq k}|p_j-p_k|=d>0$, there exists $C>0$ independent of $M\in\mathbb{N}$ and of $x\in\mathbb{R}^2$ such that

$$|\{p_j\in B_{M+1}\setminus B_M\}|\le CM.$$

Hence, we estimate

$$\left| \sum_{j\in\mathcal{P}\setminus J(x)} U_{m_j}\left(\frac{x-p_j}{\delta}\right)\right| \le C\sum_{M\in\mathbb{N}} M\mathrm{e}^{-cM/\delta} \le C\mathrm{e}^{-c/\delta}.$$

This implies the statement of the lemma. □

We are left to analyze the asymptotic behavior of u_δ as $\delta\to 0^+$. Such behavior is a direct consequence of (6.1).

Lemma 6.2. *Let u_δ be given by (6.1). The following properties hold:*

(i) $\mathrm{e}^{u_\delta}<1$ *on $\mathbb{R}^2$ and vanishes exactly at p_j with multiplicity $2m_j$, $j\in\mathcal{P}$;*

(ii) *for every compact subset K of $\mathbb{R}^2\setminus\cup_{j\in\mathcal{P}}\{p_j\}$, there exist $C,c>0$ such that $1-\mathrm{e}^{u_\delta}\le C\mathrm{e}^{-c/\delta}$ as $\delta\to 0^+$;*

(iii) $\delta^{-2}(1-\mathrm{e}^{u_\delta})\to 4\pi\sum_{j\in\mathcal{P}} m_j\delta_{p_j}$ *in the sense of distributions as $\delta\to 0^+$.*

Proof. (i) Since u_δ is a solution of (1.6), $\mathrm{e}^{u_\delta}<1$ follows by the maximum principle. Moreover, since

$$U_{m_j}\left(\frac{x-p_j}{\delta}\right) = \ln|x-p_j|^{2m_j}+v_j \tag{6.2}$$

with v_j a continuous function (see [12]), we have near p_j that $\mathrm{e}^{u_\delta}=|x-p_j|^{2m_j}f_{j,\delta}(x)$ with $f_{j,\delta}(x)$ a continuous strictly positive function. Hence, (i) is established.

(ii) Let K be a compact subset of $\mathbb{R}^2\setminus\cup_{j\in\mathcal{P}}\{p_j\}$. In view of Lemma 3.1 and Proposition 5.1, we have as $\delta\to 0^+$

$$\begin{aligned}\sup_{x\in K\cap P_j} 1-\mathrm{e}^{\varphi_j(x)U_{m_j}((x-p_j)/\delta)} &\le C\mathrm{e}^{-c/\delta}\left\|z_\delta\left(\frac{\cdot}{\delta}\right)\right\|_\infty\\ &\le C\|z_\delta\|_{\hat X_\delta}\le CR_\delta\le C\mathrm{e}^{-c/\delta}.\end{aligned}$$

Therefore, we have that for any compact set $K \subset \mathbb{R}^2 \setminus \cup_{j\in\mathcal{P}}\{p_j\}$,

$$0 \le \sup_{x\in K}(1-\mathrm{e}^{u_\delta}) \le C\sup_{j\in\mathcal{P}}\sup_{x\in K\cap P_j}(1-\mathrm{e}^{u_\delta}) \le C\mathrm{e}^{-c/\delta}.$$

(iii) Let $\varphi \in C_c^\infty(\mathbb{R}^2)$. Then,

$$-\int_{\mathbb{R}^2} u_\delta\Delta\varphi = \delta^{-2}\int_{\mathbb{R}^2}(1-\mathrm{e}^u)\varphi - 4\pi\sum_{j\in\mathcal{P}} m_j\varphi(p_j).$$

We claim that

$$\int_{\mathbb{R}^2} u_\delta\Delta\varphi \to 0 \qquad \text{as } \delta\to 0. \tag{6.3}$$

Indeed, let $j_k \in \mathcal{P}$, $k=1,\dots,n$, be such that $\operatorname{supp}\varphi \subset \cup_{k=1}^n P_{j_k}\cup K$ with K a compact subset of $\mathbb{R}^2\setminus\cup_{j\in\mathcal{P}}\{p_j\}$. Since $\sup_K|u_\delta| \le C\mathrm{e}^{-c/\delta}$, we have

$$\left|\int_K u_\delta\Delta\varphi\right| \le C\|\Delta\varphi\|_\infty \mathrm{e}^{-c/\delta} \to 0.$$

On the other hand, in view of Lemma 6.1, in P_{j_k} we have $u_\delta(x) = U_{m_{j_k}}(|x-p_{j_k}|/\delta) + O(\mathrm{e}^{-c/\delta})$. Note that $U_{m_{j_k}} \in L^1(\mathbb{R}^2)$ in view of (6.2) and Lemma 3.1. Therefore,

$$\begin{aligned}\sup_{1\le k\le n}\left|\int_{P_{j_k}} u_\delta\Delta\varphi\right| &\le \sup_{1\le k\le n}\left|\int_{P_{j_k}} U_{m_{j_k}}\left(\frac{x-p_{j_k}}{\delta}\right)\Delta\varphi\right| + O(\mathrm{e}^{-c/\delta})\\ &\le \delta^2\sup_{1\le k\le n}\|\Delta\varphi\|_\infty\|U_{m_{j_k}}\|_{L^1} + O(\mathrm{e}^{-c/\delta}) \le C\delta^2 \to 0.\end{aligned}$$

Hence, (6.3) follows, and (iii) is established. ∎

Proof of Theorem 2.1. For every $\delta\in(0,\delta_1)$, where δ_1 is given in Proposition 5.1, we obtain a solution u_δ of (1.6). Furthermore, u_δ satisfies (2.2) in view of Lemma 6.1. Finally, u_δ satisfies the asymptotic behavior as in (i)–(iii) in view of Lemma 6.2. Hence, Theorem 2.1 is completely established. ∎

Proof of Corollary 2.1. To begin, we want to prove that if p_j's are doubly periodically arranged in $\mathbb{R}^2$, then u_δ is in fact a doubly periodic solution of (1.5). Recall that the p_j's are doubly periodically arranged in $\mathbb{R}^2$ if (2.3) holds. We define $\hat{\underline{e}}_k = \underline{e}_k/\delta$, $k=1,2$. Equivalently, we show $\hat{u}_\delta(x+\hat{\underline{e}}_k) = \hat{u}_\delta(x)$ for any $x\in\mathbb{R}^2$ and for $k=1,2$. Indeed, we may assume that $\hat{\varphi}_j(x+\hat{\underline{e}}_k) = \hat{\varphi}_j(x)$, $\hat{\psi}_j(x+\hat{\underline{e}}_k) = \hat{\psi}_j(x)$ for any $j\in\mathbb{N}$, $x\in\mathbb{R}^2$, $k=1,2$. Then,

$$\hat{u}_\delta(x+\hat{\underline{e}}_k) = \sum_{j\in\mathbb{N}}\hat{\varphi}_j(x)\hat{U}_j(x) + z_\delta(x+\hat{\underline{e}}_k).$$

Hence, it is sufficient to prove that $z_\delta(x+\hat{\underline{e}}_k) = z_\delta(x)$ for every $x\in\mathbb{R}^2$ and for $k=1,2$. First, we claim that $z_\delta(\,\cdot\,+\hat{\underline{e}}_k) \in \mathcal{B}_{R_\delta}$. Indeed, for every $j\in\mathbb{N}$ there exists exactly one $j'\in\mathbb{N}$ such that

$$\|\hat{\varphi}_j z_\delta(\,\cdot\,+\hat{\underline{e}}_k)\|_{H^2} = \|\hat{\varphi}_{j'}z_\delta\|_{H^2}. \tag{6.4}$$

Hence, we obtain

$$\|z_\delta(\,\cdot\,+\hat{\underline{e}}_k)\|_{\hat{X}_\delta} = \|z_\delta\|_{\hat{X}_\delta} \le R_\delta. \tag{6.5}$$

Moreover, if $F_\delta(z_\delta) = 0$ we also have $F_\delta(z_\delta(\,\cdot\, + \hat{\underline{e}}_k)) = 0$. Therefore, $z_\delta(\,\cdot\, + \hat{\underline{e}}_k)$ is a fixed point of G_δ in $\mathcal{B}_{R_\delta}$. By uniqueness, we conclude that $z_\delta(\,\cdot\, + \hat{\underline{e}}_k) = z_\delta$, $k = 1, 2$, as asserted. At this point, the remaining statements follow recalling that in the periodic cell domain Ω, (A_δ, ϕ_δ) is given, up to gauge transformations, by (1.4). □

REFERENCES

[1] A.A. Abrikosov, *On the magnetic properties of superconductors of the second group*, Sov. Phys. JETP, 5 (1957), pp. 1174–1182.
[2] A. Aftalion, E. Sandier, and S. Serfaty, *Pinning phenomena in the Ginzburg-Landau model of superconductivity*, J. Math. Pures Appl., 80 (2001), pp. 339–372.
[3] S. Alama and L. Bronsard, *Vortices and pinning effects for the Ginzburg-Landau model in multiply connected domains*, Comm. Pure Appl. Math., to appear.
[4] N. André, P. Bauman, and D. Phillips, *Vortex pinning with bounded fields for the Ginzburg-Landau equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 705–729.
[5] S. Angenent, *The shadowing lemma for elliptic PDE*, in Dynamics of Infinite-Dimensional Systems, NATO Adv. Sci. Inst. Ser. F Comput. Syst. Sci. 37, Springer, Berlin, 1987, pp. 7–22.
[6] F. Bethuel, H. Brezis, and F. Hélein, *Ginzburg-Landau Vortices,* Birkhäuser, Boston, MA, 1994.
[7] W. E, *Dynamics of vortices in the Ginzburg-Landau theories with applications to superconductivity*, Phys. D, 77 (1994), pp. 383–404.
[8] O. García-Prada, *A direct existence proof for the vortex equations over a compact Riemann surface*, Bull. London Math. Soc., 26 (1994), pp. 88–96.
[9] D. Gilbarg and N. Trudinger, *Elliptic partial differential equations of second order*, in Classics in Mathematics, Springer, Berlin, 2001.
[10] M.Y. Hong, J. Jost, and M. Struwe, *Asymptotic limits of a Ginzburg Landau type functional*, in Geometric Analysis and Calculus of Variations, International Press, Cambridge, MA, 1996, pp. 99–123.
[11] G. 't Hooft, *A property of electric and magnetic flux in nonabelian gauge theories*, Nuclear Phys. B, 153 (1979), pp. 141–160.
[12] A. Jaffe and C. Taubes, *Vortices and Monopoles*, Birkhäuser, Boston, MA, 1980.
[13] F.H. Lin, *Some dynamical properties of Ginzburg-Landau vortices*, Comm. Pure Appl. Math., 49 (1996), pp. 323–442.
[14] J.C. Neu, *Vortices in the complex scalar field*, Phys. D, 43 (1990), pp. 385–406.
[15] M. Nolasco, *Non topological n-vortex condensates for the selfdual Chern-Simons theory*, Comm. Pure Appl. Math., 56 (2003), pp. 1752–1780.
[16] J. Rubinstein and P. Sternberg, *On the slow motion of vortices in the Ginzburg-Landau heat flow*, SIAM J. Math. Anal., 26 (1995), pp. 1452–1466.
[17] D. Stuart, *Dynamics of Abelian Higgs vortices in the near Bogomolny regime*, Comm. Math. Phys., 159 (1994), pp. 51–91.
[18] G. Tarantello, *The Analysis of Selfdual Gauge Field Vortices,* Progr. Nonlinear Differential Equations Appl., Birkhäuser, Boston, MA, in preparation.
[19] C. Taubes, *Arbitrary n-vortex solutions to the first order Ginzburg-Landau equations*, Comm. Math. Phys., 72 (1980), pp. 277–292.
[20] S. Wang and Y. Yang, Abrikosov's vortices in the critical coupling, SIAM J. Math. Anal., 23 (1992), pp. 1125–1140.
[21] M.I. Weinstein and J. Xin, *Dynamic stability of vortex solutions of Ginzburg-Landau and nonlinear Schrödinger equations*, Comm. Math. Phys., 180 (1996), pp. 389–428.
[22] Y. Yang, *Solitons in Field Theory and Nonlinear Analysis,* Springer Monogr. Math., Springer, New York, 2001.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 17–59

LOCAL GEOMETRY OF DEFORMABLE TEMPLATES*

ALAIN TROUVÉ† AND LAURENT YOUNES‡

Abstract. In this paper, we discuss a geometrical model of a space of deformable images or shapes, in which infinitesimal variations are combinations of elastic deformations (warping) and of photometric variations. Geodesics in this space are related to velocity-based image warping methods, which have proved to yield efficient and robust estimations of diffeomorphisms in the case of large deformation. Here, we provide a rigorous and general construction of this infinite dimensional "shape manifold" on which we place a Riemannian metric. We then obtain the geodesic equations, for which we show the existence and uniqueness of solutions for all times. We finally use this to provide a geometrically founded linear approximation of the deformations of shapes in the neighborhood of a given template.

Key words. infinite dimensional Riemannian manifolds, deformable templates, shape representation and recognition, warping

AMS subject classifications. Primary, 58b10; Secondary, 49J45, 68T10

DOI. 10.1137/S0036141002404838

1. Introduction.

The theoretical developments which are addressed in this paper are motivated by the theory of deformable templates, as it emerged from the work of Grenander and his collaborators in the 1980's [19, 21, 22, 20], to handle image processing problems. This theory has an abstract formulation, in which the purpose is to represent the variability within an object class by the variations in shape, or color, etc., of a single object, submitted to the action of "deformations." For instance, a model designed to describe a picture of a human face should be able to explain inter-individual variations but also variations caused by the change of expression of a given individual, and by the changing of imaging conditions, such as lighting, occultations, etc. The interesting feature in Grenander's construction is that it assigns a large part, sometimes all, of the variations to a fixed structure, describing the deformation, which is independent of the particular instance of the observed image. This structure most of the time belongs to a group, the group of deformations, which is acting on the set of objects. The specific choice of the group depends on the application and on the type of visual features which are modeled, like pixelized images [18] and discretized shapes [20, 29, 23]. In such discrete settings, the group action is used to generate variations of the constituting generators of the object (pixels for an image, segments for polygons) and therefore are modeled as finite dimensional groups, generally products of linear or affine groups. In the simple example of labeled collections of points (landmarks), the deformation may simply correspond to independent translations of each point, but when the question is raised of the similarity of two collections of landmarks, one would like to figure out the amount of deformation which is required to transform one of them into the other. When evaluating this deformation, it is clear that the lengths of the induced translations should have some impact, but that this is not the

*Received by the editors March 29, 2002; accepted for publication (in revised form) June 11, 2004; published electronically August 17, 2005.
http://www.siam.org/journals/sima/37-1/40483.html

†LAGA (UMR CNRS 7593), Institut Galilée, Université Paris XIII, Av. J-B. Clément, F-93430 Villetaneuse, France (trouve@math.univ-paris13.fr).

‡CMLA (CNRS, URA 1611), Ecole Normale Supérieure de Cachan, 61 avenue du Président Wilson, F-94235 Cachan Cedex, France (younes@cmla.ens-cachan.fr).

only factor and often not even the main factor. One would also like to draw conclusions on the smoothness of the deformation, based on the fact that, in the context of large deformations of shapes, a lower similarity must be associated to a collection of translations which point to erratic directions, compared to a more homogeneous displacement. We see, in this case, that a global point of view on the displacements is needed. Spline-based landmark matching [9, 26] specifically addresses this issue by seeking the smoothest function which interpolates the considered displacements.

When dealing with image deformation, the need to pass to the continuum is even more obvious. In this case, deformations, which should provide nonambiguous point displacements, must be diffeomorphisms on the image support. This nonambiguity constraint, however, has been relaxed in most of the early attempts to deal with this issue, working preferably with linear spaces of deformations [6, 7, 8, 2, 1, 14], which can be seen as first order approximations. Dealing explicitly with true deformations, i.e., diffeomorphisms acting on the support of images, was rigorously formalized by Riemannian metric arguments on the groups of diffeomorphisms in [32] for one-dimensional problems, and in [31] in full generality (see also [30]). Stemming from the simple representation of right invariant metrics on groups of diffeomorphisms along a path in this space, i.e., time-dependent deformations, in terms of the Eulerian velocity, this last reference built diffeomorphisms as flows associated to ODEs (a construction which was already present in [3]) and transferred the modeling effort to the linear space of velocities, i.e., of vector fields defined on the image support. Under suitable Banach space structures on these linear spaces, the extension of the ODE solutions for infinite time and the existence of minimizers to general variational problems in this space can be ensured, providing rigorous sufficient conditions for the well-posedness of many practical problems in template matching. This analysis rejoined the line of work of Miller and his collaborators on the estimation of large deformation diffeomorphisms [13, 26], in which velocity-based models have been introduced, and variational properties studied in [16]. In [27], the interest in considering a lifted group action, on the cross product of the group itself and of the image space, was demonstrated in a wide variety of applications. The final metric on the image space was obtained by projecting a right-invariant Riemannian distance designed on the product space.

The approach we follow in this paper addresses the same kind of construction as in [27], which focused on the metric aspects, but from a different point of view. Our purpose is to start from the infinitesimal analysis of small deformations of images in order to model and measure image variations and define differentiable and geodesic curves in the image space. We shall accept conditions which ensure enough smoothness on the diffeomorphisms but try whenever possible to avoid placing such smoothness assumptions on the images themselves. Such a choice, which is very important given the discontinuous nature of images, is made at the cost of increasing technicalities and notation, as will be seen in section 3, in which the basic geometry of the model is presented. Here, we define the tangent space at a given square integrable image i as an equivalent class for all possible variations resulting from an infinitesimal combination of a deformation (geometry) and of the addition of a square integrable function (photometry), yielding what can be called a morphometrical variation. We then equip it with an inner product and define from it lengths and energies of curves. This metric is based on the best tradeoff between geometrical and photometrical variations. Still, in this general setting, we show the existence of minimizing geodesics (curves of minimal energy) between any two images.

The rest of the paper is devoted to the study of geodesics and their generation from initial conditions. The motivation in this study is the possibilities it offers for prototype-based image representation and the generation of image variations and deformations from initial conditions belonging to a vector space. In this context, the geodesic equations are derived under the assumption that the deformed prototype is smooth (H^1), but with no restriction on the other endpoint. This is done in section 5.2.2. The obtained evolution equations are then generalized to a form which does not require the smoothness of the initial position and that we conjecture to represent a comprehensive class of image evolutions. The equations, under this form, are studied in section 7, where we prove that they have a unique solution over arbitrary finite time intervals. Our last result shows the local nonambiguity of this representation, at least in the smooth case: from a smooth prototype, the solutions of the geodesic equations in small time cannot coincide if they have been generated from distinct smooth initial conditions. This is done in section 9. The last section, 10, presents numerical experiments, which illustrate the feasibility of retrieving a target from the initial conditions associated to the minimizing geodesic starting from the template.

2. Notation. For further reference, we present in a single definition some of the main functional spaces we use throughout the paper.

DEFINITION 1. *Let $k, p \in \mathbb{N}_*$, $l \in \mathbb{N}$, and Ω be a bounded domain of $\mathbb{R}^k$ with C^1 boundary.*

(1) *We denote $C_c^\infty(\Omega, \mathbb{R}^p)$ the space of smooth compactly supported $\mathbb{R}^p$-valued functions on Ω.*

(2) *We denote $C^l(\overline{\Omega}, \mathbb{R}^p)$ the set of the restrictions to Ω of the l times continuously differentiable $\mathbb{R}^p$-valued functions on $\mathbb{R}^k$.*
Let $f \in C^l(\overline{\Omega}, \mathbb{R}^p)$. We define the norm $|f|_{l,\infty}$ by

$$|f|_{l,\infty} \triangleq \sum_{\alpha,\ 0 \le |\alpha| \le l} \sup_{x \in \Omega} |\frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}},$$

where for any $\alpha \triangleq (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_*^d$ we denote $|\alpha| \triangleq \sum \alpha_i$.

(3) *We denote $C_0^l(\Omega, \mathbb{R}^p)$ the completion of $C_c^\infty(\Omega, \mathbb{R}^p)$ for the norm $|\ |_{l,\infty}$.*

(4) *We denote $L^2(\Omega, \mathbb{R}^p)$ the Hilbert space of square integrable functions in $\mathbb{R}^p$ with dot product defined for $f, g \in L^2(\Omega, \mathbb{R}^p)$ by*

$$\langle f, g \rangle_2 \triangleq \int_\Omega \langle f(x), g(x) \rangle_{\mathbb{R}^p} dx.$$

(5) *We denote $H^1(\Omega, \mathbb{R}^p)$ the Hilbert space of square integrable $\mathbb{R}^p$-valued functions with square integrable first partial (generalized) derivatives. The dot product is defined for any $f, g \in H^1(\Omega, \mathbb{R}^p)$ by*

$$\langle f, g \rangle_{H^1} \triangleq \langle f, g \rangle_2 + \sum_{i=1}^k \left\langle \frac{\partial f}{\partial x_i}, \frac{\partial g}{\partial x_i} \right\rangle_2 .$$

3. Measuring distances on the image space.

3.1. Infinitesimal transformations. Let us consider a space $\mathcal{J}_W$ of functions defined on $\bar{\Omega}$, and taking values on $\mathbb{R}^d$, which will be explicitly defined later. To somewhat fix the ideas, we shall speak of elements of $\mathcal{J}_W$ as "images" and use the corresponding photometric vocabulary, although our constructions apply to generic graphs of vector-valued functions.

We want to build a distance, denoted hereafter $d_{\mathcal{J}_W}$, on $\mathcal{J}_W$ through a Riemannian analysis. Let $j \in \mathcal{J}_W$ and $h \in \mathbb{R}$, and consider a small perturbation j_h of j such that

$$j_h(x) = j(x - hv(x)) + h\sigma^2 z(x) + o(h),$$

where v is a displacement field and z is an $\mathbb{R}^d$-valued function on Ω. Here and in the following, σ^2 is a fixed positive parameter. The transformation from j to j_h is therefore divided in two complementary processes. The first, which we call the "geometric transformation," is a pure deformation of the support for which a point located at x in the first image is pushed to location $x + hv(x)$. The second process, called the "photometric transformation," is the residual, obtained by the addition of $\sigma^2 hz$. Both transformations are the main ingredients of any morphing process between two images. When j is smooth, we have

$$\frac{\partial j}{\partial h}_{|h=0} \triangleq \lim_{h\to 0} \frac{j_h - j}{h} = \sigma^2 z - dj(v). \tag{1}$$

The usual geometric interpretation is that $\gamma \triangleq \frac{\partial j}{\partial h}_{|h=0}$ is an element of the tangent space $T_j\mathcal{J}_W$, and, given our representation, it is sensible to let the length $|\gamma|_j$ depend on $w \triangleq (z, v)$ and to let w vary in some chosen vector space W. The solution cannot merely be to set $|\gamma|_j = |w|_W$, where $|\ |_W$ is a norm on W, because the representation $(z, v) \mapsto \gamma$ is not one-to-one: if $w' = (v', z')$ is such that

$$\sigma^2(z' - z) - dj(v' - v) = 0, \tag{2}$$

then the transformations along w and w' of j are infinitesimally equivalent. Hence, looking for the best tradeoff between geometric and photometric transformations, we can choose for the metric on the tangent space $T_j\mathcal{J}_W$

$$|\gamma|_j = \inf \left\{ |w|_W \mid w = (v, z),\ \gamma = \sigma^2 z - dj(v)\ \right\}. \tag{3}$$

Now, we can define formally

$$d_{\mathcal{J}_W}(j_0, j_1) \triangleq \inf \left\{ \int_0^1 \left| \frac{\partial \mathrm{j}}{\partial t} \right|_{j_t},\ \mathrm{j} \text{ path from } j_0 \text{ to } j_1 \right\}. \tag{4}$$

3.2. Differentiable structure. The previous construction is now made rigorous for $\mathcal{J}_W \triangleq L^2(\Omega, \mathbb{R}^d)$.

Remark 1. Since $L^2(\Omega, \mathbb{R}^d)$ is a Hilbert space, it has a natural structure of smooth infinite dimensional manifold. However, the differential structure we need to consider here is different from the standard L^2 structure. To see this, consider the following example: $\Omega =]0, 1[^k$, and $j_h(x) \triangleq j_0(x - hv(x))$, where

- $j_0(x) \triangleq \mathbf{1}_{x_1 \geq 1/2}$,
- $v \in C_c^\infty(\Omega, \mathbb{R}^k)$ is such that the first coordinate, v_1, of v is strictly positive at the center $c \triangleq (1/2, \ldots, 1/2)$ of Ω.

Then, $|j_h - j_0|_2/h \to +\infty$ so that j_h is not differentiable at $h = 0$ for the usual L^2 differentiable structure, whereas, by the construction above, it will be so for the differential structure on $\mathcal{J}_W$ (this is a justification for keeping the nonstandard notation $\mathcal{J}_W$ for the image space).

Our construction starts with the definition of C^1 paths on $\mathcal{J}_W$. We first need to specify the allowed geometric as well as grey-level infinitesimal transformations.

3.2.1. Infinitesimal transformation spaces.

Geometric transformation. We denote $\mathcal{B}$ the space of the displacement fields underlying the infinitesimal geometric transformation. We assume that $\mathcal{B}$ is a Hilbert space with dot product denoted by $\langle\, ,\, \rangle_{\mathcal{B}}$ and norm denoted by $|\ |_{\mathcal{B}}$. We assume throughout this paper that $\mathcal{B}$ is continuously embedded in $C_0^p(\Omega, \mathbb{R}^k)$, where $p = 1$ at least but may be larger if specified. As a reminder, we recall that $\mathcal{B}$ is continuously embedded in some Banach space $\mathcal{B}'$ (with norm $|\ |_{\mathcal{B}'}$) of functions if and only if each element v of $\mathcal{B}$ can be considered as an element of $\mathcal{B}'$ and there exists a constant C such that, for all $v \in \mathcal{B}$,

$$|v|_{\mathcal{B}'} \leq C\, |v|_{\mathcal{B}}\, .$$

Moreover, $\mathcal{B}$ is compactly embedded in $\mathcal{B}'$ if it is continuously embedded and any bounded set for the norm on $\mathcal{B}$ is relatively compact in the $\mathcal{B}'$-topology.

We shall also assume that $C_c^\infty(\Omega, \mathbb{R}^k)$ is dense in $\mathcal{B}$.

Photometric transformation. Grey-level transformations are assumed to belong to the space $L^2(\Omega, \mathbb{R}^d)$.

Finally, we denote $W \triangleq \mathcal{B} \times L^2(\Omega, \mathbb{R}^d)$ on which we place the dot product defined for $w = (v, z)$ and $w' = (v', z')$ by

$$\langle w, w' \rangle_W \triangleq \langle v, v' \rangle_{\mathcal{B}} + \sigma^2 \langle z, z' \rangle_2.$$

3.2.2. Differentiable curves and tangent space. For any smooth image j, we have, for any $u \in C_c^\infty(\Omega, \mathbb{R}^d)$ and any $w = (v, z) \in W$,

$$\langle \sigma^2 z - dj(v), u \rangle_2 = \sigma^2 \langle z, u \rangle_2 + \langle j, \operatorname{div}(u \otimes v) \rangle_2, \tag{5}$$

where $\operatorname{div}(u \otimes v) \in C_0(\Omega, \mathbb{R}^d)$ is defined by $\operatorname{div}(u \otimes v)_i = \operatorname{div}(u_i v)$. The right-hand side of the equality is well defined for arbitrary $j \in \mathcal{J}_W$, which leads us to the following definition.

Definition 2 (C^1 curves in $\mathcal{J}_W$). *Let I be an interval in $\mathbb{R}$. We say that* $\mathrm{j} : I \to \mathcal{J}_W$ *is a continuously differentiable curve if there exists* $\mathrm{w} \triangleq (\mathrm{v}, \mathrm{z}) \in C(I, W)$ *such that*

(1) $\mathrm{j} \in C(I, L^2(\Omega, \mathbb{R}^d))$ *for the usual L^2-topology,*

(2) *for any $u \in C_c^\infty(\Omega, \mathbb{R}^d)$, $t \to \langle \mathrm{j}_t, u \rangle_2$ is a continuously differentiable real-valued function and* $\frac{\partial}{\partial t} \langle \mathrm{j}_t, u \rangle_2 = \sigma^2 \langle \mathrm{z}_t, u \rangle_2 + \langle \mathrm{j}_t, \operatorname{div}(u \otimes \mathrm{v}_t) \rangle_2$.

If we define as usual tangent vectors via classes of first order equivalent curves, we can identify the tangent bundle of $\mathcal{J}_W$ from the definition of C^1 path on $\mathcal{J}_W$ as follows.

Definition 3.

(1) *For any $j \in \mathcal{J}_W$ and any $u \in C_c^\infty(\Omega, \mathbb{R}^d)$, we denote $l_{j,u}$ the continuous linear form on W (the continuity stems from the continuous embedding of $\mathcal{B}$ in $C_0^1(\Omega, \mathbb{R}^k)$) defined for any $w = (v, z) \in W$ by*

$$l_{j,u}(w) \triangleq \sigma^2 \langle z, u \rangle_2 + \langle j, \operatorname{div}(u \otimes v) \rangle_2. \tag{6}$$

(2) *We define*

$$E_j \triangleq \{\ w \in W \mid l_{j,u}(w) = 0,\ \forall u \in C_c^\infty(\Omega, \mathbb{R}^d)\ \}, \tag{7}$$

and

$$T_j \mathcal{J}_W \triangleq \{j\} \times W / E_j, \tag{8}$$

where W/E_j is the quotient space, the elements of which are denoted $\overline{w}$.

Remark 2. The use of a quotient space is a consequence of the nonuniqueness of the representation of the derivative by an element $w \in W$ as explained by (2).

We consider $T_j\mathcal{J}_W$ as a vector space where for any $\gamma = (j, \overline{w})$ and $\gamma' = (j', \overline{w}') \in T_j\mathcal{J}_W$, we have $\gamma + \lambda'\gamma' \triangleq (j, \overline{w} + \lambda\overline{w}')$. Now, if we define

$$T\mathcal{J}_W \triangleq \bigcup_{j \in \mathcal{J}_W} T_j\mathcal{J}_W,$$

$T\mathcal{J}_W$ plays the role of the tangent bundle of the manifold $\mathcal{J}_W$.

Definition 4.

(1) *We denote $\pi : T\mathcal{J}_W \to \mathcal{J}_W$ the canonical projection defined by $\pi(\gamma) = j$ for any $\gamma \triangleq (j, \overline{w}) \in T_j\mathcal{J}_W$.*

(2) *Let $\gamma \triangleq (j, \overline{w}) \in T\mathcal{J}_W$ and $w = (z, v) \in \overline{w}$. For any $u \in C_c^\infty(\Omega, \mathbb{R}^d)$, we denote*

$$\langle \gamma\, ,\, u \rangle \triangleq \sigma^2 \langle z, u \rangle_2 + \langle j, \mathrm{div}(u \otimes v) \rangle_2.$$

(Note that the right-hand side does not depend on the choice of $w \in \overline{w}$).

(3) *For any function $\gamma : I \to T\mathcal{J}_W$ where I is a real interval, we say that γ is measurable if $\pi \circ \gamma$ is measurable from I to $\mathcal{J}_W$ and for any $u \in C_c^\infty(\Omega, \mathbb{R}^d)$, $\langle \gamma_t\, ,\, u \rangle$ is measurable from I to $\mathbb{R}$.*

Returning to Definition 2, we see that C^1 curves j admit a lifting $t \mapsto \gamma_t = (\mathrm{j}_t, \overline{w}_t)$ to $T\mathcal{J}_W$ such that for all $u \in C_c^\infty(\Omega, \mathbb{R}^d)$

$$\frac{d}{dt} \langle \mathrm{j}_t\, ,\, u \rangle_2 = \langle \gamma_t\, ,\, u \rangle$$

so that it is natural to define $\frac{d\mathrm{j}_t}{dt} \triangleq \gamma_t \in T_{\mathrm{j}_t}\mathcal{J}_W$ leading to the formula

$$\frac{d}{dt} \langle \mathrm{j}_t\, ,\, u \rangle_2 = \left\langle \frac{d\mathrm{j}_t}{dt}\, ,\, u \right\rangle. \tag{9}$$

The next step, for our Riemannian construction, is to place a metric on $T_j\mathcal{J}_W$ for all $j \in \mathcal{J}_W$.

3.3. Riemannian structure.

Definition 5. *For any $j \in \mathcal{J}_W$, we define on $T_j\mathcal{J}_W$ the norm*

$$|\gamma|_j \triangleq \inf\{\ |w|_W \ |\ (j, w) \in \gamma\ \}.$$

The infimum is attained at a unique point, as stated in the following proposition.

Proposition 1. *For any $j \in \mathcal{J}_W$ and any $\gamma = (j, \overline{w}) \in T_j\mathcal{J}_W$, since $\overline{w}$ is a closed subspace of W, there exists a unique $w \in W$ denoted $\overline{p}(\gamma)$ such that*

$$\overline{p}(\gamma) \triangleq \operatorname*{Argmin}_{w \in \overline{w}} |w|_W.$$

Hence, $|\gamma|_j \triangleq |\overline{p}(\gamma)|_W$. Moreover, $\overline{p}$ is linear from $T_j\mathcal{J}_W$ to W.

Proof. Since $\overline{w}$ is a close subspace of W, it is sufficient to note that if p is the orthogonal projection from W to $E_j^\perp$, then $p(w) = 0$ for any $w \in E_j$ so that p can be factorized as a linear map $\overline{p}$ from W/E_j to $E_j^\perp$. Now, one easily checks that $\overline{p}(\gamma) \in \overline{w}$ and that $\overline{p}(\gamma)$ minimizes the norm. ∎

We can now define the geodesic distance between arbitrary points j_0, j_1 in $\mathcal{J}_W$ by

$$d_{\mathcal{J}_W}(j_0, j_1) \triangleq \inf \left\{ \int_0^1 \left| \frac{d\mathrm{j}}{dt} \right|_{\mathrm{j}_t} dt \mid \mathrm{j} \in C^1_{\mathrm{pw}}([0,1], \mathcal{J}_W),\ \mathrm{j}_0 = j_0,\ \mathrm{j}_1 = j_1 \right\}, \tag{10}$$

where $C^1_{\mathrm{pw}}([0,1], \mathcal{J}_W)$ is the set of piecewise C^1 curves in $\mathcal{J}_W$ which are straightforwardly defined from the definition of C^1 curves. This definition is the usual definition for finite dimensional Riemannian manifolds. There is, however, a measurability issue, since it is not obvious from our definition of a measurable path in $T\mathcal{J}_W$ that $t \mapsto |\gamma_t|_{\pi(\gamma_t)}$ is measurable. This issue is addressed in Proposition 2, the proof of which is provided in Appendix A.

PROPOSITION 2. *Let $\gamma : [0,1] \to T\mathcal{J}_W$ be a measurable path in $T\mathcal{J}_W$. Then, $\overline{p} \circ \gamma$ is a measurable path in W and $|\gamma|_{\pi \circ \gamma}$ is a measurable real-valued function.*

4. Groups of diffeomorphisms. Curves in W naturally generate diffeomorphisms on Ω by integration of their first component, which is a time-dependent vector field on Ω which vanishes at $\partial\Omega$. The relations between the Hilbert structure on $\mathcal{B}$ and the class of diffeomorphisms which can be generated in that way have been investigated, in particular, in [30] and [16], in which sufficient smoothness conditions on the vector field are derived to ensure existence, uniqueness, and smoothness of the flow for all time.

For $T > 0$, define the set $L^1([0,T], \mathcal{B})$ as the Banach space of measurable functions $\mathrm{v} : [0,T] \to \mathcal{B}$ such that

$$|\mathrm{v}|_{1,T} \triangleq \int_0^T |\mathrm{v}|_{\mathcal{B}}\, dt < \infty.$$

Similarly, $L^2([0,T], \mathcal{B})$ denotes the Hilbert space of square integrable functions defined on $[0,T]$ and taking values in $\mathcal{B}$, with the norm

$$|\mathrm{v}|_{2,T} \triangleq \left(\int_0^T |\mathrm{v}|^2_{\mathcal{B}}\, dt \right)^{1/2}.$$

For $\mathrm{v} \in L^1([0,T], \mathcal{B})$, consider the ODE

$$\frac{dy}{dt} = \mathrm{v}_t(y). \tag{11}$$

A global flow solution of this equation is a time-dependent family of functions $t \to \varphi_t$ such that, for all $x \in \Omega$, $\varphi_0(x) = x$ and

$$\varphi_t = \int_0^t \mathrm{v}_s \circ \varphi_s\, ds.$$

When the dependence of this flow on v must be emphasized, it is denoted by φ^{v}.

Results in [30, 16] essentially relate the existence and smoothness of such flows to embedding conditions of $\mathcal{B}$ into standard sets of continuous functions. We quote these results in the following theorem.

THEOREM 1 (Trouvé). *If $\mathcal{B}$ is continuously embedded in $C^1_0(\Omega, \mathbb{R}^k)$, then for all $T > 0$ and all $\mathrm{v} \in L^1([0,T], \mathcal{B})$, the ODE (11) can be integrated over $[0,T]$, and its associated flow φ^{v} is such that at all times $x \to \varphi^{\mathrm{v}}_t$ is a homeomorphism of Ω.*

NOTATION 1. *Assume that $\mathcal{B}$ is continuously embedded in $C_0^1(\Omega, \mathbb{R}^k)$, and introduce the map*

$$\begin{array}{rrcl} \mathbf{A}_T: & L^1([0,T],\mathcal{B}) & \to & C(\overline{\Omega},\mathbb{R}^k), \\ & \mathrm{v} & \mapsto & \varphi_T^{\mathrm{v}}. \end{array}$$

Then, the set $\mathbf{A}_1(L^1([0,1],\mathcal{B}))$ will be denoted $G_{\mathcal{B}}$.

The fact that $G_{\mathcal{B}}$ is a group is proved in [30]. Further results on these groups and on A_T can be found in Appendix C.

The relation between algebraic and metric properties of groups of diffeomorphisms and some of the fundamental equations of fluid mechanics has been the subject of several studies, starting with [5], in which the Euler equation is related to the geodesic equations of groups of diffeomorphisms with an L^2 metric on its Lie algebra (see also [3, 4, 24]). Another important equation, the Camassa–Holm equation, which describes the motion of the waves in shallow water, can be interpreted along the same lines with an H_α^1 metric on the Lie algebra [11, 17]. Here, since the energy derives from both geometric and photometric variations, the geodesic equations that we derive can be formally interpreted as conservation of momentum on a semidirect product of the group of diffeomorphisms and the space of images, as studied in [25]. However, our point of view of smooth deformations acting on nonsmooth images requires a specific approach. This is also related to developments in optimal design [28].

5. Geodesics on $\mathcal{J}$.

5.1. Minimizing geodesics. The space of C^1 curves is not well suited to deal with proofs of the existence of curves of minimal length between two images j_0 and j_1, i.e., minimizing geodesics. We introduce below the more tractable space of curves with square integrable speed.

We need first a preliminary proposition saying that square integrable paths in $T\mathcal{J}_W$ are uniquely identified by their trace on smooth space-time vector fields in $\mathbb{R}^d$. The proof of this proposition is postponed to Appendix A.

PROPOSITION 3. *Let $\gamma : [0,1] \to T\mathcal{J}_W$ be a measurable path in $T\mathcal{J}_W$. Then, if $\int_0^1 |\gamma_t|^2_{\pi(\gamma_t)} dt < +\infty$ and, for any $\mathrm{u} \in C_c^\infty(\Omega \times]0,1[, \mathbb{R}^d)$, we have $\int_0^1 \langle \gamma_t\, , \mathrm{u}_t \rangle dt = 0$, then $\gamma = 0$ a.e.*

We can now introduce the space $H^1([0,1], \mathcal{J}_W)$ of regular curves.

DEFINITION 6. *We say that a path $\mathrm{j} \in C([0,1], L^2(\Omega, \mathbb{R}^d))$ is regular if there exists a measurable path $\gamma : [0,1] \to T\mathcal{J}_W$ such that $\pi(\gamma) = \mathrm{j}$, $\int_0^1 |\gamma_t|^2 dt < \infty$, and, for any $\mathrm{u} \in C_c^\infty(]0,1[\times\Omega, \mathbb{R}^d)$, we have $-\int_0^1 \langle \mathrm{j}_t, \frac{\partial \mathrm{u}}{\partial t} \rangle_2 dt = \int_0^1 \langle \gamma_t\, , \mathrm{u}_t \rangle dt$. From Proposition 3, the path γ is uniquely defined; using the notation $\frac{\partial \mathrm{j}}{\partial t} \triangleq \gamma_t$, we get the integration by parts formula*

$$\int_0^1 \left\langle \mathrm{j}_t, \frac{\partial \mathrm{u}}{\partial t} \right\rangle_2 dt = -\int_0^1 \left\langle \frac{\partial \mathrm{j}}{\partial t}\, , \mathrm{u}_t \right\rangle dt. \tag{12}$$

We denote $H^1([0,1], \mathcal{J}_W)$ as the set of all the regular paths in $C([0,1], L^2(\Omega, \mathbb{R}^d))$.

PROPOSITION 4. *We have $C^1([0,1], \mathcal{J}_W) \subset H^1([0,1], \mathcal{J}_W)$ and both definitions of $\frac{\partial \mathrm{j}}{\partial t}$ coincide.*

Proof. Let $\mathrm{j} \in C^1([0,1], \mathcal{J}_W)$. There exists $\mathrm{w} = (\mathrm{v}, \mathrm{z}) \in C([0,1], W)$ such that for any $u \in C_c^\infty(\Omega, \mathbb{R}^d)$, $t \to \langle \mathrm{j}_t, u \rangle_2$ is C^1 and

$$\frac{\partial}{\partial t} \langle \mathrm{j}_t, \mathrm{u} \rangle_2 = \sigma^2 \langle \mathrm{z}_t, \mathrm{u} \rangle_2 + \langle \mathrm{j}_t, \mathrm{div}(\mathrm{u} \otimes \mathrm{v}_t) \rangle_2.$$

Certainly, $\mathrm{w} \in L^2([0,1], W)$. Moreover, for any $\mathrm{u} \in C_c^\infty(\Omega, \mathbb{R}^d)$ and any $f \in C_c^\infty(]0,1[, \mathbb{R})$, we have by integration by parts (we denote $f'(t) \triangleq \frac{df}{dt}$)

$$\int_0^1 \langle j_t, f'(t)\mathrm{u}\rangle_2 = \int_0^1 f'(t)\langle \mathrm{j}_t, \mathrm{u}\rangle_2 dt = -\int_0^1 f(t)\frac{d}{dt}\langle \mathrm{j}_t, \mathrm{u}\rangle_2 dt$$

so that (12) is true for $u \otimes f \in C_c^\infty(]0,1[\times\Omega, \mathbb{R}^d)$. The complete proof follows by usual density arguments. □

We carry on with an important result which characterizes regular paths in $\mathcal{J}_W$. For a path v in $L^1([0,1], \mathcal{B})$, we define for any $s, t \in [0,1]$

$$\varphi_{t,s}^{\mathrm{v}} \triangleq \varphi_s^{\mathrm{v}} \circ (\varphi_t^{\mathrm{v}})^{-1}.$$

THEOREM 2. *A path* $\mathrm{j} : [0,1] \to \mathcal{J}_W$ *is regular (resp., is in* $C^1([0,1], \mathcal{J}_W)$*) if and only if there exists* $\mathrm{w} = (\mathrm{v}, \mathrm{z}) \in L^2([0,1], W)$ *(resp.,* $\in C([0,1], W)$*) such that*

$$\mathrm{j}_t = \mathrm{j}_0 \circ \varphi_{t,0}^{\mathrm{v}} + \sigma^2 \int_0^t \mathrm{z}_s \circ \varphi_{t,s}^{\mathrm{v}} ds.$$

Proof. The proof is postponed to Appendix B. □

THEOREM 3. *Let* j_0 *and* j_1 *be in* $\mathcal{J}_W$. *Then we have*

$$d_{\mathcal{J}_W}(j_0, j_1) = \inf\left\{ \int_0^1 \left|\frac{\partial \mathrm{j}}{\partial t}\right|_{j_t} dt \mid \mathrm{j} \in H^1([0,1], \mathcal{J}_W),\ \mathrm{j}_0 = j_0,\ \mathrm{j}_1 = j_1 \right\}. \tag{13}$$

Proof. Let $\mathrm{j} \in H^1([0,1], \mathcal{J}_W)$ be a regular path from j_0 to j_1 and let $\mathrm{w} \in L^2([0,1], W)$ such that $\mathrm{w}_t = \overline{p}_{\mathrm{j}_t}(\frac{\partial \mathrm{j}}{\partial t})$ for any t. There exists a sequence $(\mathrm{w}^n = (\mathrm{v}^n, \mathrm{z}^n) \in C([0,1], W),\ n \in \mathbb{N})$ such that $\int_0^1 |\mathrm{w}_t - \mathrm{w}_t^n|_W^2 dt \to 0$. Define

$$\mathrm{j}_t^n = j_0 \circ \varphi_{t,0}^{\mathrm{v}^n} + \sigma^2 \int_0^t \mathrm{z}_s^n \circ \varphi_{t,s}^{\mathrm{v}^n} ds.$$

We get from Theorem 2 that $\mathrm{j}^n \in C^1([0,1], \mathcal{J}_W)$. Now, considering $\tilde{\mathrm{w}}^n \triangleq (\tilde{\mathrm{v}}^n, \tilde{\mathrm{z}}^n)$ with $\tilde{\mathrm{z}}_t^n \triangleq \mathrm{z}_t^n + (j_1 - \mathrm{j}_1^n) \circ \varphi_{s,1}^{\mathrm{v}^n}$ and $\tilde{\mathrm{v}}^n \triangleq \mathrm{v}^n$ we get from Theorem 9 (see Appendix C) that $\tilde{\mathrm{w}}^n \in C([0,1], W)$. Using Theorem 2, we deduce that if $\tilde{\mathrm{j}}^n$ is defined by

$$\tilde{\mathrm{j}}_t^n = j_0 \circ \varphi_{t,0}^{\mathrm{v}^n} + \sigma^2 \int_0^t \tilde{\mathrm{z}}_s^n \circ \varphi_{t,s}^{\mathrm{v}^n} ds,$$

then $\tilde{\mathrm{j}}^n \in C^1([0,1], \mathcal{J}_W)$ and $\tilde{\mathrm{j}}_1^n = j_1$. However,

$$\int_0^1 \left|\frac{\partial \tilde{\mathrm{j}}^n}{\partial t}\right|_{\tilde{\mathrm{j}}_t^n} dt \le \int_0^1 |\tilde{\mathrm{w}}_t^n|_W dt \to \int_0^1 |\mathrm{w}_t|_W dt$$

when $n \to \infty$. Therefore, we deduce that $d_{\mathcal{J}_W}(j_0, j_1) \le \int_0^1 |\frac{\partial \mathrm{j}}{\partial t}|_{\mathrm{j}_t} dt$ for any regular path from j_0 to j_1. Finally, since $C^1([0,1], \mathcal{J}_W) \subset H^1([0,1], \mathcal{J}_W)$, we get the result. □

DEFINITION 7. *Let* $j_0, j_1 \in \mathcal{J}_W$. *We say that* $\mathrm{j} \in C([0,1], L^2(\Omega, \mathbb{R}^d))$ *is a minimizing geodesic path from* j_0 *to* j_1 *if* j *is regular and*

$$\left(\int_0^1 \left|\frac{\partial \mathrm{j}}{\partial t}\right|_{\mathrm{j}_t}^2 dt\right)^{\frac{1}{2}} = d_{\mathcal{J}_W}(j_0, j_1).$$

We denote $G_{\mathcal{J}_W}(j_0, j_1)$ *as the set of the minimizing geodesic paths from* j_0 *to* j_1.

5.2. Characterization of geodesics.

5.2.1. Photometric optimality.

THEOREM 4. *Let $j_0, j_1 \in \mathcal{J}_W$ and $j \in G_{\mathcal{J}_W}(j_0, j_1)$ be a minimizing geodesic path from j_0 to j_1. Let $\mathrm{w} = (\mathrm{v}, \mathrm{z}) \in L^2([0,1], W)$ be defined by $\mathrm{w}_t \triangleq \overline{p}(\frac{\partial j}{\partial t})$ for any $t \in [0,1]$. Then $\mathrm{z} \in C([0,1], L^2(\Omega, \mathbb{R}^d))$ and for any $t \in [0,1]$ we have*

$$\mathrm{z}_t = \mathrm{z}_0 \circ \varphi^{\mathrm{v}}_{t,0} \left| d\varphi^{\mathrm{v}}_{t,0} \right|. \tag{14}$$

Proof. Let $\mathrm{j} \in H^1([0,1], \mathcal{J}_W)$ be a minimizing geodesic from j_0 to j_1, and let $\mathrm{w} = (\mathrm{v}, \mathrm{z}) \in L^2([0,1], W)$ such that for any $t \in [0,1]$, $\mathrm{w}_t = \overline{p}(\frac{d\mathrm{j}}{dt})$. For any $\mathrm{u} \in C_c^\infty(]0,1[\times \Omega, \mathbb{R}^d)$ and any $\varepsilon \in \mathbb{R}$, define

$$\tilde{\mathrm{j}}_t = j_0 \circ \varphi^{\mathrm{v}}_{t,0} + \sigma^2 \int_0^t \left(\mathrm{z}_s + \varepsilon \frac{\partial \mathrm{u}_s}{\partial s} \circ \varphi^{\mathrm{v}}_{s,1} \right) \circ \varphi_{t,s} ds.$$

Since $t \to (\mathrm{v}_t, \mathrm{z}_t + \varepsilon \frac{\partial \mathrm{u}_t}{\partial t} \circ \varphi^{\mathrm{v}}_{t,1}) \in L^2([0,1], W)$, we get from Theorem 2 that $\tilde{\mathrm{j}} \in H^1([0,1], \mathcal{J}_W)$. Moreover, $\tilde{\mathrm{j}}_0 = j_0$ and $\tilde{\mathrm{j}}_1 = j_1$ so that

$$\begin{aligned}\int_0^1 \left| \frac{d\mathrm{j}_t}{dt} \right|^2_{\mathrm{j}_t} dt = \int_0^1 \left(|\mathrm{v}_t|^2_{\mathcal{B}} + \sigma^2 |\mathrm{z}_t|_2^2 \right) dt &\leq \int_0^1 \left| \frac{d\tilde{\mathrm{j}}_t}{dt} \right|^2_{\tilde{\mathrm{j}}_t} dt \\ &\leq \int_0^1 \left(|\mathrm{v}_t|^2_{\mathcal{B}} + \sigma^2 \left| \mathrm{z}_t + \varepsilon \frac{\partial \mathrm{u}_t}{\partial t} \circ \varphi^{\mathrm{v}}_{t,1} \right|_2^2 \right) dt.\end{aligned}$$

Since ε is arbitrary, we get

$$0 = \int_0^1 \left\langle \mathrm{z}_t, \frac{\partial \mathrm{u}_t}{\partial t} \circ \varphi^{\mathrm{v}}_{t,1} \right\rangle_2 dt = \int_0^1 \left\langle \mathrm{z}_t \circ \varphi^{\mathrm{v}}_{1,t} \left| d\varphi^{\mathrm{v}}_{1,t} \right|, \frac{\partial \mathrm{u}_t}{\partial t} \right\rangle_2 dt.$$

Choosing arbitrary $\mathrm{u} \in C_c^\infty(]0,1[\times \Omega, \mathbb{R}^d)$, we get that there exists $\tilde{z}_1 \in L^2(\Omega, \mathbb{R}^d)$ such that t-a.e. we have $\mathrm{z}_t \circ \varphi^{\mathrm{v}}_{1,t} \left| d\varphi^{\mathrm{v}}_{1,t} \right| = \tilde{z}_1$. Hence, if $\tilde{\mathrm{z}}_t = \tilde{z}_1 \circ \varphi^{\mathrm{v}}_{t,1} \left| d\varphi^{\mathrm{v}}_{t,1} \right|$, we have $\tilde{\mathrm{z}} \in C([0,1), L^2([0,1], \mathbb{R}^d))$ and $\mathrm{z}_t = \tilde{\mathrm{z}}_t$ t-a.e. Note that $\tilde{\mathrm{z}}_0 \circ \varphi^{\mathrm{v}}_{1,0} \left| d\varphi^{\mathrm{v}}_{1,0} \right| = \tilde{z}_1$ so that

$$\tilde{\mathrm{z}}_t = \left(\tilde{\mathrm{z}}_0 \circ \varphi^{\mathrm{v}}_{1,0} \left| d\varphi^{\mathrm{v}}_{1,0} \right| \right) \circ \varphi^{\mathrm{v}}_{t,1} \left| d\varphi^{\mathrm{v}}_{t,1} \right| = \tilde{\mathrm{z}}_0 \circ \varphi^{\mathrm{v}}_{t,0} \left| d\varphi^{\mathrm{v}}_{t,0} \right|,$$

and the proof is ended. ∎

This leads to the following definition.

DEFINITION 8. *A regular path $\mathrm{j} \in H^1([0,1], \mathcal{J}_W)$ is called a pregeodesic path if and only if the following equations are satisfied almost everywhere in t:*

$$\begin{cases} \mathrm{j}_t = \mathrm{j}_0 \circ \varphi^{\mathrm{v}}_{t,0} + \sigma^2 \displaystyle\int_0^t \mathrm{z}_s \circ \varphi^{\mathrm{v}}_{t,s} ds, \\ \mathrm{z}_t = \mathrm{z}_0 \circ \varphi^{\mathrm{v}}_{t,0} \left| d\varphi^{\mathrm{v}}_{t,0} \right|, \\ (\mathrm{v}_t, \mathrm{z}_t) = \overline{p} \left(\dfrac{d\mathrm{j}}{dt} \right). \end{cases} \tag{15}$$

5.2.2. Study of the geodesic equation.

Directional derivatives in ^{2}. In this section, we try to clarify the last equation of system (15), at least in some situations of interest. The difficulty comes from the fact that, unless j_t is smooth enough, this equation does not, in general, specify a unique correspondence $\mathrm{z}_t \mapsto \mathrm{v}_t$.

To be more precise, let us analyze the condition that, for all t,

$$(\mathrm{v}_t, \mathrm{z}_t) = \overline{p}\left(\frac{d\mathrm{j}}{dt}\right).$$

For this purpose, we first introduce a weak version of the directional derivative $dj.v$ when $j \in \mathcal{J}_W$ and $v \in \mathcal{B}$.

Definition 9. *Let $j \in \mathcal{J}_W$.*

(1) *We define the operator $Dj : \mathcal{D}_j \to L^2(\Omega, \mathbb{R}^d)$ by*

$$\mathcal{D}_j \triangleq \{\ v \in \mathcal{B} \mid \exists C,\ s.t.\ \forall u \in C_c^\infty(\Omega, \mathbb{R}^d),\ |\langle j\,,\, \mathrm{div}(u \otimes v)\rangle_2| \le C\,|u|_2\ \},$$

and for any $v \in \mathcal{D}_j$, $Dj.v$ is the unique element in $L^2(\Omega, \mathbb{R}^d)$ such that

$$\langle Dj.v\,,\, u\rangle_2 = -\langle j\,,\, \mathrm{div}(u \otimes v)\rangle_2$$

for any $u \in C_c^\infty(\Omega, \mathbb{R}^d)$.

(2) *We define the adjoint operator $Dj^* : \mathcal{D}_j^* \to \mathcal{B}$, where*

$$\mathcal{D}_j^* \triangleq \{\ u \in L^2(\Omega, \mathbb{R}^d) \mid \exists C,\ s.t.\ \forall v \in \mathcal{D}_j\ \ |\langle Dj.v\,,\, u\rangle_2| \le C\,|v|_B\ \},$$

and, for any $u \in \mathcal{D}_j^*$, Dj^* is the unique element in $\overline{\mathcal{D}_j}$ (closure of $\mathcal{D}_j$) such that

$$\langle Dj^*.u\,,\, v\rangle_{\mathcal{B}} = \langle u\,,\, Dj.v\rangle_2 \tag{16}$$

for any $v \in \mathcal{D}_j$.

Remark 3. The existence of $Dj.v$ comes from the extension of the linear form $u \to \langle j\,,\, \mathrm{div}(u \otimes v)\rangle$ for smooth u into a continuous linear form on $L^2(\Omega, \mathbb{R}^d)$ for $v \in \mathcal{D}_j$. For the definition of the adjoint Dj^*, the adjoint is uniquely defined as an element of $\overline{\mathcal{D}_j}$ by (16) ($\mathcal{D}_j$ is not necessarily dense in $\mathcal{B}$).

Fix $j \in \mathcal{J}_W$. We may characterize elements $v \in \mathcal{D}_j$ as follows. (We denote hereafter $\varphi^v_{\cdot}$ as the flow associated with the constant speed $v_t \equiv v$ for any $t \in [0, 1]$.)

Theorem 5. *The vector field $v \in \mathcal{B}$ belongs to $\mathcal{D}_j$ if and only if there exists a square integrable function $\xi : \Omega \to \mathbb{R}^d$ such that*

$$j \circ \varphi^v_{0,t}(x) = j(x)\left|d_x\varphi^v_{0,t}\right|^{-1} + \int_0^t \xi \circ \varphi^v_{0,s}(x)\left|d_{\varphi^v_{0,s}(x)}\varphi^v_{s,t}\right|^{-1} ds. \tag{17}$$

We have in such a case $Djv = \xi - j\mathrm{div}(v)$.

Proof. We first notice that, if $v \in \mathcal{B}$,

$$-\langle j\,,\, \mathrm{div}(u \otimes v)\rangle_2 = \frac{d}{d\varepsilon}\int_\Omega \left\langle j(x)\,,\, u \circ \varphi^v_{\varepsilon,0}(x)\right\rangle \left|d_x\varphi^v_{\varepsilon,0}\right| dx = \frac{d}{d\varepsilon}\int_\Omega \left\langle j \circ \varphi^v_{0,\varepsilon}(x)\,,\, u(x)\right\rangle dx.$$

Assuming that (17) holds, the last expression yields

$$\begin{aligned}\frac{d}{d\varepsilon}\int_\Omega \langle j(x)\,,\, u(x)\rangle \left|d_x\varphi^v_{0,\varepsilon}\right|^{-1} dx + \frac{d}{d\varepsilon}\int_\Omega\int_0^\varepsilon \left\langle \xi \circ \varphi^v_{0,s}\,,\, u(x)\right\rangle \left|d_{\varphi^v_{0,s}(x)}\varphi^v_{s,\varepsilon}\right|^{-1} dx \\ = -\int_\Omega \langle j(x)\,,\, u(x)\rangle \mathrm{div}(v) dx + \int_\Omega \langle \xi(x)\,,\, u(x)\rangle dx = \langle \xi - j\mathrm{div}(v)\,,\, u\rangle_2,\end{aligned}$$

which implies that $v \in \mathcal{D}_j$ and $Djv = \xi - j\mathrm{div}(v)$.

Conversely, let $v \in \mathcal{D}_j$ and $\xi = Djv + j\mathrm{div}(v)$. Fix $u \in C^1(\Omega, \mathbb{R}^d)$. Consider the function f, defined on $[0,1]$ by $f(t) = \left\langle j \circ \varphi^v_{0,t}\, ,\, u\right\rangle_2$. Denote by $\tilde{\jmath}(t)$ the left-hand term of (17), and $g(t) = \langle \tilde{\jmath}_t\, ,\, u\rangle_2$. We have

$$\begin{aligned} g'(t) &= -\left\langle j\, ,\, u\mathrm{div}_{\varphi^v_{0,t}(x)} v \left| d_x \varphi^v_{0,t}\right|^{-1}\right\rangle_2 + \left\langle \xi \circ \varphi^v_{0,t}\, ,\, u\right\rangle_2 \\ &\qquad - \int_0^t \left\langle \xi \circ \varphi^v_{0,s} \left| d_{\varphi^v_{0,s}} \varphi^v_{s,t}\right|^{-1} ,\, u\mathrm{div}_{\varphi^v_{0,t}} v\right\rangle_2 ds \\ &= \left\langle \xi \circ \varphi^v_{0,t} - \tilde{\jmath}_t \mathrm{div}_{\varphi^v_{0,t}} v\, ,\, u\right\rangle_2. \end{aligned}$$

Since $f(t+\varepsilon) = \left\langle j \circ \varphi^v_{t,t+\varepsilon}\, ,\, u \circ \varphi^v_{t,0} \left| d\varphi^v_{t,0}\right|\right\rangle_2$, we have

$$f'(t) = \left\langle Djv\, ,\, u \circ \varphi^v_{t,0} \left| d\varphi^v_{t,0}\right|\right\rangle_2 = \left\langle (Djv) \circ \varphi^v_{0,t}\, ,\, u\right\rangle_2.$$

Therefore, computing the integral of the difference and using the definition of ξ,

$$\left\langle j \circ \varphi^v_{0,t} - \tilde{\jmath}_t\, ,\, u\right\rangle_2 = \int_0^t \left\langle j \circ \varphi^v_{0,s} - \tilde{\jmath}_s\, ,\, u\mathrm{div}_{\varphi^v_{0,t}} v\right\rangle_2 ds \le |u|_2\, |v|_{\mathcal{B}} \int_0^t \left| j \circ \varphi^v_{0,s} - \tilde{\jmath}_s\right|_2 ds.$$

Taking the supremum of the left-hand term over continuously differentiable u with L^2-norm equal to 1 yields

$$\left| j \circ \varphi^v_{0,t} - \tilde{\jmath}_t\right|_2 \le |v|_{\mathcal{B}} \int_0^t \left| j \circ \varphi^v_{0,s} - \tilde{\jmath}_s\right|_2 ds,$$

which implies $\left| j \circ \varphi^v_{0,t} - \tilde{\jmath}_t\right|_2 = 0$ for all t. ◻

An interesting consequence of this is the following lemma.

LEMMA 1. *For any $j \in L^2(\Omega, \mathbb{R}^d)$, one has $j \in D^*_j$ and, for $v \in D_j$,*

$$\langle Djv\, ,\, j\rangle_2 = -\frac{1}{2}\left\langle |j|^2\, ,\, \mathrm{div} v\right\rangle_2.$$

Proof. Indeed, let $v \in D_j$. Consider the function

$$f(t) = \int_\Omega \left| j \circ \varphi^v_{0,t}(x)\right|^2 dx.$$

Since $f(t) = \left\langle |j|^2\, ,\, |d\varphi_{t,0}|\right\rangle_2$, we have $f'(0) = -\langle |j|^2\, ,\, \mathrm{div} v\rangle_2$. Using, on the other hand, expression (17) yields $f'(0) = 2\langle Djv\, ,\, j\rangle_2$. ◻

Interpretation of the pregeodesic equations. The property that $w = (v,z) \in W$ belongs to E_j, which states that, for all $u \in C^\infty_c(\Omega, \mathbb{R}^d)$,

$$\sigma^2 \langle z, u\rangle_2 + \langle j, \mathrm{div}(u \otimes v)\rangle_2 = 0$$

is equivalent to $v \in \mathcal{D}_j$ and $\sigma^2 z - Dj.v = 0$. Consider now some tangent vector $\gamma \in T_j \mathcal{J}_W$, and study the property that, for some $w = (v,z) \in W$, one has $\overline{p}(\gamma) = w$. This implies, in particular, that, for all $(v', z') \in E_j$,

$$|v + v'|^2_{\mathcal{B}} + \sigma^2\, |z + z'|^2_2 \ge |v|^2_{\mathcal{B}} + \sigma^2\, |z|^2_2\,,$$

which is in turn equivalent to the following: for all $v' \in \mathcal{D}_j$, $\langle v\,,\,v'\rangle_{\mathcal{B}} + \langle z\,,\,Dj.v'\rangle_2 = 0$. This implies that $z \in \mathcal{D}_j^*$ and that $\langle v\,,\,v'\rangle_{\mathcal{B}} + \langle Dj^* z\,,\,v'\rangle_{\mathcal{B}} = 0$, so that $v = -Dj^* z + \gamma_\perp$, where $\gamma_\perp$ is the projection of v onto $\mathcal{D}_j^\perp$. Note that this orthogonal component does not depend on the choice of (v,z) from the equivalence class defining γ (hence the notation). We thus may conclude that $(v,z) = \overline{p}(\gamma)$ if and only if $z \in \mathcal{D}_j^*$ and

$$v = -Dj^* z + \gamma_\perp.$$

The first conclusion we may draw from this is that, whenever $\mathcal{D}_j$ is dense in $\mathcal{B}$, v_t is uniquely determined by z_t and the condition $(\mathrm{v}_t, \mathrm{z}_t) = \overline{p}(\frac{d\mathrm{j}}{dt})$. It is given by $\mathrm{v}_t = -D\mathrm{j}_t^* \mathrm{z}_t$. This is true, for example, when $\mathrm{j}_t \in H^1(\Omega, \mathbb{R}^d)$ at all times, since, in this case $\mathcal{D}_{\mathrm{j}_t} = \mathcal{B}$ (notice that, by Theorem 4, this is true along a geodesic as soon as j_0 and j_1 belong to $H^1(\Omega, \mathbb{R}^d)$).

However, this is not the general situation. As an example, consider the case when j is the indicator function of a subdomain Ω_1 of Ω with smooth boundary. If v is a vector field on Ω and u is a smooth function on Ω, we have

$$\langle j\,,\,\mathrm{div} uv\rangle = \int_{\partial\Omega_1} u\langle v\,,\,\nu_1\rangle_{\mathbb{R}^k} d\sigma_1,$$

where ν_1 is the outward normal to $\partial\Omega_1$ and σ_1 is the surface measure on $\partial\Omega_1$. This implies that djv may be identified to a singular measure supported to $\partial\Omega_1$, which does not belong to L^2 unless it vanishes. Thus, $\mathcal{D}_j$ consists exactly of vector fields on Ω which belong to $\mathcal{B}$ and have vanishing normal components on $\partial\Omega_1$. For $a \in \mathbb{R}^k$ and $x \in \mathbb{R}$, denote by $K_x a$ the element of $\mathcal{B}$ such that $\langle K_x a\,,\,u\rangle_{\mathcal{B}} = \langle u(x)\,,\,a\rangle_{\mathbb{R}^k}$. Then, $K_x \nu(x)$ belongs to $D_j^\perp$ for any $x \in \partial\Omega$, and so does any linear combination of these vector fields. We see that in this case $D_j^\perp$ is nontrivial.

This discussion implies that the pregeodesic condition for a path may be written

$$\begin{cases} \mathrm{j}_t = \mathrm{j}_0 \circ \varphi^{\mathrm{v}}_{t,0} + \sigma^2 \displaystyle\int_0^t \mathrm{z}_s \circ \varphi^{\mathrm{v}}_{t,s} ds, \\[2ex] \mathrm{z}_t = \mathrm{z}_0 \circ \varphi^{\mathrm{v}}_{t,0} \left| d\varphi^{\mathrm{v}}_{t,0} \right|, \\[2ex] \mathrm{z}_t \in D_j^*, \ \text{ and } \mathrm{v}_t - D\mathrm{j}_t^* \mathrm{z}_t \in D_{j_t}^\perp. \end{cases} \tag{18}$$

These equations are not complete yet (in the sense that they cannot be solved from the initial values (j_0, v_0, z_0)) since they provide no information on the choice of $\mathrm{v}_t - D\mathrm{j}_t^* \mathrm{z}_t$ at time t (unless of course $D_{j_t}^\perp = \{0\}$). We need to specify the mode of propagation of this singular component along a geodesic. The following computation provides a hint on possible ways to achieve this. Assume that j is pregeodesic and $(\mathrm{v}_t, \mathrm{z}_t) = \overline{p}(\frac{d\mathrm{j}}{dt})$. In such a case, we have

$$|\mathrm{z}_s|_2^2 = \int_\Omega |z_0 \circ \varphi^{\mathrm{v}}_{s,0}|^2 |d\varphi^{\mathrm{v}}_{s,0}|^2 dx = \int_\Omega |z_0|^2 |d\varphi^{\mathrm{v}}_{0,s}|^{-1} dy$$

and

$$z_0(x) = \frac{1}{\sigma^2}(j_1 \circ \varphi^{\mathrm{v}}_{0,1}(x) - j_0(x)) \int_0^1 |d\varphi^{\mathrm{v}}_{0,s}|^{-1} ds.$$

Thus

$$\int_0^1 |\mathrm{z}_s|_2^2\, ds = \frac{1}{\sigma^4} \int_\Omega \frac{\left|j_1 \circ \varphi^{\mathrm{v}}_{0,1} - j_0\right|^2}{\int_0^1 |d\varphi^{\mathrm{v}}_{0,s}|^{-1} ds}.$$

Making the change of variables $y = \varphi^{\mathrm{v}}_{0,1}(x)$ yields

$$\int_0^1 |z_s|_2^2\, ds = \frac{1}{\sigma^4} \int_\Omega \frac{\left|j_1 - j_0 \circ \varphi^{\mathrm{v}}_{1,0}\right|^2}{\int_0^1 |d_{\varphi^{\mathrm{v}}_{1,0}(x)} \varphi^{\mathrm{v}}_{0,s}|^{-1} |d\varphi^{\mathrm{v}}_{1,0}(x)|^{-1} ds},$$

i.e.,

$$\int_0^1 |\mathrm{z}_s|_2^2\, ds = \frac{1}{\sigma^4} \int_\Omega \frac{\left|j_1 - j_0 \circ \varphi^{\mathrm{v}}_{1,0}\right|^2}{\int_0^1 |d\varphi^{\mathrm{v}}_{1s}|^{-1} ds},$$

and the geodesic energy is given by

$$\int_0^1 |\mathrm{v}_s|_{\mathcal{B}}^2\, ds + \frac{1}{\sigma^2} \int_\Omega \frac{\left|j_1 - j_0 \circ \varphi^{\mathrm{v}}_{1,0}\right|^2}{\int_0^1 |d\varphi^{\mathrm{v}}_{1,s}|^{-1} ds}. \tag{19}$$

We can obtain more precise information on the geodesic by studying variations of this expression with respect to v. This will be handled below, under a smoothness assumption on j_0. Before this, we need some notation for the reproducing kernel on $\mathcal{B}$. They will be useful throughout the paper.

Kernels for the inner-product on $\mathcal{B}$.

PROPOSITION 5. *There exists a continuous operator K (resp., K_∇) on $L^1(\Omega, \mathbb{R}^k)$ (resp., $L^1(\Omega, \mathbb{R})$) with values in $\mathcal{B}$ such that, for all $u \in L^1(\Omega, \mathbb{R}^k)$ (resp., $u \in L^1(\Omega, \mathbb{R})$), for all $v \in \mathcal{B}$,*

$$\langle Ku\, ,\, v\rangle_{\mathcal{B}} = \langle u, v\rangle_2,$$

and

$$\langle K_\nabla u\, ,\, v\rangle_{\mathcal{B}} = -\langle u, \mathrm{div} v\rangle_2.$$

Proof of Proposition 5. Let $u \in L^1(\Omega, \mathbb{R}^k)$. Since we assume that $\mathcal{B}$ is continuously embedded in C_0^1, the linear form defined on $\mathcal{B}$ by $v \mapsto \langle u, v\rangle_2$ is continuous because $|\langle u, v\rangle_2| \le |u|_1\, |v|_\infty$. Therefore, there exists a unique element in $\mathcal{B}$, denoted Ku, such that, for all $v \in \mathcal{B}$, $\langle Ku\, ,\, v\rangle_{\mathcal{B}} = \langle u, v\rangle_2$ and continuity comes from the inequality $\langle Ku\, ,\, v\rangle_{\mathcal{B}} \le |u|_1\, |v|_\infty \le \mathrm{cst}\ |u|_1\, |v|_{\mathcal{B}}$.

The same proof holds for K_∇, since $|\mathrm{div} v|_\infty$ is also controlled by $|v|_{\mathcal{B}}$. ∎

It can be remarked that, for smooth u, $K_\nabla u = K(\nabla u)$.

Remark 4. When j is smooth (e.g., $j \in H^1(\Omega, \mathbb{R}^d)$), the operator Dj^* introduced in the previous paragraph is given by $Dj^* z = K(dj^*.z)$, in which dj^* is the standard matrix adjoint of dj. Indeed, we have in this case

$$\langle z\, ,\, Dj.v\rangle_2 = \langle z\, ,\, dj.v\rangle_2 = \langle dj^*.z\, ,\, v\rangle_2 = \langle K(dj^*.z)\, ,\, v\rangle_{\mathcal{B}}.$$

Characterization with a smooth endpoint. We study the effect of small variations in v on the geodesic energy (19), under the additional hypothesis that $j_0 \in H^1(\Omega, \mathbb{R}^d)$. Thus, fix $h \in L^2([0,1], \mathcal{B})$, and consider a perturbation $\mathrm{v} + \varepsilon h$ of v. We compute the corresponding variation of the geodesic energy. The variation of the first term being $2\int_0^1 \langle \mathrm{v}_t \, , \, h_t \rangle_{\mathcal{B}} dt$, we can focus on the second term, namely,

$$U^\varepsilon \triangleq \frac{1}{\sigma^2} \int_\Omega \frac{\left| j_0 \circ \varphi_{1,0}^{\mathrm{v}+\varepsilon h} - j_1 \right|^2}{\int_0^1 |d\varphi_{1s}^{\mathrm{v}+\varepsilon h}|^{-1} ds} dx.$$

The variations of U^ε are given in Lemma 2, which is proved in Appendix E.

LEMMA 2. *We have, at $\varepsilon = 0$,*

$$\frac{dU^\varepsilon}{d\varepsilon} = \sigma^2 \int_0^1 \left\langle \Big(K_\nabla (q_t^{\mathrm{v}} \left|\mathrm{z}_t\right|^2) + K(\left|\mathrm{z}_t\right|^2 \nabla q_t^{\mathrm{v}}) \Big) + 2K([d\varphi_{t,0}^{\mathrm{v}}]^* d_{\varphi_{t,0}^{\mathrm{v}}} j_0^* \mathrm{z}_t), h_t \right\rangle_{\mathcal{B}} dt, \tag{20}$$

with $q_t^{\mathrm{v}} = \int_0^t |d\varphi_{t,s}^{\mathrm{v}}|^{-1} ds$.

We can deduce from this our additional condition for a regular path to be a minimizing geodesic: for almost all $t \in [0,1]$,

$$\mathrm{v}_t + \frac{1}{2} (K D_t^{\mathrm{v}} + K_\nabla C_t^{\mathrm{v}})_{\mathcal{B}} = 0,$$

where

$$D_t^{\mathrm{v}} \triangleq \sigma^2 \left|\mathrm{z}_t\right|^2 \nabla q_t^{\mathrm{v}} + 2[d\varphi_{t,0}^{\mathrm{v}}]^* d_{\varphi_{t,0}^{\mathrm{v}}} j_0^* \mathrm{z}_t,$$

and

$$C_t^{\mathrm{v}} \triangleq \sigma^2 q_t^{\mathrm{v}} \left|\mathrm{z}_t\right|^2.$$

It may be interesting to check that this condition boils down to the one we have obtained before for smooth trajectories, namely,

$$\mathrm{v}_t + K(dj_t^* \mathrm{z}_t) = 0.$$

It suffices to notice that, for pregeodesic trajectories, $j_t = j_0 \circ \varphi_{t,0}^{\mathrm{v}} + \sigma^2 \mathrm{z}_t q_t^{\mathrm{v}}$ and that, when z_t is smooth,

$$K D_t^{\mathrm{v}} + K_\nabla C_t^{\mathrm{v}} = K(D_t^{\mathrm{v}} + \nabla C_t^{\mathrm{v}}).$$

We now define geodesic paths (not necessarily minimizing).

DEFINITION 10. *Let $j_0 \in H^1(\Omega, \mathbb{R}^d)$. A regular path $\mathrm{j} \in H^1([0,1], \mathcal{J}_W)$ starting at j_0 is called a geodesic path if and only if there exists $\mathrm{w} = (\mathrm{v}, \mathrm{z}) \in L^2([0,1], W)$ such that the following equations are satisfied almost everywhere in t:*

$$\begin{cases} \mathrm{j}_t = \mathrm{j}_0 \circ \varphi_{t,0}^{\mathrm{v}} + \sigma^2 \displaystyle\int_0^t \mathrm{z}_s \circ \varphi_{t,s}^{\mathrm{v}} ds, \\ \mathrm{z}_t = \mathrm{z}_0 \circ \varphi_{t,0}^{\mathrm{v}} \left| d\varphi_{t,0}^{\mathrm{v}} \right|, \\ \mathrm{v}_t + K([d\varphi_{t,0}^{\mathrm{v}}]^* d_{\varphi_{t,0}^{\mathrm{v}}} j_0^* \mathrm{z}_t) + \dfrac{\sigma^2}{2} \Big(K_\nabla (q_t^{\mathrm{v}} \left|\mathrm{z}_t\right|^2) + K(\left|\mathrm{z}_t\right|^2 \nabla q_t^{\mathrm{v}}) \Big) = 0, \end{cases} \tag{21}$$

with $q^{\mathrm{v}}_t = \int_0^t |d\varphi^{\mathrm{v}}_{t,s}|^{-1} ds$.

These equations are complete; it will be shown in section 7 that initial conditions (j_0, z_0) uniquely specify the solutions. It is interesting to check that geodesics as defined in (21) also are pregeodesics. For this, we first show that, for all times t, $z_t \in \mathcal{D}^*_{j_t}$. Noting that the first equation in (21) may also be written

$$\mathrm{j}_t = \mathrm{j}_0 \circ \varphi^{\mathrm{v}}_{t,0} + \sigma^2 \mathrm{z}_t q_t$$

it is clear that $\mathcal{D}_{j_t} = D_{z_t}$, and $z_t \in D^*_{z_t}$ is proved in Lemma 1. The same lemma also provides the fact that, for $w \in \mathcal{D}_{z_t}$,

$$\langle z_t\,,\, Dj_t w\rangle_2 = \langle z_t\,,\, d(\mathrm{j}_0 \circ \varphi^{\mathrm{v}}_{t,0})w\rangle_2 + \sigma^2 \Big\langle |z_t|^2 \,\nabla q_t\,,\, w\Big\rangle_2 - \frac{\sigma^2}{2}\Big\langle |z_t|^2\,,\, \mathrm{div}(q^{\mathrm{v}}_t w)\Big\rangle,$$

and this is equal to $-\langle \mathrm{v}_t\,,\, w\rangle_{\mathcal{B}}$ by definition of K and K_∇. We thus obtain the fact that $v_t + Dj_t^* z_t \in D^\perp_{j_t}$ as required.

We shall prove existence of solutions for a broader class of evolution equations, extending the range of initial values v_0. Consider the term $\mathrm{u}_t = K([d\varphi^{\mathrm{v}}_{t,0}]^* d_{\varphi^{\mathrm{v}}_{t,0}} j_0^* \mathrm{z}_0 \circ \varphi^{\mathrm{v}}_{t,0}\,|d\varphi^{\mathrm{v}}_{t,0}|)$ which appears in the third equation of (21). We have, letting $\omega_0 = -dj_0^* z_0$, and, for $w \in \mathcal{B}$,

$$\begin{aligned}\langle \mathrm{u}_t\,,\, w\rangle_{\mathcal{B}} &= \Big\langle [d\varphi^{\mathrm{v}}_{t,0}]^* d_{\varphi^{\mathrm{v}}_{t,0}} j_0 \mathrm{z}_0 \circ \varphi^{\mathrm{v}}_{t,0}\,|d\varphi^{\mathrm{v}}_{t,0}|\,,\, w\Big\rangle_{L^2}\\ &= \Big\langle d_{\varphi^{\mathrm{v}}_{t,0}} j_0 \mathrm{z}_0 \circ \varphi^{\mathrm{v}}_{t,0}\,|d\varphi^{\mathrm{v}}_{t,0}|\,,\, d\varphi^{\mathrm{v}}_{t,0} w\Big\rangle_{L^2}\\ &= \langle \omega_0\,,\, (d\varphi^{\mathrm{v}}_{0,t})^{-1} w \circ \varphi^{\mathrm{v}}_{0,t}\rangle_{L^2}.\end{aligned}$$

We know, by Appendix C, that $\varphi^{\mathrm{v}}_{0,t}$ belongs to $C^p(\overline{\Omega})$ as soon as $\mathcal{B}$ is continuously embedded in $C^p_0(\Omega, \mathbb{R}^k)$, which implies in this case (with $p \geq 1$) that

$$\big|(d\varphi^{\mathrm{v}}_{0,t})^{-1} w \circ \varphi^{\mathrm{v}}_{0,t}\big|_{p-1,\infty} \leq \mathrm{Const}\, |w|_{\mathcal{B}}\,,$$

the constant depending on $|v|_{1,\mathcal{B}}$. But this implies in turn that, if the L^2 inner product is replaced by the action of any continuous functional, ω_0, on $C^{p-1}_0(\Omega, \mathbb{R}^k)$, which will be denoted

$$\big(\omega_0, (d\varphi^{\mathrm{v}}_{0,t})^{-1} w \circ \varphi^{\mathrm{v}}_{0,t}\big),$$

there exists an element of $\mathcal{B}$ that we shall still denote u_t such that

$$\langle u_t\,,\, w\rangle_{\mathcal{B}} = \big(\omega_0, (d\varphi^{\mathrm{v}}_{0,t})^{-1} w \circ \varphi^{\mathrm{v}}_{0,t}\big).$$

With this notation, we may formulate the following definition.

DEFINITION 11. *Let* $j_0 \in L^2(\Omega, \mathbb{R}^d)$. *Let* ω_0 *be a continuous linear functional on* $C^{p-1}(\Omega, \mathbb{R}^k)$ *and* $z_0 \in L^2(\Omega, \mathbb{R}^d)$. *A regular path* $\mathrm{j} \in H^1([0,1], \mathcal{J}_W)$ *starting at* j_0 *with initial direction* (ω_0, z_0) *is called a generalized geodesic if and only if, for all* $u \in \mathcal{D}_{j_0}$, *one has*

$$(\omega_0, u) + \langle z_0\,,\, Dj_0 u\rangle = 0,$$

and there exists $\mathrm{w} = (\mathrm{v}, \mathrm{z}) \in L^2([0,1], W)$ *such that the following equations are satisfied almost everywhere in* t*:*

$$
\begin{cases}
\mathrm{j}_t = \mathrm{j}_0 \circ \varphi^{\mathrm{v}}_{t,0} + \sigma^2 \displaystyle\int_0^t \mathrm{z}_s \circ \varphi^{\mathrm{v}}_{t,s} ds, \\
\mathrm{z}_t = \mathrm{z}_0 \circ \varphi^{\mathrm{v}}_{t,0} \left| d\varphi^{\mathrm{v}}_{t,0} \right|, \\
\mathrm{v}_t - \mathrm{u}^{\mathrm{v}}_t + \dfrac{\sigma^2}{2} \Big(K_\nabla (q^{\mathrm{v}}_t \left|\mathrm{z}_t\right|^2) + K(\left|\mathrm{z}_t\right|^2 \nabla q^{\mathrm{v}}_t) \Big) = 0, \\
\forall w \in \mathcal{B} \ \langle \mathrm{u}^{\mathrm{v}}_t \,,\, w \rangle_{\mathcal{B}} = \big(\omega_0, (d\varphi^{\mathrm{v}}_{0,t})^{-1} w \circ \varphi^{\mathrm{v}}_{0,t} \big)
\end{cases}
\tag{22}
$$

with $q^{\mathrm{v}}_t = \int_0^t |d\varphi^{\mathrm{v}}_{t,s}|^{-1} ds$.

Recall that when j_0 is smooth, the only choice is $\omega_0 = dj_0^* z_0$, and if z_0 is also smooth, the system may be written under the simple form

$$
\begin{cases}
\mathrm{j}_t = \mathrm{j}_0 \circ \varphi^{\mathrm{v}}_{t,0} + \sigma^2 \displaystyle\int_0^t \mathrm{z}_s \circ \varphi^{\mathrm{v}}_{t,s} ds, \\
\mathrm{z}_t = \mathrm{z}_0 \circ \varphi^{\mathrm{v}}_{t,0} \left| d\varphi^{\mathrm{v}}_{t,0} \right|, \\
\mathrm{v}_t + \sigma^2 K(dj_t^* z_t) = 0.
\end{cases}
\tag{23}
$$

As an example of the nonsmooth applications we have in mind, assume that j_0 is a binary, plane image, which is the indicator function of the interior of a connected open subset Ω_1 of Ω with smooth boundary $\partial\Omega_1$. We have seen that any element $w \in \mathcal{D}_{j_0}$ should be tangent to $\partial\Omega_1$ and that in this case $Dj_0 w = 0$ and $\mathcal{D}^*_{j_0} = L^2(\Omega, \mathbb{R})$. We therefore may choose z_0 arbitrarily in L^2, and (ω_0, w) should vanish for $w \in \mathcal{D}_{j_0}$, which is true, for example, when ω_0 is defined by

$$
(\omega_0, w) = \int_{\partial\Omega_1} \langle w \,,\, \nu_1 \rangle d\sigma_1,
$$

where ν_1 is the outward normal to $\partial\Omega_1$ and σ_1 is the surface measure on $\partial\Omega_1$.

6. Existence of minimizing geodesics. The next theorem states that minimizing geodesics always exist between two elements of $\mathcal{J}_W$.

THEOREM 6. *Assume that* $\mathcal{B}$ *is compactly embedded in* $C^1_0(\Omega, \mathbb{R}^d)$*, and let* $j_0, j_1 \in \mathcal{J}_W$*. Then* $G_{\mathcal{J}_W}(j_0, j_1)$ *is nonempty.*

Proof. Let $(\mathrm{j}^n)_{n\in\mathbb{N}}$ be a minimizing family of paths in $H^1([0,1], \mathcal{J}_W)$ from j_0 to j_1; for any $n \in \mathbb{N}$, let $\mathrm{w}^n_t \triangleq \overline{p}(\frac{d\mathrm{j}^n}{dt})$ so that $(\mathrm{w}^n)_{n\in\mathbb{N}}$ is a bounded sequence in $L^2([0,1], W)$. Up to the extraction of a subsequence, we can assume that w^n converges weakly to a w^∞ in $L^2([0,1], W)$. By lower semicontinuity, we have

$$
\int_0^1 |\mathrm{w}^\infty_t|^2_W dt \leq d_{\mathcal{J}_W}(j_0, j_1).
$$

By a time change argument, which is classical in the proof that minimizing geodesics travel at constant speed (see [12]), we may furthermore assume that $|\mathrm{w}^n_t|_W$ is uniformly bounded by, say, $d_{\mathcal{J}_W}(j_0, j_1) + 1$. Denoting $w^n = (\mathrm{v}^n, \mathrm{z}^n)$, consider $\mathrm{j}'_t \triangleq j_0 \circ \varphi^\infty_{t,0} + \sigma^2 \int_0^t \mathrm{z}^\infty_s \circ \varphi^\infty_{t,s} ds$, where φ^∞ is the flow associated to v^∞. Since j'

is a regular path, it is sufficient to prove that $\mathrm{j}_1' = j_1$. However, if φ^n denotes the flow associated with v^n, we know, from Theorem 9, that $\varphi^n_{1,0}$ converges uniformly to $\varphi^\infty_{1,0}$ so that $j_0 \circ \varphi^n_{t,0} \to j_0 \circ \varphi^\infty_{t,0}$ in $L^2(\Omega, \mathbb{R}^d)$. Now, let $u \in C_c^\infty(\Omega, \mathbb{R}^d)$. We have $\int_0^1 \langle \mathrm{z}^n_s \circ \varphi^n_{1,s}, u\rangle_2 ds = \int_0^1 \langle \mathrm{z}^n_s, u \circ \varphi^n_{s,1} \left|d\varphi^n_{s,1}\right|\rangle_2 ds$. Since u has bounded derivatives and using Theorem 9 implies the uniform convergence of φ^n_{s1} to φ^∞_{s1} and the pointwise convergence of the derivatives (because of the uniform boundedness of $|v^n_s|_{\mathcal{B}}$), we have

$$\int_0^1 \langle \mathrm{z}^n_s, u \circ \varphi^n_{s,1} \left|d\varphi^n_{s,1}\right|\rangle_2 ds - \int_0^1 \langle \mathrm{z}^n_s, u \circ \varphi^\infty_{s,1} \left|d\varphi^\infty_{s,1}\right|\rangle_2 ds \to 0. \tag{24}$$

Moreover, from the weak convergence of z^n to z^∞, we get

$$\int_0^1 \langle \mathrm{z}^n_s, u \circ \varphi^\infty_{s,1} \left|d\varphi^\infty_{s,1}\right|\rangle_2 ds \to \int_0^1 \langle \mathrm{z}^\infty_s, u \circ \varphi^\infty_{s,1} \left|d\varphi^\infty_{s,1}\right|\rangle_2 ds, \tag{25}$$

so that finally $\langle \mathrm{j}_1 - \mathrm{j}_1', u\rangle_2 = 0$ for any $u \in C_c^\infty(\Omega, \mathbb{R}^d)$. Hence $\mathrm{j}' \in H^1([0,1], \mathcal{J}_W)$ and the result is proved. ∎

7. Initial value problem for the geodesic equation. We have the following theorem.

THEOREM 7. *Assume that $\mathcal{B}$ is continuously embedded in $C_0^p(\Omega, \mathbb{R}^p)$ for $p \geq 3$. Then, for all $T > 0$, there exists a unique solution $(\mathrm{v}, \mathrm{j}, \mathrm{z})$ of (21) over $[0,T]$, with initial values $j_0 \in H^1(\Omega, \mathbb{R}^d)$, $z_0 \in L^2(\Omega, \mathbb{R}^d)$, and $\omega_0 \in C^{p-1}([0,1], \mathbb{R}^k)'$ (where $C^{p-1}([0,1], \mathbb{R}^k)'$ denotes the topological dual of $C^{p-1}([0,1], \mathbb{R}^k)$ with the norm $|\omega| \triangleq \sup_{|v|_{p-1,\infty} \leq 1} (\omega, v)$) which continuously depends on these initial conditions.*

Continuity of the solution $(\mathrm{v}, \mathrm{j}, \mathrm{z})$ as a function of (j_0, z_0) is meant according to $H^1(\Omega, \mathbb{R}^d) \times L^2(\Omega, \mathbb{R}^d)$-norms for the initial conditions, $L^2([0,T], W)$-norm for (v, z), and $C([0,1], L^2(\Omega, \mathbb{R}^d))$-norm for j.

8. Proof of Theorem 7. To prove Theorem 7, we show the existence of solutions for short time and then extend them to all time. Fix $T > 0$. For a given $\mathrm{v} \in L^2([0,T], \mathcal{B})$, let $\Psi(\mathrm{v}) \in L^2([0,T], \mathcal{B})$ be defined by

$$\begin{cases} \Psi(\mathrm{v})_t = \mathrm{u}^v_t - \dfrac{\sigma^2}{2}\left(K_\nabla(q^{\mathrm{v}}_t \left|\mathrm{z}^{\mathrm{v}}_t\right|^2) + K(\left|\mathrm{z}^{\mathrm{v}}_t\right|^2 \nabla q^{\mathrm{v}}_t)\right), \\ \mathrm{z}^{\mathrm{v}}_t = \mathrm{z}_0 \circ \varphi^{\mathrm{v}}_{t,0} \left|d\varphi^{\mathrm{v}}_{t,0}\right|, \\ \langle \mathrm{u}^v_t\, ,\, w\rangle_{\mathcal{B}} = \left(\omega_0, (d\varphi^{\mathrm{v}}_{0,t})^{-1} w \circ \varphi^{\mathrm{v}}_{0,t}\right). \end{cases} \tag{26}$$

To estimate the Lipschitz coefficient of Ψ, we introduce $\mathrm{v}, \mathrm{v}' \in L^1([0,T], \mathcal{B})$ and compute the variation of each term in $\Psi(\mathrm{v})_t - \Psi(\mathrm{v}')_t$. Fix $w \in \mathcal{B}$ with $|w|_{\mathcal{B}} = 1$. We have

$$\begin{aligned} \langle \Psi(\mathrm{v})_t\, ,\, w\rangle_{\mathcal{B}} =& \frac{\sigma^2}{2}\left\langle \left|z^v_t\right|^2, \, q^{\mathrm{v}}_t \operatorname{div}(w)\right\rangle_2 - \frac{\sigma^2}{2}\left\langle \left|z^v_t\right|^2, \, dq^{\mathrm{v}}_t w\right\rangle_2 + \left(\omega_0, (d\varphi^{\mathrm{v}}_{0,t})^{-1} w \circ \varphi^{\mathrm{v}}_{0,t}\right) \\ =& \left(\omega_0, (d\varphi^{\mathrm{v}}_{0,t})^{-1} w \circ \varphi^{\mathrm{v}}_{0,t}\right) + \frac{\sigma^2}{2}\left\langle |z_0|^2, \, \left|d\varphi^{\mathrm{v}}_{0,t}\right|^{-1} q^{\mathrm{v}}_t \circ \varphi^{\mathrm{v}}_{0,t} \operatorname{div}(w) \circ \varphi^{\mathrm{v}}_{0,t}\right\rangle_2 \\ & - \frac{\sigma^2}{2}\left\langle |z_0|^2, \, \left|d\varphi^{\mathrm{v}}_{0,t}\right|^{-1} d_{\varphi^{\mathrm{v}}_{0,t}} q^{\mathrm{v}}_t w \circ \varphi^{\mathrm{v}}_{0,t}\right\rangle_2. \end{aligned} \tag{27}$$

We have

$$q_t^{\mathrm{v}} \circ \varphi_{0,t}^{\mathrm{v}}(x) = \int_0^t \left| d_{\varphi_{0,t}^{\mathrm{v}}(x)} \varphi_{t,s}^{\mathrm{v}} \right|^{-1} ds = \int_0^t \left| d_{\varphi_{0,s}^{\mathrm{v}}(x)} \varphi_{s,t}^{\mathrm{v}} \right| ds = \int_0^t \frac{\left| d_x \varphi_{0,t}^{\mathrm{v}} \right|}{\left| d_x \varphi_{0,s}^{\mathrm{v}} \right|} ds,$$

and letting $\xi_{s,t}^{\mathrm{v}} = \frac{\left| d_x \varphi_{0,t}^{\mathrm{v}} \right|}{\left| d_x \varphi_{0,s}^{\mathrm{v}} \right|}$ and $\lambda_t^{\mathrm{v}}(w) = (d\varphi_{0,t}^{\mathrm{v}})^{-1} w \circ \varphi_{0,t}^{\mathrm{v}}$,

(28)

$$\begin{aligned} \langle \Psi(\mathrm{v})_t \,,\, w \rangle_{\mathcal{B}} = \frac{\sigma^2}{2} \int_0^t \Big\langle |z_0|^2 \,,\, (\left| d\varphi_{0,s}^{\mathrm{v}} \right|^{-1} \operatorname{div}(w) \circ \varphi_{0,t}^{\mathrm{v}} - \left| d\varphi_{0,t}^{\mathrm{v}} \right|^{-1} \left\langle \nabla \xi_{s,t}^{\mathrm{v}} \,,\, \lambda_t^{\mathrm{v}}(w) \right\rangle) \Big\rangle ds \\ +(\omega_0, \lambda_t^{\mathrm{v}}(w)). \end{aligned}$$

This implies

$$\begin{aligned} |\Psi(\mathrm{v}')_t - \Psi(\mathrm{v})_t|_{\mathcal{B}} \le \frac{\sigma^2}{2} |z_0|_2^2 \sup \Bigg\{ \int_0^t \Bigg(& \left| d\varphi_{0,s}^{\mathrm{v}} \right|^{-1} \operatorname{div}(w) \circ \varphi_{0,t}^{\mathrm{v}} \\ & - \left| d\varphi_{0,s}^{\mathrm{v}'} \right|^{-1} \operatorname{div}(w) \circ \varphi_{0,t}^{\mathrm{v}'} \Bigg) ds : |w|_{\mathcal{B}} = 1 \Bigg\} \\ + \frac{\sigma^2}{2} |z_0|_2^2 \sup \Bigg\{ \int_0^t \Bigg(& \left| d\varphi_{0,t}^{\mathrm{v}} \right|^{-1} \left\langle \nabla \xi_{s,t}^{\mathrm{v}} \,,\, \lambda_t^{\mathrm{v}}(w) \right\rangle \\ & - \left| d\varphi_{0,t}^{\mathrm{v}'} \right|^{-1} \left\langle \nabla \xi_{s,t}^{\mathrm{v}'} \,,\, \lambda_t^{\mathrm{v}'}(w) \right\rangle \Bigg) ds : |w|_{\mathcal{B}} = 1 \Bigg\} \\ + |\omega_0| \sup \Bigg\{ & \left| \lambda_t^{\mathrm{v}}(w) - \lambda_t^{\mathrm{v}'}(w) \right|_{p-1,\infty} : |w|_{\mathcal{B}} = 1 \Bigg\}. \end{aligned}$$

The problem is thus reduced to the estimation of variations, with respect to v, of $\lambda_t^{\mathrm{v}}(w)$, $\nabla \xi_{s,t}^{\mathrm{v}}$ and of $\left| d\varphi_{0,s}^{\mathrm{v}} \right|^{-1} \operatorname{div}(w) \circ \varphi_{0,t}^{\mathrm{v}}$. They involve differentials of φ^{v}, $\varphi^{\mathrm{v}'}$, and w up to the second degree. The inclusion of $\mathcal{B}$ in $C^3([0,1], \mathbb{R}^k)$ and an application of Lemmas 7 and 11 in the appendix directly lead to the estimate

$$|\Psi(\mathrm{v})_t - \Psi(\mathrm{v}')_t|_{\mathcal{B}} \le C \left(\sigma^2 |z_0|_2^2 + |\omega_0| \right) |\mathrm{v} - \mathrm{v}'|_{1,T} \, e^{C' \max(|\mathrm{v}|_{1,T}, |\mathrm{v}'|_{1,T})}, \tag{29}$$

and finally

$$\begin{aligned} |\Psi(\mathrm{v}) - \Psi(\mathrm{v}')|_{2,T} &\le C\sqrt{T} \left(\sigma^2 |z_0|_2^2 + |\omega_0| \right) |\mathrm{v} - \mathrm{v}'|_{1,T} \, e^{C' \max(|\mathrm{v}|_{1,T}, |\mathrm{v}'|_{1,T})} \\ &\le CT \left(\sigma^2 |z_0|_2^2 + |\omega_0| \right) |\mathrm{v} - \mathrm{v}'|_{2,T} \, e^{C'\sqrt{T} \max(|\mathrm{v}|_{2,T}, |\mathrm{v}'|_{2,T})}. \end{aligned} \tag{30}$$

Therefore, Ψ is q-Lipschitz with $q < 1$ for T small enough, and its unique fixed point yields a unique solution of (21). This is stated below.

LEMMA 3. *There exists a time $T > 0$ depending only on $|z_0|_2$ and $|j_0|_{H^1}$ such that a unique solution of* (21) *exists on* $[0,T]$.

We now show that this solution can be extended to all times. For this, we prove that there exists a unique fixed point for Ψ at all times. Denote by Ψ_T this mapping when defined on $L^2([0,T], \mathcal{B})$. Clearly, if v is a fixed point of Ψ_T, its restriction to $[0,S]$ is a fixed point of Ψ_S. Thus, if T_0 is the largest T such that Ψ_S has a unique fixed point v^S in $L^2([0,S], \mathcal{B})$ for any $S < T$, then each v^T for $T < T_0$ is an extension of v^S whenever $S \le T$. We can show that $T_0 = \infty$ by showing that, if $T_0 < \infty$, then there exists $\varepsilon > 0$ (depending only on T_0 and the initial conditions) such that, for all $T < T_0$, there exists a unique extension of v^T to $[T, T+\varepsilon]$. Fix such a T; the issue of extending a fixed point of Ψ_T on $[T, T+\varepsilon]$ can be rephrased as a fixed point problem for small time with the following notation. For $\mathrm{v} \in L^2([0,T], \mathcal{B})$ and $\mathrm{v}' \in L^2([0,\varepsilon], \mathcal{B})$, define $\mathrm{v} \vee \mathrm{v}' \in L^2([0, T+\varepsilon], \mathcal{B})$, equal to v on $[0,T]$ and equal to $(t \mapsto \mathrm{v}'(t-T))$ on $[T, T+\varepsilon]$. Introduce the function $\Psi^\varepsilon : L^2([0,\varepsilon], \mathcal{B}) \to L^2([0,\varepsilon], \mathcal{B})$ defined by

$$\Psi^\varepsilon(\mathrm{v})(t) = \Psi_{T+\varepsilon}(\mathrm{v}^T \vee \mathrm{v})(t-T).$$

For $t > T$,

$$q_t^{\mathrm{v}^T \vee \mathrm{v}} = q_1^{\mathrm{v}^T} + q_{t-T}^{\mathrm{v}},$$

$\mathrm{z}_t = \mathrm{z}_T \circ \varphi_{Tt}^{\mathrm{v}} |d\varphi_{Tt}^{\mathrm{v}}|^{-1}$, and

$$\begin{aligned}
\langle \mathrm{u}_t^{\mathrm{v}}\,,\, w\rangle_{\mathcal{B}} &= \big(\omega_0, (d\varphi_{0,t}^{\mathrm{v}})^{-1} w \circ \varphi_{0,t}^{\mathrm{v}}\big) \\
&= \Big(\omega_0, (d\varphi_{0,T}^{\mathrm{v}})^{-1} (d_{\varphi_{0,T}^{\mathrm{v}}} \varphi_{Tt}^{\mathrm{v}})^{-1} w \circ \varphi_{Tt}^{\mathrm{v}} \circ \varphi_{0,T}^{\mathrm{v}}\Big) \\
&= \big(\omega_T, d\varphi_{Tt}^{\mathrm{v}})^{-1} w \circ \varphi_{Tt}^{\mathrm{v}}\big)
\end{aligned}$$

with $(\omega_T, w) = \big(\omega_0, (d\varphi_{0,T}^{\mathrm{v}})^{-1} w \circ \varphi_{0,T}^{\mathrm{v}}\big)$. It is clear that the study of Ψ^ε can follow exactly the lines of the study of Ψ_T, yielding a unique fixed point if ε is small enough, the size of admissible ε being controlled by the L^2-norms of z_T and the norm of ω_T as a linear form on $C_0^{p-1}(\Omega, \mathbb{R}^k)$. These norms can in turn be bounded by the L^2-norms of z_0 and the norm of ω_0, respectively, multiplied by a continuous function of $\max(|\varphi_{0,T}^{\mathrm{v}^T}|_{1,\infty}, |\varphi_{T,0}^{\mathrm{v}^T}|_{1,\infty})$. Proving that this is uniformly bounded for $T < T_0$ is therefore sufficient to get the contradiction we aim for, that is, that the solution can be uniquely extended beyond T_0.

So, everything relies on proving the uniform boundedness of $\varphi_{0,T}^{\mathrm{v}^T}$, $\varphi_{T,0}^{\mathrm{v}^T}$, and their derivatives over Ω. By Lemmas 7 and 9, these quantities are bounded by functions of $|\mathrm{v}^T|_{1,T}$ so that we have to prove that these can be bounded uniformly in T. However, it suffices to use the facts that v^T satisfies a geodesic equation and that geodesics travel at constant speed. More precisely, defining, for $t \le T < T_0$,

$$\psi_t = \left|\mathrm{v}_t^T\right|_{\mathcal{B}}^2 + \sigma^2 \left|\mathrm{z}_t\right|_2^2,$$

we have (recall that this does not depend on T as soon as $T \ge t$) $\psi_t \equiv \psi(0)$ so that

$$\left|\mathrm{v}^T\right|_{1,T} \le T\psi(0)$$

for all T. It is well known that *minimizing* geodesics have constant speed, but we must check that this property remains true for all the solutions of (22). This is proved in the appendix and is stated, for further reference, in the next lemma.

LEMMA 4. *If (j, v, z) is a solution of system* (22) *on $[0, T[$, then $|v_t|_{\mathcal{B}}^2 + \sigma^2 |z_t|_2^2$ is constant with respect to time.*

To prove the continuity of the solution, let (v, j, z) and (v', j', z') be two solutions of system (22) with initial conditions (ω_0, z_0) and (ω_0', z_0'), respectively. Using, in particular, the computation leading from (28) to (29), it is not to difficult to obtain the estimate

$$\begin{aligned}
|v_t - v_t'|_{\mathcal{B}} \le C &\left(|\omega_0 - \omega_0'| + \frac{\sigma^2}{2} \left||z_0|^2 - |z_0'|^2\right|\right) e^{C|\mathrm{v}|_{1,T}} \\
&+ C\left(\sigma^2 |z_0'|_2^2 + |\omega_0'|\right) |\mathrm{v} - \mathrm{v}'|_{1,T}\, e^{C|\mathrm{v}'|_{1,T}}.
\end{aligned}$$

As we just have shown, $|\mathrm{v}|_{1,T} = T\,|\mathrm{v}_0|_{\mathcal{B}}$, and this is smaller (up to a universal multiplicative constant) than $|\omega_0|$ so that

$$\begin{aligned} |v_t - v_t'|_{\mathcal{B}} \le{}& C\left(|\omega_0 - \omega_0'| + \frac{\sigma^2}{2}\left||z_0|^2 - |z_0'|^2\right|\right)e^{CT|\omega_0|} \\ &+ C\left(\sigma^2\,|z_0'|_2^2 + |\omega_0'|\right)|\mathrm{v} - \mathrm{v}'|_{1,T}\,e^{CT|\omega_0'|}. \end{aligned}$$

Gronwall's lemma now allows us to conclude that, for some constant C which may now depend on $T, |\omega_0|\,, |\omega_0'|\,, |z_0|_2$, and $|z_0'|_2$,

$$|v_t - v_t'|_{\mathcal{B}} \le C\left(|\omega_0 - \omega_0'| + \frac{\sigma^2}{2}\left||z_0|^2 - |z_0'|^2\right|\right). \tag{31}$$

9. Normal coordinates in $\ ^{1}$**.** We now consider the question, which motivated this paper, of whether the previous construction could be used as an indexing device for characterizing the deformations and variations of an object relative to a prototype.

Fix an image $j_0 \in H^1(\Omega, \mathbb{R}^d)$. The computationally simplest way to describe an image j in a neighborhood of j_0 is by the difference $j - j_0$. However, one cannot be satisfied with this representation which takes no account of the metric we have placed on $\mathcal{J}_W$. Among local charts related to the metric, normal coordinates on a Riemannian manifold are radial flattenings of this manifold onto its tangent space in the sense that radial lines in this space correspond to geodesics on the manifold. They provide a very efficient linear representation of the manifold and of its metric. Existence of such coordinates is a standard theorem in finite dimensions, and our purpose is to check how much of this result remains valid in our infinite dimensional framework.

In the previous sections, another candidate for local coordinates has emerged, which turns out to be closely related (it is in fact dual) to normal coordinates. We have proved that, for a fixed $j_0 \in H^1(\Omega, \mathbb{R}^d)$, one can associate to any $z_0 \in L^2(\Omega, \mathbb{R}^d)$ a unique solution of system (21). We introduce the function $M_{j_0} : L^2 \to L^2$, which assigns to $z_0 \in L^2$ the "image" j_1, where j_t is the solution of (21) at time t.

The following theorem shows that M_{j_0} shares some features of local coordinates on $\mathcal{J}_W$.

THEOREM 8. *Let $\mathcal{B}_{H^1}(0, \varepsilon)$ denote the open ball in $H^1(\Omega, \mathbb{R}^d)$ containing all $z_0 \in H^1(\Omega, \mathbb{R}^d)$ such that $|z_0|_{H^1} < \varepsilon$. Then, for all $j_0 \in H^1(\Omega, \mathbb{R}^d)$, there exists $\varepsilon > 0$ such that M_{j_0} restricted to $\mathcal{B}_{H^1}(0, \varepsilon)$ is continuous and one-to-one onto its image, equipped with the L^2-topology.*

Proof of Theorem 8. Continuity of $M_{j_0} : L^2(\Omega, \mathbb{R}^d) \to L^2(\Omega, \mathbb{R}^d)$ is a consequence of Theorem 7, and it trivially implies the continuity of the restriction $M_{j_0} : H^1(\Omega, \mathbb{R}^d) \to L^2(\Omega, \mathbb{R}^d)$ for the $H^1(\Omega, \mathbb{R}^d)$-topology. We show that this map is one-to-one in a neighborhood of 0. We first have the following lemma.

LEMMA 5. *Let $j_0, z_0, \tilde{z}_0 \in H^1(\Omega, \mathbb{R}^d)$, with $\max(|z_0|_{H^1}\,, |\tilde{z}_0|_{H^1}) \le 1$. Denote $\tilde{\mathrm{v}}$ the time-dependent vector field along the solution of (21) with initial condition $(j_0, \tilde{z}_0)$. Then, there exist a constant C and a function ε which depend only on j_0 such that, for $t > 0$,*

$$\begin{aligned} &\left|(M_{j_0}(t\tilde{z}_0) - M_{j_0}(tz_0)) \circ \varphi_{0,t}^{\tilde{\mathrm{v}}} - t[\sigma^2(\tilde{z}_0 - z_0) + dj_0 K(dj_0^*(\tilde{z}_0 - z_0))]\right|_2 \\ &\le Ct\,|\tilde{z}_0 - z_0|_2\,\varepsilon(t), \end{aligned}$$

and $\lim_{t\to 0}\varepsilon(t) = 0$.

The proof of Lemma 5 is given in Appendix G. To prove Theorem 8, we first remark that

$$\left|\sigma^2(\tilde z_0 - z_0) + dj_0 K(dj_0^*(\tilde z_0 - z_0))\right|_2 \geq \sigma^2 \left|\tilde z_0 - z_0\right|_2$$

so that

$$\left|(M_{j_0}(t\tilde z_0) - M_{j_0}(tz_0)) \circ \varphi^{\tilde{\mathrm{v}}}_{0,t}\right|_2 \geq t\sigma^2 \left|\tilde z_0 - z_0\right|_2 \left(1 - \frac{C}{\sigma^2}\varepsilon(j_0, t)\right),$$

and the lower bound is nonvanishing as soon as t is small enough. □

Remark that we have, for $j_1, j_2 \in H^1(\Omega, \mathbb{R}^d)$, the inequality

$$d(j_1, j_2) \leq \frac{1}{\sigma} \left|j_1 - j_2\right|_2$$

since the right-hand side is an upper bound of the length of the curve $\mathrm{j}_t = (1-t)j_1 + tj_2$ (since choosing $\mathrm{v} \equiv 0$ and $\sigma^2 \mathrm{z} \equiv j_2 - j_1$, we have $\mathrm{w}_t \triangleq (v_t, z_t) \in \frac{\partial j_t}{\partial t}$ and $\sigma^2 \int_0^1 \left|z_t\right|_2^2 = \left|j_2 - j_1\right|_2^2 / \sigma^2$). So continuity of M_{j_0} for the d-topology on its image is also true.

According to Lemma 5, normal coordinates (which are time derivatives at $t = 0$ of geodesics) are related to M by the relation (we use the standard exponential notation)

$$\exp_{j_0}(Sz_0) = M_{j_0}(z_0),$$

where S is defined by

$$Sz \triangleq \sigma^2 z + Dj_0 K(Dj_0^* z).$$

This indicates that a good approximation of the metric in terms of the z-coordinate would be

$$\left|z_1 - z_2\right|_{j_0}^2 = \langle z_1 - z_2\,,\, S(z_1 - z_2)\rangle_2,$$

which satisfies

$$\left|z_1\right|_{j_0} = d(j_0, M_{j_0}(z_1))$$

in a neighborhood of 0.

10. Experiments. In this section, we propose a preliminary set of experiments to illustrate the information contained in the z-coordinate described above. Experiments in Figures 1, 2, and 3 were conducted in two steps: given two images j_0 and j_1, we first computed the minimizing geodesic between them, yielding a trajectory $(\mathrm{j}_t, \mathrm{z}_t, \mathrm{v}_t)$ and the corresponding flow φ^{v}_t. Then, using j_0 again, and the obtained value z_0 on the minimizing geodesic, we computed the solution of (21) until time $t = 1$. The obtained values $(\mathrm{j}'_t, \mathrm{z}'_t, \mathrm{v}'_t)$ could then be compared with those which have been computed along the geodesics. In Figure 4, the initial j_0 is the same as in Figures 2 and 3, but the z_0 is the average of the two so that it does not correspond to any precomputed geodesic in the image space. The result is quite interesting, because it still possesses characteristics of a human face and can be compared to the result of a simple linear combination of the target images in Figures 2 and 3.

The numerical implementation of both operations (minimization of the geodesic energy and integration of (21)) must be done with some care in order, in particular, to avoid instabilities due to the conservation part of the energy. Details will be provided in a forthcoming paper.

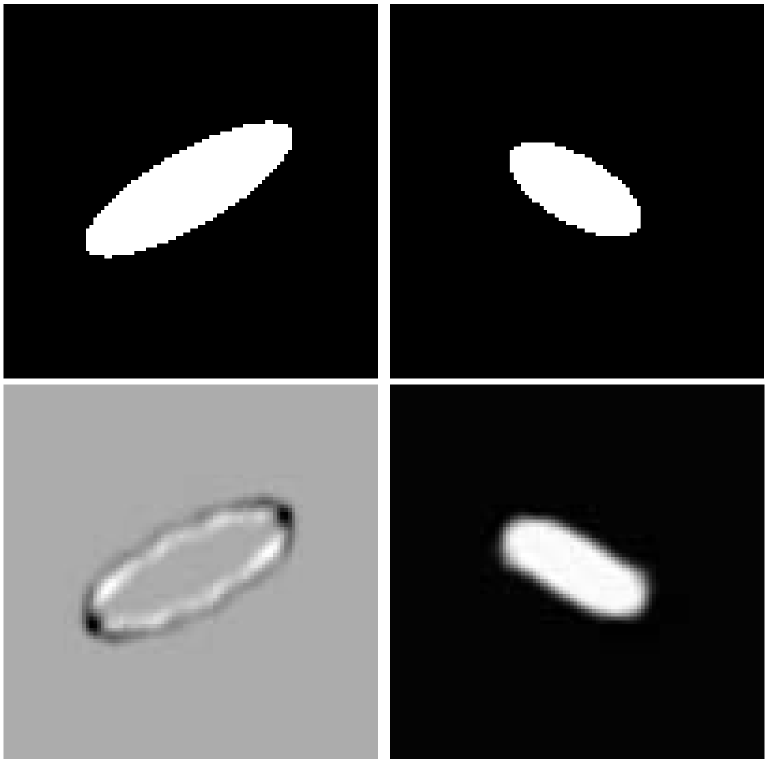


FIG. 1. *From top to bottom and from left to right: Initial image, target image, z-coordinate, reconstructed target image.*

Appendix A. Proofs of Propositions 2 and 3.

A.1. Proof of Proposition 2. The proof relies on a sequence of standard measurability arguments, of which we sketch only the main steps. First let $(w_n)_{n\in\mathbb{N}}$ be a Hilbert basis of W. Since, for any $u \in C_c^\infty(\Omega, \mathbb{R}^d)$ and $w = (v, z) \in W$, $j \to l_{j,u}(w)$ (which has been defined in (6) by $\sigma^2\langle z, u\rangle_2 + \langle j, \operatorname{div}(u \otimes v)\rangle_2$) is continuous from $L^2(\Omega, \mathbb{R}^d)$ to $\mathbb{R}$, the map

$$j \mapsto w_{j,u} \triangleq \sum_{n\geq 0} l_{j,u}(w_n) w_n$$

is measurable from $L^2(\Omega, \mathbb{R}^d)$ to W. By construction, we have, for $w \in W$,

$$\langle w\,,\, w_{ju}\rangle_W = \sum_{n\geq 0} l_{j,u}(w_n)\langle w\,,\, w_n\rangle_W = l_{j,u}(w).$$

Thus, for $\gamma \in T_j\mathcal{J}_W$, we have

$$\overline{p}(\gamma) = \operatorname{Argmin}\left\{|w|_W : \langle w\,,\, w_{j,u}\rangle = \langle \gamma\,,\, u\rangle \text{ for all } u \in C_c^\infty(\Omega, \mathbb{R}^d)\right\}.$$

Introducing a family $(u_n)_{n\in\mathbb{N}}$ in $C_c^\infty(\Omega, \mathbb{R}^d)$ which is dense in $H_0^1(\Omega, \mathbb{R}^d)$, the previous expression may be replaced by

$$\overline{p}(\gamma) = \operatorname{Argmin}\left\{|w|_W : \langle w\,,\, w_{j,u_n}\rangle = \langle \gamma\,,\, u_n\rangle \text{ for all } n \geq 0\right\}.$$

FIG. 2. *From top to bottom and from left to right: Initial image, target image, z-coordinate, reconstructed target image.*

For $N \in \mathbb{N}$ and $\lambda > 0$, we define

$$\overline{p}^{N,\lambda}(\gamma) = \text{Argmin}\left\{ |w|_W^2 + \lambda \sum_{n=0}^{N} \left(\langle w\,,\, w_{j,u_n} \rangle_W - \langle \gamma\,,\, u_n \rangle \right)^2 \right\}. \tag{32}$$

Clearly, we must have $\overline{p}^{N,\lambda}(\gamma) = \sum_{i=1}^{N} x_i w_{j,u_i}$, where

$$x = \underset{x' \in \mathbb{R}^{N+1}}{\text{Argmin}} \left\{ \left| \sum_{n=0}^{N} x'_n w_{j,u_n} \right|_W^2 + \lambda \sum_{n=0}^{N} \left(\sum_{n'=1}^{N} x'_{n'} \langle w_{j,u_{n'}}\,,\, w_{j,u_n} \rangle_W - \langle \gamma\,,\, u_i \rangle \right)^2 \right.$$
$$\left. + \ \frac{1}{\lambda} |x'|^2 \right\}.$$

For $\lambda > 0$, the optimal x is given by $x = (A + I/\lambda)^{-1} y$, where $y \in \mathbb{R}^{N+1}$ is such that $y_i = \langle \gamma\,,\, u_i \rangle$ and A is an $(N+1) \times (N+1)$ matrix with coefficients given

FIG. 3. *From top to bottom and from left to right: Initial image, target image, z-coordinate, reconstructed target image.*

by $a_{n,n'} = \left\langle w_{j,u_{n'}}\,,\, w_{j,u_n}\right\rangle_W$. This implies that, if γ_t is a measurable path, the function $t \mapsto \overline{p}^{N,\lambda}(\gamma_t)$ is measurable. The measurability of $\overline{p}(\gamma_t)$ is a consequence of the pointwise convergence of $\overline{p}^{N,N}(\gamma_t)$ to $\overline{p}(\gamma_t)$, which comes from the following argument: for all N and λ, we have $\left|\overline{p}^{N,\lambda}(\gamma)\right|_W \leq |\overline{p}(\gamma)|_W$, since the last term in (32) vanishes for $w = \overline{p}(\gamma)$. For the same reason,

$$\sum_{n=0}^{N} \left(\left\langle \overline{p}^{N,\lambda}(\gamma)\,,\, w_{j,u_n}\right\rangle - \langle \gamma\,,\, u_n\rangle\right)^2 \leq \frac{1}{\lambda}\,|\overline{p}(\gamma)|_W\,,$$

which implies that for all n, $\left\langle \overline{p}^{N,N}(\gamma)\,,\, w_{j,u_n}\right\rangle \to \langle \gamma\,,\, u_n\rangle$ when N tends to infinity. Moreover, for any weakly converging subsequence extracted from $\overline{p}^{N,N}(\gamma)$ (which forms a weakly compact set in W), we have, and denoting w^* its limit, $|w^*|_W \leq \liminf \left|\overline{p}^{N,N}(\gamma)\right|_W \leq |\overline{p}(\gamma)|_W$, and, for all n, $\langle w^*\,,\, w_{j,u_n}\rangle = \langle \gamma\,,\, u_n\rangle$ by weak convergence, which is only possible when $w^* = \overline{p}(\gamma)$.

Hence $t \mapsto \overline{p}(\gamma_t)$ is measurable if γ_t is measurable, and the proof of Proposition 2 is ended.

FIG. 4. *From top to bottom and from left to right: Initial image, target image, z-coordinate, obtained by averaging the z-coordinate in Figures* 2 *and* 3, *and obtained target image.*

A.2. Proof of Proposition 3. We deduce from Proposition 2 that it is sufficient to prove the next proposition.

PROPOSITION 6. *Let $w \in L^2([0,1], W)$ such that for any $u \in C_c^\infty(\Omega \times]0,1[, \mathbb{R}^d)$ we have*

$$\int_0^1 \left(\sigma^2 \langle z_t, u_t \rangle_2 + \langle j_t, \operatorname{div}(u_t \otimes v_t) \rangle_2\right) \, dt = 0. \tag{33}$$

Then almost everywhere in t, $w_t \in E_{j_t}$.

Proof. Let $(u_n)_{n \in \mathbb{N}}$ be a family in $C_c^\infty(\Omega, \mathbb{R}^d)$ dense in $C_c^\infty(\Omega, \mathbb{R}^d)$ for the $H^1(\Omega, \mathbb{R}^d)$-norm. If we prove that for any $n \in \mathbb{N}$, the function c_n defined by $c_n(t) \triangleq \sigma^2 \langle z_t, u_n \rangle_2 + \langle j_t, \operatorname{div}(u_n \otimes v_t) \rangle_2$ is vanishing almost everywhere, then by density, there exists a negligible set $\mathcal{N}$ such that for any $t \in [0,1] \setminus \mathcal{N}$ and any $u \in C_c^\infty(\Omega, \mathbb{R}^d)$

$$\sigma^2 \langle z_t, u \rangle_2 + \langle j_t, \operatorname{div}(u \otimes v_t) \rangle_2 = 0,$$

which implies Proposition 6. Hence, let us consider $n \in \mathbb{N}$. For any $f \in C_c^\infty([0,1],\mathbb{R})$, if $u(t,x) \triangleq f(t)u_n(x)$, we have from (33) that

$$\int_0^1 c_n(t)f(t)dt = 0$$

so that, by standard arguments, we get $c_n = 0$ almost everywhere. ☐

Appendix B. Proof of Theorem 2. We start the ($\Leftarrow$) part in the case $L^2([0,1],W)$.

LEMMA 6. *Let* $\mathrm{w} = (\mathrm{z},\mathrm{v}) \in L^2([0,1],W)$. *Let us define for any* $t \in [0,1]$

$$\mathrm{j}_t \triangleq \mathrm{j}_0 \circ \varphi_{t,0} + \sigma^2 \int_0^t \mathrm{z}_s \circ \varphi_{t,s} ds,$$

where φ_t *is the flow at time* t *associated with* v. *Then* j *is regular.*

Proof. Let us notice first that

$$\mathrm{j}_{t+h} = \mathrm{j}_t \circ \varphi_{t+h,t} + \sigma^2 \int_t^{t+h} \mathrm{z}_s \circ \varphi_{t+h,s} dt. \tag{34}$$

From equality (34), the continuity in $\mathcal{J}_W$ of j is straightforward.

It is sufficient to prove that for any $u \in C_c^\infty(\Omega\times]0,1[,\mathbb{R}^d)$, we have

$$-\int_0^1 \left\langle \mathrm{j}_t, \frac{\partial u}{\partial t} \right\rangle_2 dt = \int_0^1 \left(\sigma^2 \langle \mathrm{z}_t, u_t\rangle_2 + \langle \mathrm{j}_t, \mathrm{div}(u_t \otimes \mathrm{v}_t)\rangle_2\right) dt. \tag{35}$$

Indeed, if (35) is proved, if for any $t \in [0,1]$ we denote $\gamma_t \triangleq (\mathrm{j}_t, \overline{\mathrm{w}}_t)$, we have for any $u \in C_c^\infty(\Omega,\mathbb{R}^d)$, $t \to \langle \gamma_t\, ,\, u\rangle = \sigma^2\langle \mathrm{z}_t, u\rangle_2 + \langle \mathrm{j}_t, \mathrm{div}(u\otimes \mathrm{v}_t)\rangle_2$ measurable, and $|\gamma_t|_{\mathrm{j}_t} \le |\mathrm{w}_t|_W$ so that $\int_0^1 |\gamma_t|_t^2 dt \le \int_0^1 |\mathrm{w}_t|_W^2 dt < +\infty$, and the lemma is proved.

We have

$$-\int_0^1 \left\langle \mathrm{j}_t, \frac{\partial u}{\partial t}\right\rangle_2 dt = -\lim_{h\to 0}\int_0^1 \left\langle \mathrm{j}_t, \frac{u_t - u_{t-h}}{h}\right\rangle_2 dt = \lim_{h\to 0}\int_0^1 \left\langle \frac{\mathrm{j}_{t+h}-\mathrm{j}_t}{h}, u_t\right\rangle_2 dt$$

so that

$$-\int_0^1 \left\langle \mathrm{j}_t, \frac{\partial u}{\partial t}\right\rangle_2 = \lim_{h\to 0}\frac{1}{h}\int_0^1 \left(\langle \mathrm{j}_t\circ\varphi_{t+h,t} - \mathrm{j}_t, u_t\rangle_2 dt + \sigma^2 \int_t^{t+h} \langle \mathrm{z}_s \circ \varphi_{t+h,s}, u_t\rangle_2 ds\right) dt.$$

However, $\langle \mathrm{j}_t\circ\varphi_{t+h,t} - \mathrm{j}_t, u_t\rangle_2 = \int_t^{t+h} \langle \mathrm{j}_t \circ \varphi_{s,t}, \mathrm{div}(u_t\otimes \mathrm{v}_s)\rangle_2 ds$ so that

$$-\int_0^1 \left\langle \mathrm{j}_t, \frac{\partial u}{\partial t}\right\rangle_2 dt = \lim_{h\to 0}\int_0^1 \frac{1}{h}\left(\int_t^{t+h} \langle \mathrm{j}_t\circ\varphi_{s,t}, \mathrm{div}(u_t\otimes \mathrm{v}_s)\rangle_2 + \sigma^2\langle \mathrm{z}_s\circ\varphi_{t+h,s}, u_t\rangle_2 ds\right) dt$$

$$= \lim_{h\to 0}\int_0^1 \frac{1}{h}\left(\int_t^{t+h} \langle \mathrm{j}_t, \mathrm{div}(u_t\otimes \mathrm{v}_s)\circ\varphi_{t,s}|d\varphi_{t,s}|\rangle_2 + \sigma^2\langle \mathrm{z}_s, u_t\circ\varphi_{s,t+h}|d\varphi_{s,t+h}|\rangle_2 ds\right) dt.$$

Since j_t is uniformly bounded on L^2 and $|\varphi_{t,s} - I|_{1,\infty} = \epsilon(|t-s|)$ (since $\mathcal{B}$ is continuously embedded in $C^1(\overline{\Omega},\mathbb{R}^k)$), there exists $C > 0$ such that

$$\left|\int_0^1 \frac{1}{h}\int_t^{t+h} \langle \mathrm{j}_t, \mathrm{div}(u_t\otimes \mathrm{v}_s)\circ(\varphi_{t,s}|d\varphi_{t,s}| - I)\rangle_2 ds dt\right| \le C\epsilon(h)\int_0^1 |\mathrm{v}_t|_{1,\infty}\, dt \tag{36}$$

$$\le C'\epsilon(h)\left(\int_0^1 |\mathrm{w}_t|_W^2 dt\right)^{1/2}. \tag{37}$$

Now, using again the fact that j_t is uniformly bounded in L^2 and fact that $C([0,1],L^2(\Omega,\mathbb{R}^k))$ is dense in $L^2([0,1],L^2(\Omega,\mathbb{R}^k))$, we get

$$\left|\int_0^1 \frac{1}{h}\int_t^{t+h} \langle \mathrm{j}_t, \mathrm{div}(u_t\otimes \mathrm{v}_s) - \mathrm{div}(u_t\otimes \mathrm{v}_t)\rangle_2 ds dt\right| \tag{38}$$
$$\leq C\int_0^1 \frac{1}{h}\int_t^{t+h} |\mathrm{div}(u_t\otimes \mathrm{v}_s) - \mathrm{div}(u_t\otimes \mathrm{v}_t)|_2 ds dt$$
$$\to 0 \text{ when } h\to 0.$$

At this point we have proved that

$$\lim_{h\to 0}\frac{1}{h}\int_0^1 \langle \mathrm{j}_t\circ\varphi_{t+h,t} - \mathrm{j}_t, u_t\rangle_2 dt = \int_0^1 \langle \mathrm{j}_t, \mathrm{div}(u_t\otimes \mathrm{v}_t)\rangle_2 dt. \tag{39}$$

Still using the fact that $|\varphi_{t,s}-I|_{1,\infty} = \epsilon(|t-s|)$ and the fact that $|u_t|_{1,\infty}$ is uniformly bounded, we have

$$\sigma^2\int_0^1 \frac{1}{h}\int_t^{t+h} \langle \mathrm{z}_s, u_t\circ(\varphi_{s,t+h}|d\varphi_{s,t+h}| - I)\rangle_2 \; ds dt \leq C\sigma^2\epsilon(h)\int_0^1 |\mathrm{z}_s|_2 ds \tag{40}$$
$$\leq C\sigma\epsilon(h)\left(\int_0^1 |\mathrm{w}_t|_2^2\right)^{1/2}. \tag{41}$$

Finally, since $|u_t|_\infty$ is uniformly bounded, we get

$$\lim_{h\to 0}\left|\int_0^1 \frac{1}{h}\int_t^{t+h} \langle \mathrm{z}_s - \mathrm{z}_t, u_t\rangle_2 ds dt\right| \leq \lim_{h\to 0} C\int_0^1\int_t^{t+h} |\mathrm{z}_s-\mathrm{z}_t|_2 \; ds dt = 0.$$

Hence the proof of the lemma is ended. □

Let us consider the ($\Rightarrow$) part of Theorem 2 for $H^1([0,1],\mathcal{J}_W)$. Let $j\in H^1([0,1],\mathcal{J}_W)$ be a regular path, and let $\mathrm{w}_t = \overline{p}(\frac{\partial j}{\partial t})$ for any $t\in[0,1]$. We get from Proposition 2 that $\mathrm{w}\in L^2([0,1],W)$. Hence, let us define the new path j' by

$$\mathrm{j}'_t = \mathrm{j}_0\circ\varphi_{t,0} + \sigma^2\int_0^t \mathrm{z}_s\circ\varphi_{t,s} ds,$$

where φ is the flow associated with v. From the ($\Leftarrow$) part, we get that j' is regular and that $\frac{\partial \mathrm{j}'}{\partial t} = \frac{\partial j}{\partial t}$. Now let $u_0\in C_c^\infty(\Omega,\mathbb{R}^d)$. For any $f\in C_c^\infty(]0,1[,\mathbb{R})$ if $u(t,x) = u_0(x)f(t)$ for any $x\in\Omega$ and $t\in[0,1]$, we have from the integration by parts formula for a regular path

$$\int_0^1 r(t)f'(t)dt = 0,$$

where $r(t) = \langle \mathrm{j}_t, u\rangle_2 - \langle \mathrm{j}'_t, u\rangle_2$. Since r is continuous and $r(0)=0$, we get $r\equiv 0$. Considering arbitrary u_0, we get finally $\mathrm{j}_t = \mathrm{j}'_t$ for any $t\in[0,1]$.

Since the ($\Rightarrow$) part for $C^1([0,1],\mathcal{J}_W)$ is a straightforward consequence of the definition of C^1 curves and of the ($\Rightarrow$) part for $H^1([0,1],\mathcal{J}_W)$, we consider the ($\Leftarrow$) part for $\mathrm{w}\in C([0,1],W)$. We get from the corresponding part for $L^2([0,1],W)$ that (35) is still true. For any $f\in C_c^\infty([0,1],\mathbb{R})$ and any $u\in C_c^\infty(\Omega,\mathbb{R}^d)$ we have

$$-\int_0^t f'(t)\langle \mathrm{j}_t, u\rangle_2 dt = \int_0^1 f(t)\left(\sigma^2\langle \mathrm{z}_t, u_t\rangle_2 + \langle \mathrm{j}_t, \mathrm{div}(u\otimes \mathrm{v}_t)\rangle_2\right)dt.$$

One easily checks that $t \to \sigma^2\langle \mathrm{z}_t, u_t\rangle_2 + \langle \mathrm{j}_t, \mathrm{div}(u \otimes \mathrm{v}_t)\rangle_2$ is continuous as well as $t \to \langle \mathrm{j}_t, u\rangle_2$ so that, considering smooth approximates of step functions, we deduce that

$$\langle \mathrm{j}_s, u\rangle_2 = \langle \mathrm{j}_0, u\rangle_2 + \int_0^s \left(\sigma^2\langle \mathrm{z}_t, u_t\rangle_2 + \langle \mathrm{j}_t, \mathrm{div}(u \otimes \mathrm{v}_t)\rangle_2\right) dt,$$

and the result is proved.

Appendix C. Regularity results for A . In this section, we collect a few useful results on how the regularity of the diffeomorphism $\mathbf{A}_T(\mathrm{v}) = \varphi_T^{\mathrm{v}}$ may be related to the norm on $\mathcal{B}$, provided this norm can in turn control a sufficient number of derivatives; the first result deals with boundedness. In the following, we assume at least that $\mathcal{B}$ is continuously embedded in $C_0^1(\Omega, \mathbb{R}^k)$ so that $\mathbf{A}_T$ is well defined for all T. In this case,

$$\varphi_T^{\mathrm{v}}(x) = x + \int_0^T \mathrm{v}_s(\varphi_s^{\mathrm{v}}(x))ds.$$

If v_s had p space derivatives for all s, a formal differentiation of this equality yields

$$d^p\varphi_T^{\mathrm{v}} = d^p\mathrm{id} + \int_0^T d^p(\mathrm{v}_s \circ \varphi_s^{\mathrm{v}})ds. \tag{42}$$

This can be proved rigorously from rather standard arguments in the study of ODEs and is stated in the next lemma, for which we provide a proof for completeness, because of the small complication due to the fact that we have only an L^1 control with respect to the t variable, instead of the usual uniform one.

LEMMA 7. *If $p \geq 1$ and $\mathcal{B}$ is embedded in $C_0^p(\Omega, \mathbb{R}^k)$, then, for all $\mathrm{v} \in L^1([0,T], \Omega)$, φ^{v} is p times differentiable and, for all $q \leq p$,*

$$\frac{\partial}{\partial t} d^q\varphi_t^{\mathrm{v}} = d^q(\mathrm{v}_t \circ \varphi_t^{\mathrm{v}}).$$

Moreover, there exist constants C, C' such that, for all $\mathrm{v} \in L^1([0,T], \Omega)$,

$$\sup_{s\in[0,T]} |\varphi_s^{\mathrm{v}}|_{p,\infty} \leq Ce^{C'|\mathrm{v}|_{1,T}}. \tag{43}$$

Proof. For further reference, we first state Gronwall's lemma.

LEMMA 8 (Gronwall). *Asume that α and β are two positive, continuous functions on the interval $[0,c]$ and that*

$$w(t) \leq \alpha(t) + \int_0^t \beta(s)w(s)ds.$$

Then,

$$w(t) \leq \alpha(t) + \int_0^t \alpha(s)\beta(s)e^{\int_s^t \beta(u)du}ds.$$

The continuity of $x \mapsto \varphi^{\mathrm{v}}_{0,t}(x)$ is a direct consequence of this lemma since, for $x, y \in \Omega$,

$$\begin{aligned}|\varphi^{\mathrm{v}}_{0,t}(x) - \varphi^{\mathrm{v}}_{0,t}(y)| &= \left| x - y + \int_0^t \left(\mathrm{v}_s(\varphi^{\mathrm{v}}_{0,s}(x)) - \mathrm{v}_s(\varphi^{\mathrm{v}}_{0,s}(y))\right) ds \right| \\ &\leq |x-y| + \int_0^t \|\mathrm{v}_s\|_{1,\infty} |\varphi^{\mathrm{v}}_{0,s}(x) - \varphi^{\mathrm{v}}_{0,s}(y)| ds,\end{aligned}$$

and Gronwall's lemma implies

$$|\varphi^{\mathrm{v}}_{0,t}(x) - \varphi^{\mathrm{v}}_{0,t}(y)| \leq |x-y| \exp(C\,|\mathrm{v}|_{1,T}). \tag{44}$$

Assume $p = 1$ and pass now to the differential of $\varphi^{\mathrm{v}}_{0,t}$. Fix $x \in \Omega$ and introduce the linear differential equation, formally obtained in (42) for $p = 1$,

$$\frac{\partial W_t}{\partial t} = d_{\varphi^{\mathrm{v}}_{0,t}(x)} \mathrm{v}_t W_t \tag{45}$$

with initial condition $W(0) = \delta \in \mathbb{R}^k$. We skip the argument ensuring the existence and uniqueness of a solution of this equation on $[0, 1]$ and proceed to identifying it as $W_t = d_x \varphi^{\mathrm{v}}_{0,t} \delta$. Denote

$$a_\varepsilon(t) = \left(\varphi^{\mathrm{v}}_{0,t}(x + \varepsilon\delta) - \varphi^{\mathrm{v}}_{0,t}(x)\right)/\varepsilon - W_t.$$

For $\alpha > 0$, introduce

$$\mu_t(\alpha) = \max\left\{|d_x \mathrm{v}_t - d_y \mathrm{v}_t| : x, y \in \Omega, |x - y| \leq \alpha\right\}.$$

The function $d_x \mathrm{v}_t \in C^1_0(\Omega)$ being uniformly continuous on the compact set $\overline{\Omega}$, we have $\lim_{\alpha \to 0} \mu_t(\alpha) = 0$. We may write

$$\begin{aligned}a_\varepsilon(t) = &\frac{1}{\varepsilon} \int_0^t \left(\mathrm{v}_s(\varphi^{\mathrm{v}}_{0,s}(x + \varepsilon\delta)) - \mathrm{v}_s(\varphi^{\mathrm{v}}_{0,s}(x))\right) ds - \int_0^t d_{\varphi^{\mathrm{v}}_{0,s}(x)} \mathrm{v}_s W_s ds \\ = &\int_0^t d_{\varphi^{\mathrm{v}}_{0,s}(x)} \mathrm{v}_s a_\varepsilon(s) ds \\ &+ \frac{1}{\varepsilon} \int_0^t (\mathrm{v}_s(\varphi^{\mathrm{v}}_{0,s}(x + \varepsilon\delta)) - \mathrm{v}_s(\varphi^{\mathrm{v}}_{0,s}(x)) - \varepsilon d_{\varphi^{\mathrm{v}}_{0,s}(x)} \mathrm{v}_s(\varphi^{\mathrm{v}}_{0,s}(x + \varepsilon\delta) - \varphi^{\mathrm{v}}_{0,s}(x))) ds.\end{aligned}$$

Since for all $y, y' \in \Omega$

$$|\mathrm{v}_t(y') - \mathrm{v}_t(y) - d_y \mathrm{v}_t(y' - y)| \leq \mu_s(|y' - y|)\,|y' - y|,$$

we may write

$$|a_\varepsilon(t)| \leq \int_0^t |\mathrm{v}_s|_{1,\infty} |a_\varepsilon(s)|\, ds + C(\mathrm{v})\,|\delta| \int_0^1 \mu_s(\varepsilon C(\mathrm{v})\,|\delta|) ds$$

for some constant $C(\mathrm{v})$ which depends only on v. The fact that $a_\varepsilon(t)$ tends to 0 when $\varepsilon \to 0$ now is a direct consequence of Gronwall's lemma and of the fact that

$$\lim_{\alpha \to 0} \int_0^1 \mu_s(\alpha) ds = 0,$$

which is true by the dominated convergence theorem, since μ_s pointwise converges to 0 and $\mu_s(\alpha) \leq 2\,|\mathrm{v}|_{1,\infty}$. This proves Lemma 7 in the case $p = 1$. The rest of the proof is by induction: let $q_0 \leq p$, $q_0 > 1$, and assume that the result is proved for all $q < q_0$:

$$\frac{\partial}{\partial t} d^q \varphi_t^{\mathrm{v}} = d^q(\mathrm{v}_t \circ \varphi_t^{\mathrm{v}}).$$

This implies that for $\delta_1, \ldots, \delta_q \in \mathbb{R}^k$, we may write

$$\frac{\partial}{\partial t} d^q \varphi_t^{\mathrm{v}}(\delta_1, \ldots, \delta_q) = d_{\varphi_{0,t}^{\mathrm{v}}} \mathrm{v}_t d^q \varphi_t^{\mathrm{v}}(\delta_1, \ldots, \delta_q) + \sum_{l=2}^{q} d^l \mathrm{v}_t(\delta_1^{(l)}, \ldots, \delta_l^{(l)}),$$

each vector $\delta_k^{(l)}$ being a linear combination (with universal coefficients) of terms of the kind $d^{l'} \varphi_{0,t}^{\mathrm{v}}(\delta_{i_1}, \ldots, \delta_{i_{l'}})$ with $l' \leq q + 1 - l$ (this result on the differentials of the composition of two functions can be easily proved by induction). This is a linear equation in $d^q \varphi_t^{\mathrm{v}}(\delta_1, \ldots, \delta_q)$, which is valid for $q = q_0 - 1$, and the proof of its validity for q_0 follows exactly the same lines as for $p = 1$.

This expression also shows (using Gronwall's lemma) that $|d^q \varphi_t^{\mathrm{v}}|_\infty$ may be bounded by an expression of the kind

$$|\mathrm{v}|_{1,T}\, \tilde{C}\left(|d\varphi_t^{\mathrm{v}}|_\infty, \ldots, \left|d^{q-1}\varphi_t^{\mathrm{v}}\right|_\infty\right) \exp\left(C\,|\mathrm{v}|_{1,T}\right),$$

where $\tilde{C}$ is a polynomial, which in turn implies (43). $\square$

The same estimate is true for $(\varphi^{\mathrm{v}})^{-1}$.

Lemma 9. *If $p \geq 1$ and $\mathcal{B}$ is continuously embedded in $C_0^p(\Omega, \mathbb{R}^k)$, there exist constants C, C' such that, for all $\mathrm{v} \in L^1([0,T], \Omega)$,*

$$\sup_{s \in [0,T]} \left|(\varphi_s^{\mathrm{v}})^{-1}\right|_{p,\infty} \leq C e^{C' |\mathrm{v}|_{1,T}}.$$

Lemma 9 is a consequence of Lemma 7 and of the fact that $(\varphi_t^{\mathrm{v}})^{-1} = \varphi_t^{\mathrm{w}}$ with $\mathrm{w}_s = -\mathrm{v}_{t-s}$ on $[0,\, t]$.

We now pass to sufficient conditions for Lipschitz continuity of $\mathbf{A}_T$. For this, let $\mathrm{v}, \mathrm{v}' \in L^1([0,T], \mathcal{B})$. For $\xi \in [0,1]$, denote $\mathrm{v}^\xi = (1-\xi)\mathrm{v} + \xi \mathrm{v}'$ and $\varphi^\xi = \varphi^{\mathrm{v}^\xi}$.

Lemma 10.

$$\frac{\partial}{\partial \xi} \varphi_{s,t}^{\mathrm{v}^\xi}(x) = \int_s^t d_{\varphi_{s,u}^{\mathrm{v}^\xi}(x)} \varphi_{ut}^{\mathrm{v}^\xi}(\mathrm{v}'_u - \mathrm{v}_u) \circ \varphi_{s,u}^{\mathrm{v}^\xi}(x) du. \tag{46}$$

Proof. Let us first start with a formal differentiation of

$$\frac{\partial \varphi_{s,t}^{\mathrm{v}^\xi}}{\partial t} = v_t^\xi \circ \varphi_{s,t}^{\mathrm{v}^\xi}$$

with respect to ξ, which yields

$$\frac{\partial}{\partial t} \frac{\partial}{\partial \xi} \varphi_{s,t}^{\mathrm{v}^\xi} = (v'_t - v_t) \circ \varphi_t^{\mathrm{v}^\xi} + d_{\varphi_{s,t}^{\mathrm{v}^\xi}} v_t^\xi \frac{d}{d\xi} \varphi_{s,t}^{\mathrm{v}^\xi},$$

which naturally leads us to introduce the solution of the differential equation

$$\frac{\partial}{\partial t} W_t = (v'_t - v_t) \circ \varphi_t^{\mathrm{v}^\xi} + d_{\varphi_{s,t}^{\mathrm{v}^\xi}} v_t^\xi W_t \tag{47}$$

with initial condition $W_s = 0$. Noting that we have already encountered this equation without the constant term in (45), the solution of which is of the form $d_x\varphi^{\mathrm{v}^\xi}_{0,t}\delta$, a standard argument by variation of the constant shows that the solution of (47) is given by the right-hand term of (46). Therefore, the proof boils down to show that the interversion of derivatives underlying the formal argument above can be made rigorous.

For this, it clearly suffices to consider the problem in the vicinity of $\xi = 0$. The proof in fact follows the same lines as the proof of Lemma 7: letting W_t be the solution of (47), we let

$$a_\xi(t) = \left(\varphi^{\mathrm{v}^\xi}_{s,t}(x) - \varphi^{\mathrm{v}}_{s,t}(x)\right)/\xi - W_t$$

and express it under the form, letting $h_u = v'_u - v_u$,

$$\begin{aligned} a_\xi(t) = &\int_s^t d_{\varphi^{\mathrm{v}}_{s,u}} \mathrm{v}_u a_\xi(u) du + \int_s^t (h_u(\varphi^{\mathrm{v}^\xi}_{s,u}(x)) - h_u(\varphi^{\mathrm{v}}_{s,u}(x)))du \\ &+ \frac{1}{\xi}\int_s^t \left(\mathrm{v}_u(\varphi^{\mathrm{v}^\xi}_{s,u}(x)) - \mathrm{v}_u(\varphi^{\mathrm{v}}_{s,u}(x)) - \xi d_{\varphi^{\mathrm{v}}_{s,u}(x)}\mathrm{v}_u \left(\varphi^{\mathrm{v}^\xi}_{s,u}(x) - \varphi^{\mathrm{v}}_{s,u}(x)\right)\right) du. \end{aligned}$$

The proof can proceed exactly as that of Lemma 7, provided it has been shown that $|\varphi^{\mathrm{v}^\xi}_{s,u}(x) - \varphi^{\mathrm{v}}_{s,u}(x)|$ tends to 0 with ξ, which is again a direct consequence of Gronwall's lemma and of the inequality

$$\left|\varphi^{\mathrm{v}^\xi}_{s,t}(x) - \varphi^{\mathrm{v}}_{s,t}(x)\right| \le \int_s^t |\mathrm{v}_u|_{1,\infty} \left|\varphi^{\mathrm{v}^\xi}_{s,u}(x) - \varphi^{\mathrm{v}}_{s,u}(x)\right| du + \xi \int_s^t |h_u|_\infty du. \qquad \square$$

This lemma implies, in particular, that

$$\varphi^{\mathrm{v}'}_{s,t}(x) - \varphi^{\mathrm{v}}_{s,t}(x) = \int_0^1 \int_s^t d_{\varphi^{\mathrm{v}^\xi}_{s,u}(x)} \varphi^{\mathrm{v}^\xi}_{ut}(\mathrm{v}'_u - \mathrm{v}_u) \circ \varphi^{\mathrm{v}^\xi}_{s,u}(x) du d\xi, \tag{48}$$

which almost immediately leads to the following result (by computing differentials and applying Lemma 7).

LEMMA 11. *Assume that $\mathcal{B}$ is continuously embedded in $C^p_0(\Omega)$. If $\mathrm{v}, \mathrm{v}' \in L^1([0,T], \mathcal{B})$, we have, for $t \le T$,*

$$\left|\varphi^{\mathrm{v}}_t - \varphi^{\mathrm{v}'}_t\right|_{p-1,\infty} \le C_p \left|\mathrm{v} - \mathrm{v}'\right|_{1,t} e^{C_p(|\mathrm{v}|_{1,t} + |\mathrm{v}'|_{1,t})}$$

for some constant C_p which depends only on p.

The same results apply on $L^2([0,1], \mathcal{B})$, since $|\mathrm{v}|_{1,T} \le \sqrt{T}\,|\mathrm{v}|_{2,T}$, but, in this space, weak continuity is true under more general conditions.

THEOREM 9 (Trouvé, Dupuis, et al). *Assume that $\mathcal{B}$ is continuously embedded in $C^p_0(\Omega, \mathbb{R}^k)$. Then the map*

$$\begin{array}{rccl} \tilde{\mathbf{A}}_T : & L^2([0,T], \mathcal{B}) & \to & C^p([0,T] \times \overline{\Omega}, \mathbb{R}^k), \\ & \mathrm{v} & \mapsto & \varphi^{\mathrm{v}}_{\cdot}(.) \end{array}$$

is continuous for the weak *topology on $L^2([0,1], \mathcal{B})$ and the norm $|.|_{T,p-1,\infty}$ on $C^p([0,T] \times \overline{\Omega}, \mathbb{R}^k)$ defined by $|\varphi|_{T,p-1,\infty} = \text{ess.sup}(|\varphi_t|_{p-1,\infty}, t \in [0,T])$.*

Moreover, assume that the embedding is compact, that v^n *converges weakly to* v*, and that there exists a constant* A *such that, for all* n *and almost all* $s \in [0,1]$, $|v_s^n|_{\mathcal{B}} \leq A$*. Then, for all* $x \in \Omega$ *and* $t \in [0,T]$,

$$d_x^p \varphi_t^{\mathrm{v}^n} \to d_x^p \varphi_t^{\mathrm{v}}.$$

Recall that v_n converges to v in the weak topology on $L^2([0,1],\mathcal{B})$ if and only if, for all $\mathrm{w} \in L^2([0,1],\mathcal{B})$,

$$\lim_{n\to\infty} \int_0^1 \int_\Omega \langle \mathrm{v}_n(t)\,,\,\mathrm{w}(t)\rangle_{\mathcal{B}} dt = \int_0^1 \int_\Omega \langle \mathrm{v}(t)\,,\,\mathrm{w}(t)\rangle_{\mathcal{B}} dt.$$

Proof. The proof of this theorem, which is sketched here for completeness, relies on the remark that, since v_n weakly converges, it is bounded in $L^2([0,1],\mathcal{B})$, and Lemma 7 readily implies that (φ^{v_n}) and their space derivatives up to order $p-1$ are equicontinuous sequences in space. Equicontinuity in time comes by applying the Cauchy–Schwarz inequality to

$$d^q \varphi_t^{\mathrm{v}} - d^q \varphi_s^{\mathrm{v}} = \int_s^t d^q(\mathrm{v}_u \circ \varphi_u^{\mathrm{v}}) du.$$

Ascoli's theorem implies compactness of (φ^{v_n}) for the $|.|_{T,p-1,\infty}$-topology, and it remains to identify a limit of any converging subsequence as φ^{v}. Denoting this limit by ψ, one deduces from

$$\varphi_{0,t}^{\mathrm{v}^n}(x) = \int_0^t \mathrm{v}_s^n(\varphi_{0,s}^{\mathrm{v}^n}(x)) ds$$

and the convergence of $\varphi_{0,t}^{\mathrm{v}^n}$ to φ_t the fact that

$$\psi_t(x) = \int_0^t \mathrm{v}_s^n(\psi_s(x)) ds + o(n),$$

and the conclusion comes after the remark that $\mathrm{w} \mapsto \int_0^t \mathrm{w}_s(\psi_s(x)) ds$ is a continuous linear functional on $L^2([0,1],\mathcal{B})$ so that the weak convergence of v^n to v implies that

$$\psi_t(x) = \int_0^t \mathrm{v}_s(\psi_s(x)) ds$$

and $\psi_t = \varphi_{0,t}^{\mathrm{v}}$.

We now prove the pointwise convergence of the pth derivative. We know that

$$\frac{d}{dt} d_x^p \varphi_t^{\mathrm{v}} = d_{\varphi_t^{\mathrm{v}}} \mathrm{v} d_x^p \varphi_t^{\mathrm{v}} + Q_t^{\mathrm{v}}(x),$$

where $Q_t^{\mathrm{v}}(x)$ depends on the derivatives of v evaluated at $\varphi_t^{\mathrm{v}}(x)$ and on the $p-1$ first space derivatives of φ_t^{v}. We may therefore write

$$\begin{aligned} d_x^p \varphi_t^{\mathrm{v}} - d_x^p \varphi_t^{\mathrm{v}^n} = \int_0^t d_{\varphi_s^{\mathrm{v}^n}} \mathrm{v}(d_x^p \varphi_s^{\mathrm{v}} - d_x^p \varphi_s^{\mathrm{v}^n}) ds + \int_0^t (d_{\varphi_s^{\mathrm{v}}} \mathrm{v} - d_{\varphi_s^{\mathrm{v}^n}} \mathrm{v}^n) d_x^p \varphi_s^{\mathrm{v}} ds \\ + \int_0^t (Q_s^{\mathrm{v}}(x) - Q_s^{\mathrm{v}^n}(x)) ds. \end{aligned}$$

The first integral may be bounded by $C(|\mathrm{v}^n|_{1,T})\int_0^t \left|d_x^p\varphi_s^{\mathrm{v}} - d_x^p\varphi_s^{\mathrm{v}^n}\right| ds$, and the result will be a consequence of Gronwall's lemma, provided we show that the remaining terms tend to 0. Consider the second integral, which may be written

$$\int_0^t (d_{\varphi_s^{\mathrm{v}}}\mathrm{v} - d_{\varphi_s^{\mathrm{v}}}\mathrm{v}^n)d_x^p\varphi_s^{\mathrm{v}}ds + \int_0^t (d_{\varphi^{\mathrm{v}_s}}\mathrm{v}^n - d_{\varphi_s^{\mathrm{v}^n}}\mathrm{v}^n)d_x^p\varphi_s^{\mathrm{v}}ds.$$

The first term tends to 0 because

$$w \mapsto \int_0^t d_{\varphi_s^{\mathrm{v}}(x)}w d_x^p\varphi_s^{\mathrm{v}}ds$$

is a continuous linear functional on $L^2([0,t],\mathcal{B})$ and v^n weakly converges to v in this space. To estimate the second one, introduce, for $A, \varepsilon > 0$, the number

$$C(A,\varepsilon) = \max\left\{|d_x w - d_y w| : x, , y \in \Omega, |x-y| \le \varepsilon, |w|_{\mathcal{B}} \le A\right\}.$$

The compact embedding assumption implies that, A being fixed, $C(A,\varepsilon)$ tends to 0 when ε tends to 0. Using this notation, we have

$$\begin{aligned}\int_0^t (d_{\varphi^{\mathrm{v}_s}}\mathrm{v}^n - d_{\varphi_s^{\mathrm{v}^n}}\mathrm{v}^n)d_x^p\varphi_s^{\mathrm{v}}ds &\le \int_0^t C\left(|\mathrm{v}_s^n|_{\mathcal{B}}, |\varphi_s^{\mathrm{v}} - \varphi_s^{\mathrm{v}_n}|_\infty\right)|d_x^p\varphi_s^{\mathrm{v}}|_\infty ds \\ &\le \int_0^t C\left(A, |\varphi_s^{\mathrm{v}} - \varphi_s^{\mathrm{v}_n}|_\infty\right)|\mathrm{v}|_{\mathcal{B}}\, ds,\end{aligned}$$

where $A = \text{ess.sup}\left\{|\mathrm{v}_s^n|_{\mathcal{B}}, n \ge 0, s \in [0,1]\right\}$. The last upper bound now tends to 0, by dominated convergence.

Finally, a generic term of Q_t^{v} being

$$d^k_{\varphi_t^{\mathrm{v}}(x)}\mathrm{v}_t(d^{i_1}\varphi_t^{\mathrm{v}}, \dots, d^{i_k}\varphi_t^{\mathrm{v}}),$$

we can use the same argument to prove its pointwise convergence. ∎

Appendix D. Action of diffeomorphisms on images. The next theorem provides results concerning the regularity of the action of diffeomorphisms on $L^2(\Omega,\mathbb{R}^d)$ and $H^1(\Omega,\mathbb{R}^d)$.

THEOREM 10.

(i) *Let φ be a diffeomorphism of Ω such that φ and φ^{-1} have uniformly bounded first derivatives on Ω. Then, if $i \in L^2(\Omega,\mathbb{R}^d)$ (resp., $i \in H^1(\Omega,\mathbb{R}^d)$), also $i\circ\varphi \in L^2(\Omega,\mathbb{R}^d)$ (resp., $i\circ\varphi \in H^1(\Omega,\mathbb{R}^d)$ and $d_x(i\circ\varphi) = d_{\varphi(x)}i.d_x\varphi$).*

(ii) *Moreover, for all $M > 0$ and for all $i \in L^2(\Omega,\mathbb{R}^d)$, there exists a function $\varepsilon_M(i,\eta)$ such that, for all φ, φ' such that*

$$\max\left(|\varphi|_{1,\infty}, \left|\varphi^{-1}\right|_{1,\infty}, |\varphi'|_{1,\infty}, \left|\varphi'^{-1}\right|_{1,\infty}\right) \le M,$$

we have

$$|i\circ\varphi' - i\circ\varphi|_2 \le \varepsilon_M(i, |\varphi - \varphi'|_{1,\infty}),$$

and $\varepsilon_M(i,\eta) \to 0$ when $\eta \to 0$. The same statement is true for $i \in H^1(\Omega,\mathbb{R}^d)$, the $L^2(\Omega,\mathbb{R}^d)$-norm being replaced by the $H^1(\Omega,\mathbb{R}^d)$-norm.

Proof. We start with (i) and give the proof for $H^1(\Omega, \mathbb{R}^d)$, since it contains exactly the arguments which are valid for $L^2(\Omega, \mathbb{R}^d)$. Fix φ and let L_φ be defined by $L_\varphi(i) = i \circ \varphi$. The vector space $C_d^\infty = C^\infty(\Omega, \mathbb{R}^d)$ of restrictions to Ω of infinitely differentiable functions on $\mathbb{R}^k$ taking values in $\mathbb{R}^d$ is dense in $H^1(\Omega, \mathbb{R}^d)$ [10]. The linear map L_φ is continuous from C_d^∞ (with the topology induced by $H^1(\Omega, \mathbb{R}^d)$) to $H^1(\Omega, \mathbb{R}^d)$; indeed, for $i \in C_d^\infty$,

$$|L_\varphi(i)|_{H^1}^2 = |i \circ \varphi|_2^2 + |d_\varphi i \, d\varphi|_2^2 \leq |i|_2^2 \left| d\varphi^{-1} \right|_\infty + |di|_2^2 |d\varphi|_\infty^2 |d\varphi^{-1}|_\infty \leq C \, |i|_{H^1}^2$$

since the first derivatives of φ and φ^{-1} are bounded. Thus, L_φ restricted to C_d^∞ can be extended to a continuous function $\tilde{L}_\varphi$ on $H^1(\Omega, \mathbb{R}^d)$. If $i \in H^1(\Omega, \mathbb{R}^d)$ and i_n is a sequence of elements of C_d^∞ which converges to i when n tends to infinity (so that $i_n \circ \varphi \to \tilde{L}_\varphi(i)$ in $H^1(\Omega, \mathbb{R}^d)$), then, because convergence in $H^1(\Omega, \mathbb{R}^d)$ implies convergence in $L^2(\Omega, \mathbb{R}^d)$, a subsequence of i_n can be extracted which converges almost everywhere to i and such that $i_n \circ \varphi$ converges almost everywhere to $\tilde{L}_n(i)$. If $N \subset \Omega$ has null Lebesgue measure, then it is also the case for $\varphi^{-1}(N)$ (by boundedness of $\left| d\varphi^{-1} \right|$), so that $i_n \circ \varphi$ also converges almost everywhere to $i \circ \varphi$, yielding $\tilde{L}_\varphi = L_\varphi$. Now, since the map $i \to di$ is obviously continuous from H^1 to L^2, so is $i \to d(L_\varphi(i))$. But, since this map coincides with $i \to d_\varphi i \, d\varphi$ on C_d^∞, and this last map is also continuous on $H^1(\Omega, \mathbb{R}^d)$ (by the previous computation), we get equality over all $H^1(\Omega, \mathbb{R}^d)$, again by density of C_d^∞.

For (ii), we first consider the $L^2(\Omega, \mathbb{R}^d)$ case. Let i, φ', φ, and M be as in the theorem, and fix $s \in C^\infty(\Omega, \mathbb{R}^d)$; we have

$$|i \circ \varphi' - i \circ \varphi|_2 \leq |i \circ \varphi' - s \circ \varphi'|_2 + |s \circ \varphi - s \circ \varphi'|_2 + |i \circ \varphi - s \circ \varphi|_2 .$$

First notice that

$$|i \circ \varphi' - s \circ \varphi'|_2^2 = \int_\Omega \left| d{\varphi'}^{-1} \right| |i - s|^2 dx \leq C \left| {\varphi'}^{-1} \right|_{1,\infty} |i - s|_2^2$$

for some constant C. For the middle term, we have

$$\begin{aligned} |s \circ \varphi - s \circ \varphi'|_2 &\leq \int_0^1 \left| d_{\varphi + t(\varphi' - \varphi)} s (\varphi' - \varphi) \right|_2 dt \\ &\leq |\varphi' - \varphi|_\infty \int_0^1 \left| d_{\varphi + t(\varphi' - \varphi)} s \right|_2 dt \\ &\leq C(M) \, |ds|_2 \, |\varphi' - \varphi|_\infty . \end{aligned}$$

We thus get

$$|i \circ \varphi' - i \circ \varphi|_2 \leq C(M) \left(|i - s|_2 + |ds|_2 \, |\varphi' - \varphi|_\infty \right) .$$

Letting

$$\varepsilon_M(i, \eta) \triangleq C(M) \inf_{s \in C^\infty(\Omega)} \left(|i - s|_2 + |ds|_2 \, \eta \right)$$

yields the conclusion of the theorem in the $L^2(\Omega, \mathbb{R}^d)$ case, the $H^1(\Omega, \mathbb{R}^d)$ case being handled similarly. $\square$

Appendix E. Proof of Lemma 2. We must compute the derivative at $\varepsilon = 0$ of

$$U^\varepsilon = \frac{1}{\sigma^2}\int_\Omega \frac{\left|j_0\circ\varphi_{1,0}^{\mathrm{v}+\varepsilon h} - j_1\right|^2}{\int_0^1 |d\varphi_{1,s}^{\mathrm{v}+\varepsilon h}|^{-1}ds}dx.$$

First, we notice the equation

$$\sigma^2 z_t(x) = \frac{j_1\circ\varphi_{t,1}^{\mathrm{v}} - j_0\circ\varphi_{t,0}^{\mathrm{v}}}{\int_0^1 |d\varphi_{t,s}^{\mathrm{v}}|^{-1}ds}, \tag{49}$$

which implies that (differentiating at $\varepsilon = 0$)

$$\frac{dU^\varepsilon}{d\varepsilon} = -2\left\langle z_1\,,\, d_{\varphi_{1,0}^{\mathrm{v}}}j_0\frac{d}{d\varepsilon}\varphi_{1,0}^{\mathrm{v}+\varepsilon h}\right\rangle_2 - \sigma^2\int_0^1\left\langle |z_1|^2\,,\,\frac{d}{d\varepsilon}\left|d\varphi_{1,s}^{\mathrm{v}+\varepsilon h}\right|^{-1}\right\rangle_2 ds.$$

Starting with the first term and using Lemma 10, we have

$$\begin{aligned}\left\langle z_1\,,\, d_{\varphi_{1,0}^{\mathrm{v}}}j_0\frac{d}{d\varepsilon}\varphi_{1,0}^{\mathrm{v}+\varepsilon h}\right\rangle_2 &= -\int_0^1\left\langle d_{\varphi_{1,0}^{\mathrm{v}}}j_0^* z_1\,,\, d_{\varphi_{1t}^{\mathrm{v}}}\varphi_{t,0}^{\mathrm{v}}h_t\circ\varphi_{1t}^{\mathrm{v}}\right\rangle_2 dt\\ &= -\int_0^1\left\langle d_{\varphi_{t,0}^{\mathrm{v}}}j_0^* z_1\circ\varphi_{t1}^{\mathrm{v}}\,|d\varphi_{t1}^{\mathrm{v}}|\,,\, d\varphi_{t,0}^{\mathrm{v}}h_t\right\rangle_2 dt\\ &= -\int_0^1\left\langle (d\varphi_{t,0}^{\mathrm{v}})^* d_{\varphi_{t,0}^{\mathrm{v}}}j_0^* z_t\,,\, h_t\right\rangle_2 dt\\ &= -\int_0^1\left\langle K\left((d\varphi_{t,0}^{\mathrm{v}})^* d_{\varphi_{t,0}^{\mathrm{v}}}j_0^* z_t\right)\,,\, h_t\right\rangle_{\mathcal{B}} dt\end{aligned}$$

because of the identity $z_t = z_1\circ\varphi_{t1}^{\mathrm{v}}\,|d\varphi_{t1}^{\mathrm{v}}|$.

We now pass to the second term, for which we use the equality

$$\left|d\varphi_{t,s}^{\mathrm{v}+\varepsilon h}\right|^{-1} = \exp\left[\int_s^t \operatorname{div}(\mathrm{v}_u+\varepsilon h_u)\circ\varphi_{t,u}^{\mathrm{v}+\varepsilon h}du\right],$$

which is a consequence of Lemma 7 and standard computations on the derivative of the determinant. This implies that

$$\begin{aligned}\frac{d}{d\varepsilon}\left(\left|d\varphi_{t,s}^{\mathrm{v}+\varepsilon h}\right|^{-1}\right) &= \left|d\varphi_{t,s}^{\mathrm{v}}\right|^{-1}\int_s^t \operatorname{div}(h_u)\circ\varphi_{t,u}^{\mathrm{v}}du\\ &\quad+\left|d\varphi_{t,s}^{\mathrm{v}}\right|^{-1}\int_s^t d_{\varphi_{t,u}^{\mathrm{v}}}\operatorname{div}(\mathrm{v}_u)\int_t^u d_{\varphi_{t\tau}^{\mathrm{v}}}\varphi_{\tau u}^{\mathrm{v}}h_\tau\circ\varphi_{t\tau}^{\mathrm{v}}d\tau du\\ &= \left|d\varphi_{t,s}^{\mathrm{v}}\right|^{-1}\int_s^t \operatorname{div}(h_u)\circ\varphi_{t,u}^{\mathrm{v}}du\\ &\quad-\left|d\varphi_{t,s}^{\mathrm{v}}\right|^{-1}\int_s^t\int_s^\tau d_{\varphi_{t,u}^{\mathrm{v}}}\operatorname{div}(\mathrm{v}_u)d_{\varphi_{t\tau}^{\mathrm{v}}}\varphi_{\tau u}^{\mathrm{v}}h_\tau\circ\varphi_{t\tau}^{\mathrm{v}}du d\tau.\end{aligned}$$

We may notice that

$$\left\langle\nabla\,|d\varphi_{\tau s}^{\mathrm{v}}|^{-1}\,,\,\xi\right\rangle = |d\varphi_{\tau s}^{\mathrm{v}}|^{-1}\int_s^\tau d_{\varphi_{\tau u}^{\mathrm{v}}}(\operatorname{div}v_u)d\varphi_{\tau u}^{\mathrm{v}}(\xi)du$$

to identify the last term as

$$\left|d\varphi^{\mathrm{v}}_{t,s}\right|^{-1}\int_s^t \left|d_{\varphi^{\mathrm{v}}_{t\tau}}\varphi^{\mathrm{v}}_{\tau s}\right|\left\langle\nabla_{\varphi^{\mathrm{v}}_{t\tau}}\left|d\varphi^{\mathrm{v}}_{\tau s}\right|^{-1},\, h_\tau\circ\varphi^{\mathrm{v}}_{t\tau}\right\rangle d\tau$$

so that

$$\begin{aligned}\frac{d}{d\varepsilon}\left(\left|d\varphi^{\mathrm{v}+\varepsilon h}_{t,s}\right|^{-1}\right) &= \left|d\varphi^{\mathrm{v}}_{t,s}\right|^{-1}\int_s^t \operatorname{div}(h_u)\circ\varphi^{\mathrm{v}}_{t,u}du \\ &\quad - \int_s^t \left|d\varphi^{\mathrm{v}}_{t\tau}\right|^{-1}\left\langle\nabla_{\varphi^{\mathrm{v}}_{t\tau}}\left|d\varphi^{\mathrm{v}}_{\tau s}\right|^{-1},\, h_\tau\circ\varphi^{\mathrm{v}}_{t\tau}\right\rangle d\tau.\end{aligned}$$

Therefore,

$$\begin{aligned}\int_0^1\left\langle |z_1|^2,\, \frac{d}{d\varepsilon}\left|d\varphi^{\mathrm{v}+\varepsilon h}_{1,s}\right|^{-1}\right\rangle_2 ds &= \int_0^1\int_s^1\left\langle |z_1|^2,\, \left|d\varphi^{\mathrm{v}}_{1s}\right|^{-1}\operatorname{div}(h_u)\circ\varphi^{\mathrm{v}}_{1u}\right\rangle_2 duds \\ &\quad -\int_0^1\int_s^1\left\langle |z_1|^2,\, \left|d\varphi^{\mathrm{v}}_{1u}\right|^{-1}\left\langle\nabla_{\varphi^{\mathrm{v}}_{1u}}\left|d\varphi^{\mathrm{v}}_{us}\right|^{-1},\, h_u\circ\varphi^{\mathrm{v}}_{1u}\right\rangle\right\rangle_2 duds \\ &= \int_0^1\int_s^1\left\langle |z_u|^2,\, \left|d\varphi^{\mathrm{v}}_{u1}\right|^{-1}\left|d_{\varphi^{\mathrm{v}}_{u1}}\varphi^{\mathrm{v}}_{1s}\right|^{-1}\operatorname{div}(h_u)\right\rangle_2 duds \\ &\quad -\int_0^1\int_s^1\left\langle |z_u|^2,\, \left\langle\nabla\left|d\varphi^{\mathrm{v}}_{us}\right|^{-1},\, h_u\right\rangle\right\rangle_2 duds.\end{aligned}$$

Introducing

$$q^{\mathrm{v}}_u \triangleq \int_0^u \left|d\varphi^{\mathrm{v}}_{us}\right|^{-1}ds,$$

this may be written

$$\begin{aligned}\int_0^1\left\langle |z_1|^2,\, \frac{d}{d\varepsilon}\left|d\varphi^{\mathrm{v}+\varepsilon h}_{1,s}\right|^{-1}\right\rangle_2 ds &= \int_0^1\left\langle q^{\mathrm{v}}_u\,|z_u|^2,\, \operatorname{div}(h_u)\right\rangle_2 du \\ &\quad -\int_0^1\left\langle |z_u|^2,\, \langle\nabla q^{\mathrm{v}}_u,\, h_u\rangle\right\rangle_2 du \\ &= -\int_0^1\left\langle K_\nabla(q^{\mathrm{v}}_u\,|z_u|^2),\, h_u\right\rangle_{\mathcal{B}} du \\ &\quad -\int_0^1\left\langle K(|z_u|^2\,\nabla q^{\mathrm{v}}_u),\, h_u\right\rangle_{\mathcal{B}} du.\end{aligned}$$

Now, defining functions

$$C^{\mathrm{v}}_t \triangleq \sigma^2 q^{\mathrm{v}}_t\,|z_t|^2 \tag{50}$$

and

$$D^{\mathrm{v}}_t \triangleq \sigma^2\,|z_t|^2\,\nabla q^{\mathrm{v}}_t + 2[d\varphi^{\mathrm{v}}_{t,0}]^* d_{\varphi^{\mathrm{v}}_{t,0}} j_0^* z_t, \tag{51}$$

Proposition 5 implies

$$\frac{dU^\varepsilon}{d\varepsilon} = \int_0^1 \langle h_t\,,\, K.D^{\mathrm{v}}_t + K_\nabla C^{\mathrm{v}}_t\rangle_{\mathcal{B}} dt,$$

which is the conclusion of Lemma 2.

Appendix F. Proof of Lemma 4. We prove that solutions of system (22) travel at constant speed and therefore compute the derivative of $|v_t|^2_{\mathcal{B}} + \sigma^2 |z_t|^2_2$ for such a solution. Starting with the second term, we have $z_t = z_0 \circ \varphi^{\mathrm{v}}_{t,0} \left| d\varphi^{\mathrm{v}}_{t,0} \right|$, which implies, after a change of variables,

$$|z_t|^2_2 = \int_\Omega |z_0|^2_2 \left| d\varphi^{\mathrm{v}}_{0,t} \right|^{-1} dx.$$

Using the identity

$$\left| d\varphi^{\mathrm{v}}_{s,t} \right| = \exp\left(\int_s^t \operatorname{div}(v_u) \circ \varphi^{\mathrm{v}}_{s,u} du \right), \tag{52}$$

we obtain

$$\frac{d}{dt} |z_t|^2_2 = - \int_\Omega |z_0|^2_2 \left| d\varphi^{\mathrm{v}}_{0,t} \right|^{-1} \operatorname{div}(v_t) \circ \varphi^{\mathrm{v}}_{0,t} dx. \tag{53}$$

To study the variation of $|\mathrm{v}_t|^2_{\mathcal{B}}$, we start with the computation of the derivative of $\langle \mathrm{v}_t \,,\, w \rangle$ for a fixed $w \in \mathcal{B}$. Applying formula (28) for a solution of (22) yields

$$\begin{aligned}\langle \mathrm{v}_t \,,\, w \rangle_{\mathcal{B}} = {} & \frac{\sigma^2}{2} \int_0^1 \Big\langle |z_0|^2 \,,\, (\left| d\varphi^{\mathrm{v}}_{0,s} \right|^{-1} \operatorname{div}(w) \circ \varphi^{\mathrm{v}}_{0,t} - \left| d\varphi^{\mathrm{v}}_{0,t} \right|^{-1} \left\langle \nabla \xi^{\mathrm{v}}_{s,t} \,,\, \lambda^{\mathrm{v}}_t(w) \right\rangle) \Big\rangle ds \\ & + (\omega_0, \lambda^{\mathrm{v}}_t(w))\end{aligned}$$

with $\xi^{\mathrm{v}}_{s,t} = \left| d\varphi^{\mathrm{v}}_{0,t} \right| / \left| d\varphi^{\mathrm{v}}_{0,s} \right|$ and $\lambda^{\mathrm{v}}_t(w) = (d\varphi^{\mathrm{v}}_{0,t})^{-1} w \circ \varphi^{\mathrm{v}}_{0,t}$. From formula (52), we have

$$\xi^{\mathrm{v}}_{s,t} = \exp\left(\int_s^t (\operatorname{div} v_u) \circ \varphi^{\mathrm{v}}_{0u} du \right),$$

which implies that

$$d\xi^{\mathrm{v}}_{s,t} = \xi^{\mathrm{v}}_{s,t} \int_s^t d_{\varphi^{\mathrm{v}}_{0u}} (\operatorname{div} v_u) d\varphi^{\mathrm{v}}_{0u} du$$

so that

$$\begin{aligned}\langle \mathrm{v}_t \,,\, w \rangle_{\mathcal{B}} &= (\omega_0, \lambda^{\mathrm{v}}_t(w)) + \frac{\sigma^2}{2} \int_0^t \Big\langle |z_0|^2 \,,\, \left| d\varphi^{\mathrm{v}}_{0,s} \right|^{-1} \operatorname{div}(w) \circ \varphi^{\mathrm{v}}_{0,t} \Big\rangle ds \\ &\quad - \frac{\sigma^2}{2} \int_0^t \int_s^t \Big\langle |z_0|^2 \,,\, \left| d\varphi^{\mathrm{v}}_{0,s} \right|^{-1} d_{\varphi^{\mathrm{v}}_{0u}} (\operatorname{div} v_u) d\varphi^{\mathrm{v}}_{0u} \lambda^{\mathrm{v}}_t(w) \Big\rangle du ds \\ &= (\omega_0, \lambda^{\mathrm{v}}_t(w)) + \frac{\sigma^2}{2} \int_0^t \Big\langle |z_0|^2 \,,\, \left| d\varphi^{\mathrm{v}}_{0,s} \right|^{-1} \operatorname{div}(w) \circ \varphi^{\mathrm{v}}_{0,t} \Big\rangle ds \\ &\quad - \frac{\sigma^2}{2} \int_0^t \int_0^u \Big\langle |z_0|^2 \,,\, \left| d\varphi^{\mathrm{v}}_{0,s} \right|^{-1} d_{\varphi^{\mathrm{v}}_{0u}} (\operatorname{div} v_u) d\varphi^{\mathrm{v}}_{0u} \lambda^{\mathrm{v}}_t(w) \Big\rangle ds du.\end{aligned}$$

We now compute the time differential of each term which appears in this expression. Denote $\overline{\lambda}^{\mathrm{v}}_t(w) = \frac{d}{dt} \lambda^{\mathrm{v}}_t(w)$. We have

$$\overline{\lambda}^{\mathrm{v}}_t(w) = \frac{d}{dt} \left((d\varphi^{\mathrm{v}}_{0,t})^{-1} w \circ \varphi_{0,t} \right) = -(d\varphi^{\mathrm{v}}_{0,t})^{-1} d_{\varphi^{\mathrm{v}}_{0,t}} v_t w \circ \varphi_{0,t} + (d\varphi^{\mathrm{v}}_{0,t})^{-1} d_{\varphi^{\mathrm{v}}_{0,t}} w v_t \circ \varphi_{0,t}.$$

Next, we have

$$\frac{d}{dt}\int_0^t \left\langle |z_0|^2 \,,\, \left|d\varphi^{\mathrm{v}}_{0,s}\right|^{-1} \mathrm{div}(w)\circ\varphi^{\mathrm{v}}_{0,t}\right\rangle ds = \left\langle |z_0|^2 \,,\, \left|d\varphi^{\mathrm{v}}_{0,t}\right|^{-1} \mathrm{div}(w)\circ\varphi^{\mathrm{v}}_{0,t}\right\rangle + \int_0^t \left\langle |z_0|^2 \,,\, \left|d\varphi^{\mathrm{v}}_{0,s}\right|^{-1} \nabla_{\varphi^{\mathrm{v}}_{0,t}}(\mathrm{div}(w))v_t\circ\varphi^{\mathrm{v}}_{0,t}\right\rangle ds$$

and

$$\begin{aligned}\frac{d}{dt}\int_0^t\int_0^u &\left\langle |z_0|^2 \,,\, \left|d\varphi^{\mathrm{v}}_{0,s}\right|^{-1} d_{\varphi^{\mathrm{v}}_{0u}}(\mathrm{div}v_u)d\varphi^{\mathrm{v}}_{0u}\lambda^{\mathrm{v}}_t(w)\right\rangle dsdu\\ &= \int_0^t \left\langle |z_0|^2 \,,\, \left|d\varphi^{\mathrm{v}}_{0,s}\right|^{-1} d_{\varphi^{\mathrm{v}}_{0,t}}(\mathrm{div}v_t)d\varphi^{\mathrm{v}}_{0,t}\lambda^{\mathrm{v}}_t(w)\right\rangle dsdu\\ &+ \int_0^t\int_0^u \left\langle |z_0|^2 \,,\, \left|d\varphi^{\mathrm{v}}_{0,s}\right|^{-1} d_{\varphi^{\mathrm{v}}_{0u}}(\mathrm{div}v_u)d\varphi^{\mathrm{v}}_{0u}\overline{\lambda}^{\mathrm{v}}_t(w)\right\rangle dsdu.\end{aligned}$$

Putting everything together, we have

$$\begin{aligned}\frac{d}{dt}\langle \mathrm{v}_t \,,\, w\rangle_{\mathcal{B}} &= \left(\omega_0, \overline{\lambda}^{\mathrm{v}}_t(w)\right) + \frac{\sigma^2}{2}\left\langle |z_0|^2 \,,\, \left|d\varphi^{\mathrm{v}}_{0,t}\right|^{-1} \mathrm{div}(w)\circ\varphi^{\mathrm{v}}_{0,t}\right\rangle\\ &+ \frac{\sigma^2}{2}\int_0^t \left\langle |z_0|^2 \,,\, \left|d\varphi^{\mathrm{v}}_{0,s}\right|^{-1} d_{\varphi^{\mathrm{v}}_{0,t}}(\mathrm{div}w)v_t\circ\varphi^{\mathrm{v}}_{0,t}\right\rangle dsdu\\ &- \frac{\sigma^2}{2}\int_0^t \left\langle |z_0|^2 \,,\, \left|d\varphi^{\mathrm{v}}_{0,s}\right|^{-1} d_{\varphi^{\mathrm{v}}_{0,t}}(\mathrm{div}v_t)d\varphi^{\mathrm{v}}_{0,t}\lambda^{\mathrm{v}}_t(w)\right\rangle dsdu\\ &- \frac{\sigma^2}{2}\int_0^t\int_0^u \left\langle |z_0|^2 \,,\, \left|d\varphi^{\mathrm{v}}_{0,s}\right|^{-1} d_{\varphi^{\mathrm{v}}_{0u}}(\mathrm{div}v_u)d\varphi^{\mathrm{v}}_{0u}\overline{\lambda}^{\mathrm{v}}_t(w)\right\rangle dsdu.\end{aligned}$$

A little care must be taken in writing, as we did above, $\frac{d}{dt}(\omega_0, \lambda^{\mathrm{v}}_t(w)) = (\omega_0\overline{\lambda}^{\mathrm{v}}_t(w))$, since this requires proving that $(\lambda^{\mathrm{v}}_{t+\varepsilon}(w) - \lambda^{\mathrm{v}}_t(w))/\varepsilon$ converges to $\overline{\lambda}^{\mathrm{v}}_t(w)$ for the $(p-1,\infty)$-norm. This is indeed true in our case, because of the fact that $w\in\mathcal{B}$ allows us to control the uniform norm of its differentials up to order p, and the differentials of φ^{v}_t up to the same order are solutions of a linear differential equation which ensures their uniform continuity.

We now use the identity (which is justified below)

$$\frac{d}{dt}\left|\mathrm{v}_t\right|^2_{\mathcal{B}} = 2\lim_{\varepsilon\to 0}\langle \mathrm{v}_{t+\varepsilon} - \mathrm{v}_t \,,\, \mathrm{v}_t\rangle_{\mathcal{B}}/\varepsilon, \tag{54}$$

which implies that, to compute the time differential of $\left|\mathrm{v}_t\right|^2_{\mathcal{B}}$, it suffices to use the obtained expression for the derivative of $\langle \mathrm{v}_t \,,\, w\rangle_{\mathcal{B}}$ with $w = \mathrm{v}_t$ and multiply it by 2. Since $\overline{\lambda}^{\mathrm{v}}_t(v_t) = 0$, and because of (53), we see that all terms cancel, yielding $\frac{d}{dt}(\left|\mathrm{v}_t\right|^2_{\mathcal{B}} + \sigma^2\left|z_t\right|^2_2) = 0$.

To show (54), one writes

$$(\left|\mathrm{v}_t\right|^2_{\mathcal{B}} - \left|\mathrm{v}_t\right|^2_{\mathcal{B}} - \langle \mathrm{v}_{t+\varepsilon} - \mathrm{v}_t \,,\, \mathrm{v}_t\rangle_{\mathcal{B}})/\varepsilon = \left|\mathrm{v}_{t+\varepsilon} - \mathrm{v}_t\right|^2_{\mathcal{B}}/\varepsilon,$$

and the result is obtained by proving that, for $w\in\mathcal{B}$,

$$\left|\langle \mathrm{v}_{t+\varepsilon} - \mathrm{v}_t \,,\, w\rangle_{\mathcal{B}}\right| = O(\varepsilon)\left|w\right|_{\mathcal{B}},$$

which can be done by a direct estimation of $\frac{d}{dt}\langle \mathrm{v}_t \,,\, w\rangle_{\mathcal{B}}$.

Appendix G. Proof of Lemma 5. It suffices to prove this result for smooth $z_0, \tilde{z}_0, j_0$. It is straightforward that $M_{j_0}(tz_0) = \mathrm{j}_t$, where j is the solution of (21) with initial conditions (j_0, z_0). Let $\tilde{\mathrm{j}}_t = M_{j_0}(t\tilde{z}_0)$. Introduce also the corresponding $(\mathrm{v}_t, \mathrm{z}_t)$ and $(\tilde{\mathrm{v}}_t, \tilde{\mathrm{z}}_t)$.

Introduce the notation $\eta = \tilde{\mathrm{j}} - \mathrm{j}$, $\zeta = \tilde{\mathrm{z}} - \mathrm{z}$, and $\alpha = \tilde{\mathrm{v}} - \mathrm{v}$. Since we have assumed smooth trajectories, we may write

$$\frac{\partial \mathrm{j}_t}{\partial t} = \sigma^2 \mathrm{z}_t - d\mathrm{j}_t \mathrm{v}_t$$

and

$$\frac{\partial \mathrm{z}_t}{\partial t} = -\mathrm{div}(\mathrm{z}_t \otimes \mathrm{v}_t)$$

and similar equations for the trajectory with initial condition $(j_0, \tilde{z}_0)$. Computing the differences along both trajectories yields

$$\left\{ \begin{array}{l} \frac{\partial \eta_t}{\partial t} + d\eta_t \tilde{\mathrm{v}}_t = \sigma^2 \zeta_t - d\mathrm{j}_t \alpha_t, \\ \\ \frac{\partial \zeta_t}{\partial t} + \mathrm{div}(\zeta_t \otimes \tilde{\mathrm{v}}_t) = -\mathrm{div}(\mathrm{z}_t \otimes \alpha_t). \end{array} \right. \tag{55}$$

Since

$$\begin{aligned} \frac{\partial}{\partial t}\left[\left|d\varphi^{\tilde{\mathrm{v}}}_{0,t}\right| \zeta_t \circ \varphi^{\tilde{\mathrm{v}}}_{0,t}\right] &= \left|d\varphi^{\tilde{\mathrm{v}}}_{0,t}\right| \left(\frac{\partial \zeta_t}{\partial t} + \mathrm{div}(\zeta_t \otimes \tilde{\mathrm{v}}_t)\right) \circ \varphi^{\tilde{\mathrm{v}}}_{0,t} \\ &= -\left|d\varphi^{\tilde{\mathrm{v}}}_{0,t}\right| \mathrm{div}(\mathrm{z}_t \otimes \alpha_t) \circ \varphi^{\tilde{\mathrm{v}}}_{0,t}, \end{aligned}$$

the second term yields

$$\zeta_s \circ \varphi^{\tilde{\mathrm{v}}}_{0,s} = \left|d\varphi^{\tilde{\mathrm{v}}}_{0,s}\right|^{-1} (\tilde{z}_0 - z_0) - \left(\int_0^s \left|d\varphi^{\tilde{\mathrm{v}}}_{s,u}\right| \mathrm{div}(\mathrm{z}_u \otimes \alpha_u) \circ \varphi^{\tilde{\mathrm{v}}}_{s,u} du\right) \circ \varphi^{\tilde{\mathrm{v}}}_{0,s},$$

and the first one implies

$$\eta_t \circ \varphi^{\tilde{\mathrm{v}}}_{0,t} = \sigma^2 \int_0^t \zeta_s \circ \varphi^{\tilde{\mathrm{v}}}_{0,s} ds - \int_0^t [d\mathrm{j}_s \alpha_s] \circ \varphi^{\tilde{\mathrm{v}}}_{0,s} ds.$$

Replacing ζ in the last equation gives

(56)
$$\begin{aligned} \eta_t \circ \varphi^{\tilde{\mathrm{v}}}_{0,t} &= t[\sigma^2(\tilde{z}_0(.) - z_0(.)) - dj_0(\tilde{v}_0 - v_0)] \\ &\quad - \sigma^2 \int_0^t \left(\int_0^s \left|d\varphi^{\tilde{\mathrm{v}}}_{s,u}\right| \mathrm{div}(\mathrm{z}_u \otimes \alpha_u) \circ \varphi^{\tilde{\mathrm{v}}}_{s,u} du\right) \circ \varphi^{\tilde{\mathrm{v}}}_{0,s} ds \\ &\quad - \int_0^t \left\{[d\mathrm{j}_s \alpha_s] \circ \varphi^{\tilde{\mathrm{v}}}_{0,s} - dj_0(\tilde{v}_0 - v_0)\right\} ds + \int_0^t \left(\left|d\varphi^{\tilde{\mathrm{v}}}_{0,s}\right|^{-1} - 1\right) (\tilde{z}_0 - z_0) ds \end{aligned}$$

so that Lemma 5 reduces to evaluating the L^2-norm of the last three integrals. We shall use the fact that, for a function $f \in L^2([0,1] \times \Omega, \mathbb{R}^d)$,

$$\left|\int_0^t f_s ds\right|_2 \leq \int_0^t |f_s|_2 \, ds.$$

For $t \in [0,1]$, we also have, from (31), with $\omega_0 = dj_0^* z_0$ and $\omega_0' = dj_0^* z_0'$ (here and in the following, we denote by const any quantity which depends only on j_0, z_0 and $\tilde{z}_0$),

$$|\alpha_t|_{\mathcal{B}} \le \text{const } |z_0 - \tilde{z}_0|_2 \,. \tag{57}$$

This implies that

$$\begin{aligned}
&\left|\int_0^t \left(\int_0^s \left|d\varphi^{\tilde{\mathrm{v}}}_{s,u}\right| \operatorname{div}(\mathrm{z}_u \otimes \alpha_u) \circ \varphi^{\tilde{\mathrm{v}}}_{s,u} du\right) \circ \varphi^{\tilde{\mathrm{v}}}_{0,s} ds\right|_2 \\
&\quad \le \int_0^t \int_0^s \left|\left|d_{\varphi^{\tilde{\mathrm{v}}}_{0,s}} \varphi^{\tilde{\mathrm{v}}}_{s,u}\right| \operatorname{div}(\mathrm{z}_u \otimes \alpha_u) \circ \varphi^{\tilde{\mathrm{v}}}_{0,u}\right|_2 du ds \\
&\quad = \int_0^t \int_0^s \left|\left|d\varphi^{\tilde{\mathrm{v}}}_{0,s}\right|^{-1} \operatorname{div}(\mathrm{z}_u \otimes \alpha_u)\right|_2 ds \\
&\quad \le \text{const } \int_0^t \int_0^s |\operatorname{div}(\mathrm{z}_u \otimes \alpha_u)|_2 \, du ds \\
&\quad \le \text{const } \int_0^t \int_0^s |\mathrm{z}_u|_{H^1} \, |\alpha_u|_{\mathcal{B}} \, du ds.
\end{aligned}$$

The relation $\mathrm{z}_t = z_0 \circ \varphi^{\mathrm{v}}_{t,0} \left|d\varphi^{\mathrm{v}}_{t,0}\right|$ implies that $|\mathrm{z}_u|_{H^1} \le \text{const } |z_0|_{H^1}$ so that

$$\left|\int_0^t \left(\int_0^s \left|d\varphi^{\tilde{\mathrm{v}}}_{s,u}\right| \operatorname{div}(\mathrm{z}_u \otimes \alpha_u) \circ \varphi^{\tilde{\mathrm{v}}}_{s,u} du\right) \circ \varphi^{\tilde{\mathrm{v}}}_{0,s} ds\right|_2 \le \text{const } t^2 \, |\tilde{z}_0 - z_0|_2 \,.$$

A similar estimate is valid for the last integral in (56), since $||d\varphi^{\mathrm{v}}_{0,s}|^{-1} - 1|_\infty \le \text{const } s$. We finally consider the second integral in this equation.

Since

$$\mathrm{j}_s = j_0 \circ \varphi^{\mathrm{v}}_{s,0} + \sigma^2 z_0 \circ \varphi^{\mathrm{v}}_{s,0} \int_0^s |d\varphi^{\mathrm{v}}_{u0}| \circ \varphi^{\mathrm{v}}_{s,u} du,$$

we have, letting $\gamma_s = \varphi^{\mathrm{v}}_{s,0} \circ \varphi^{\tilde{\mathrm{v}}}_{0,s}$,

$$d_{\varphi^{\tilde{\mathrm{v}}}_{0,s}} \mathrm{j}_s \, \alpha \circ \varphi^{\tilde{\mathrm{v}}}_{0,s} = d_{\gamma_s} j_0 \, d_{\varphi^{\tilde{\mathrm{v}}}_{0,s}} \varphi^{\mathrm{v}}_{s,0} \, \alpha \circ \varphi^{\tilde{\mathrm{v}}}_{0,s} + R_s,$$

and it is easy to check that $|R_t|_2 \le \text{const } t \, |z_0|_{H^1} \, |\tilde{z}_0 - z_0|_2$. We need to estimate

$$\begin{aligned}
&\int_0^t \left(d_{\gamma_s} j_0 \, d_{\varphi^{\tilde{\mathrm{v}}}_{0,s}} \varphi^{\mathrm{v}}_{s,0} \, \alpha_s \circ \varphi^{\tilde{\mathrm{v}}}_{0,s} - dj_0 \alpha_0\right) ds \\
&\qquad = \int_0^t \left(d_{\gamma_s} j_0 (d_{\varphi^{\tilde{\mathrm{v}}}_{0,s}} \varphi^{\mathrm{v}}_{s,0} \, \alpha_s \circ \varphi^{\tilde{\mathrm{v}}}_{0,s} - \alpha_0)\right) ds + \int_0^t (d_{\gamma_s} j_0 \, \alpha_0 - dj_0 \, \alpha_0) \, ds.
\end{aligned} \tag{58}$$

Start with the first term, for which we must bound, for the L^∞-norm, the difference $d_{\varphi^{\tilde{\mathrm{v}}}_{0,s}} \varphi^{\mathrm{v}}_{s,0} \, \alpha \circ \varphi^{\tilde{\mathrm{v}}}_{0,s} - \alpha_0$ or, equivalently,

$$d\varphi^{\mathrm{v}}_{s,0} \alpha_s - \alpha_0 \circ \varphi^{\tilde{\mathrm{v}}}_{s,0}.$$

It is simple to check, from (57) and estimates we have used several times on the variations of the diffeomorphisms, that $(d\varphi^{\mathrm{v}}_{s,0} - I)\alpha_s$ and $\alpha_0 \circ \varphi^{\tilde{\mathrm{v}}}_{s,0} - \alpha_0$ are bounded by const $s \, |\tilde{z}_0 - z_0|_2$. We now proceed to an upper bound for $\alpha_s - \alpha_0$, for which we need to return to the expression obtained in (28), which yields

$$\begin{aligned}
\langle \mathrm{v}_s - \mathrm{v}_0, w\rangle_{\mathcal{B}} = \frac{\sigma^2}{2} \int_0^s \Big\langle |z_0|^2 \,, (|d\varphi^{\mathrm{v}}_{0u}|^{-1} \operatorname{div}(w) \circ \varphi^{\mathrm{v}}_{0,s} \\
- \left|d\varphi^{\mathrm{v}}_{0,s}\right|^{-1} \langle \nabla \xi^{\mathrm{v}}_{us} \,, \, \lambda^{\mathrm{v}}_s(w)\rangle)\Big\rangle \, du + \langle z_0 \,, \, dj_0(\lambda^{\mathrm{v}}_s(w) - w)\rangle_2
\end{aligned}$$

so that

$$\begin{aligned}\langle \alpha_s - \alpha_0\,,\, w\rangle_{\mathcal{B}} &= \frac{\sigma^2}{2}\int_0^s \Big\langle \left|\tilde z_0\right|^2,\, (\left|d\varphi^{\tilde{\mathrm{v}}}_{0u}\right|^{-1}\operatorname{div}(w)\circ\varphi^{\tilde{\mathrm{v}}}_{0,s} - \left|d\varphi^{\tilde{\mathrm{v}}}_{0,s}\right|^{-1}\left\langle \nabla\xi^{\tilde{\mathrm{v}}}_{us}\,,\,\lambda^{\tilde{\mathrm{v}}}_s(w)\right\rangle)\Big\rangle du\\ &\quad- \frac{\sigma^2}{2}\int_0^s \Big\langle \left|z_0\right|^2,\, (\left|d\varphi^{\mathrm{v}}_{0u}\right|^{-1}\operatorname{div}(w)\circ\varphi^{\mathrm{v}}_{0,s} - \left|d\varphi^{\mathrm{v}}_{0,s}\right|^{-1}\left\langle \nabla\xi^{\mathrm{v}}_{us}\,,\,\lambda^{\mathrm{v}}_s(w)\right\rangle)\Big\rangle du\\ &\quad+ \left\langle \tilde z_0\,,\, dj_0(\lambda^{\tilde{\mathrm{v}}}_s(w)-w)\right\rangle_2 - \left\langle z_0\,,\, dj_0(\lambda^{\mathrm{v}}_s(w)-w)\right\rangle_2.\end{aligned}$$

The difference of the first two integrals takes the form

$$\frac{\sigma^2}{2}\int_0^s \left(\left\langle \tilde z_0\,,\, Q^{\tilde{\mathrm{v}}}_{us}(w)\right\rangle - \left\langle z_0\,,\, Q^{\mathrm{v}}_{us}(w)\right\rangle\right) du \tag{59}$$

with $Q^{\mathrm{v}}_{us}(w) = \left|d\varphi^{\mathrm{v}}_{0u}\right|^{-1}\operatorname{div}(w)\circ\varphi^{\mathrm{v}}_{0,s} - \left|d\varphi^{\mathrm{v}}_{0,s}\right|^{-1}\left\langle\nabla\xi^{\mathrm{v}}_{us}\,,\,\lambda^{\mathrm{v}}_s(w)\right\rangle$. From Lemmas 7 and 11, and from (57), we obtain the fact that $\left|Q^{\tilde{\mathrm{v}}}_{us}(w) - Q^{\mathrm{v}}_{us}(w)\right| \le \text{const }\left|\tilde z_0 - z_0\right|_2 \left|w\right|_{\mathcal{B}}$ so that the quantity in (59) is bounded by const $s\left|\tilde z_0 - z_0\right|_2$. Writing

$$\begin{aligned}\left\langle \tilde z_0\,,\, dj_0(\lambda^{\tilde{\mathrm{v}}}_s(w)-w)\right\rangle_2 - \left\langle z_0\,,\, dj_0(\lambda^{\mathrm{v}}_s(w)-w)\right\rangle_2 &= \left\langle \tilde z_0 - z_0\,,\, dj_0(\lambda^{\tilde{\mathrm{v}}}_s(w)-w)\right\rangle_2\\ &\quad+ \left\langle z_0\,,\, dj_0(\lambda^{\tilde{\mathrm{v}}}_s(w)-\lambda^{\mathrm{v}}_s(w))\right\rangle_2\end{aligned}$$

and using $\left|\lambda^{\tilde{\mathrm{v}}}_s(w) - w\right|_\infty \le$ const s (which is deduced from Lemma 7 and a computation of the differential of $\lambda^{\tilde{\mathrm{v}}}_s(w)$ with respect to s) and $\left|\lambda^{\tilde{\mathrm{v}}}_s(w) - \lambda^{\mathrm{v}}_s(w)\right|_\infty \le$ const $s\left|\tilde z_0 - z_0\right|_2$ (from Lemma 11 and (57)), we finally conclude that

$$\left|d_{\varphi^{\tilde{\mathrm{v}}}_{0,s}}\varphi^{\mathrm{v}}_{s,0}\,\alpha\circ\varphi^{\tilde{\mathrm{v}}}_{0,s} - \alpha_0\right|_\infty \le \text{const } s\left|\tilde z_0 - z_0\right|_2,$$

which implies that the first integral in the right-hand term of (58) is bounded by const $t^2\left|\tilde z_0 - z_0\right|_2$.

Consider now the last term of (58), namely,

$$\int_0^t (dj_0\circ\gamma_s - dj_0)\,\alpha_0 ds.$$

Since $\left|\alpha_0\right|_\infty \le C\left|\tilde z_0 - z_0\right|_2$, we must estimate $\left|d_{\gamma_s}j_0 - dj_0\right|_2$. By Theorem 10, this is a function of the kind

$$\varepsilon_M(dj_0, \left|\gamma_s - \mathrm{Id}\right|_\infty) = \varepsilon_M(dj_0, \left|\varphi^{\tilde{\mathrm{v}}}(s) - \varphi^{\mathrm{v}}(s)\right|_\infty),$$

where M depends only on $\left|j_0\right|_{H^1}, \left|z_0\right|_2, \left|\tilde z_0\right|_2$. Since $|\varphi^{\tilde{\mathrm{v}}}_{0,s} - \varphi^{\mathrm{v}}_{0,s}|_\infty = O(s)$, we get (with another function ε)

$$\int_0^t (dj_0\circ\gamma_s - dj_0)\,\alpha_0 ds \le \varepsilon(j_0,t)t\left|\tilde z_0 - z_0\right|_2.$$

We need finally to consider the last line of (56) which can be easily bounded from above by $\varepsilon(j_0,t)t\left|\tilde z_0 - z_0\right|_2$. We now can collect the estimates we have obtained to conclude the proof of Lemma 5.

REFERENCES

[1] Y. Amit, U. Grenander, and M. Piccioni, *Structural image restoration through deformable templates*, J. Amer. Statist. Assoc., 86 (1989), pp. 376–387.

[2] Y. Amit and P. Piccioni, *A non-homogeneous Markov process for the estimation of Gaussian random fields with nonlinear observations*, Ann. Probab., 19 (1991), pp. 1664–1678.

[3] I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1989.
[4] V. Arnold and B. Khesin, *Topological methods in hydrodynamics*, in Annual Review Fluid of Mechanics, 24, Annual Review, Palo Alto, CA, 1992, pp. 145–166.
[5] V. I. Arnold, *Sur la géometrie differentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits*, Ann. Inst. Fourier (Grenoble), 1 (1966), pp. 319–361.
[6] R. Bajcsy and C. Broit, *Matching of deformed images*, in Proceedings of the 6th International Conference on Pattern Recognition, Munich, Germany, 1982, pp. 351–353.
[7] R. Bajcsy and S. Kovacic, *Multiresolution elastic matching*, Comp. Vision, Graphics, and Image Proc., 46 (1989), pp. 1–21.
[8] F. L. Bookstein, *Principal warps: Thin plate splines and the decomposition of deformations*, IEEE Trans. Pattern Anal. Mach. Intell., 11 (1989), pp. 567–585.
[9] F. L. Bookstein, *Morphometric Tools for Landmark Data; Geometry and Biology*, Cambridge University Press, Cambridge, UK, 1991.
[10] H. Brezis, *Analyse Fonctionnelle, Théorie et Applications*, Masson, Paris, 1983.
[11] R. Camassa and D. D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.
[12] M. P. D. Carmo, *Riemannian Geometry*, Birkhaüser, Basel, 1992.
[13] E. Christensen, R. D. Rabbitt, and M. I. Miller, *Deformable templates using large deformation kinematics*, IEEE Trans. Image Process., 1996, pp. 1435–1447.
[14] T. Cootes, C. Taylor, D. Cooper, and J. Graham, *Active shape models: Their training and application*, Comp. Vis. and Image Understanding, 61 (1995), pp. 38–59.
[15] C. David and S. W. Zucker, *Potentials, valleys, and dynamic global coverings*, Int. J. of Comp. Vision, 5 (1990), pp. 219–238.
[16] P. Dupuis, U. Grenander, and M. Miller, *Variational problems on flows of diffeomorphisms for image matching*, Quart. Appl. Math., 56 (1998), pp. 587–600.
[17] C. Foias, D. D. Holm, and E. S. Titi, *The Navier-Stokes-alpha model of fluid turbulence*, Phys. D, 152 (2001), pp. 505–519.
[18] D. Geman and S. Geman, *Stochastic relaxation, gibbs distribution and bayesian restoration of images*, IEEE Trans. Pattern Anal. Mach. Intell., 6 (1984), pp. 721–741.
[19] U. Grenander, *Lectures in Pattern Theory*, Appl. Math. Sci. 33, Springer-Verlag, New York, 1981.
[20] U. Grenander, *General Pattern Theory*, Oxford Science Publications, Oxford University Press, New York, 1993.
[21] U. Grenander and D. M. Keenan, *Towards automated image understanding*, J. Appl. Stat., 16 (1989), pp. 207–221.
[22] U. Grenander and D. M. Keenan, *On the shape of plane images*, SIAM J. Appl. Math., 53 (1993), pp. 1072–1094.
[23] U. Grenander and M. I. Miller, *Representations of knowledge in complex systems (with discussion section)*, J. Roy. Statist. Soc. Ser. B, 56 (1994), pp. 549–603.
[24] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Springer-Verlag, New York, 1994.
[25] J. E. Marsden, T. Ratiu, and A. Weinstein, *Semidirect products and reduction in mechanics*, Trans. Amer. Math. Soc., 281 (1984), pp. 147–177.
[26] M. I. Miller, S. C. Joshi, and G. E. Christensen, *Large deformation fluid diffeomorphisms for landmark and image matching*, in Brain Warping, A. Toga, ed., Academic Press, New York, 1999, pp. 115–131.
[27] M. I. Miller, and L. Younes, *Group action, diffeomorphism and matching: A general framework*, Int. J. Comp. Vis, 41 (2001), pp. 61–84.
[28] F. Murat and J. Simon, *Étude de problèmes d'optimal design*, in Optimization Techniques: Modeling and Optimization in the Service of Man, Lecture Notes in Comput. Sci. 41, Springer-Verlag, New York, 1975, pp. 54–62.
[29] B. D. Ripley and A. I. Sutherland, *Finding spiral structures in images of galaxies*, Phil. Trans. Roy. Soc. A, 332 (1990), pp. 477–485.
[30] A. Trouvé, *Infinite Dimensional Group Action and Pattern Recognition*, Tech. report, DMI, Ecole Normale Supérieure, Paris, France, 1995.
[31] A. Trouvé, *Diffeomorphism groups and pattern matching in image analysis*, Int. J. Comp. Vis., 28 (1998), pp. 213–221.
[32] L. Younes, *Computable elastic distances between shapes*, SIAM J. Appl. Math., 58 (1998), pp. 565–586.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 60–82

MATHEMATICAL ANALYSIS OF A NONLINEAR PARABOLIC EQUATION ARISING IN THE MODELLING OF NON-NEWTONIAN FLOWS*

ERIC CANCÈS†, ISABELLE CATTO‡, AND YOUSRA GATI†

Abstract. The mathematical properties of a nonlinear parabolic equation arising in the modelling of concentrated suspension flows are investigated. The peculiarity of this equation is that it may degenerate into a hyperbolic equation (in fact, a linear advection equation). Depending on the initial data, at least two situations can be encountered: the equation may have a unique solution in a convenient class, or it may have infinitely many solutions. The present article is the theoretical side of a joint project with rheologists, aiming at better understanding the flows of complex fluids.

Key words. nonlinear parabolic equation, degenerate parabolic equation, viscosity solutions, non-Newtonian flows, complex fluids, concentrated suspensions

AMS subject classifications. 35K50, 35K55, 35K65, 35Q35, 76A05

DOI. 10.1137/S0036141003430044

1. Introduction and description of the model.

Complex fluids are substances that are neither really liquid nor really solid in the classical sense. They include melt polymers, colloids, emulsions, foams, gels, liquid crystals, suspensions, and other materials that form flowable microstructures. Modelling the flow of such fluids is a very intricate problem which is far from being solved up to now. The model we are interested in is an attempt to recover the rheological behavior of the particular type of concentrated suspensions. Examples of such suspensions are numerous and can be found, e.g., in food (pastes), cosmetics (tooth paste), medicine (blood), and building industry (cement). In contrast to some complex fluids such as polymeric liquids for which elaborate rheological models, based on fine mesoscopic physical descriptions, are available, the modelling of concentrated suspensions is still in its infancy.

When simple fluids are sheared, stress and shear rate are linked by a linear relation. The linear response coefficients are well understood and their relation to the microstructure of the fluid are known [7]. On the contrary, complex systems, such as concentrated suspensions of hard or soft spheres, exhibit very nonlinear flow properties far from being understood. These nonlinear properties occur not only at high shear rates, where one expects that linear response theory does not apply, but also at very low shear rates. Let us for instance consider a concentrated suspension of particles. At low concentrations, thermally induced structural relaxations of the particles positions occur. The system behaves as a Newtonian fluid at low shear rates[1]. But, when the concentration is increased, the energetical cost of particles reorganizations

*Received by the editors June 19, 2003; accepted for publication (in revised form) November 5, 2004; published electronically August 17, 2005.
http://www.siam.org/journals/sima/37-1/43004.html
†CERMICS, Ecole Nationale des Ponts et Chaussées and INRIA, 6 & 8 avenue Blaise Pascal, Cité Descartes, 77455 Marne-la-Vallée Cedex 2, France (cances@cermics.enpc.fr, gati@cermics.enpc.fr).
‡CEREMADE, UMR CNRS 7534, Université Paris IX-Dauphine, Place du Maréchal de Lattre de Tassigny, F-75775 Paris Cedex 16, France (catto@ceremade.dauphine.fr).
[1]At higher shear rates, due to hydrodynamics interactions, the relation between the stress and the shear rate may be sublinear, and the system is said to be shear-thinning. We neglect all hydrodynamics interactions in the following description.

is much higher than thermal energy, and structural relaxations are arrested. This implies that, once at rest, the system is not at thermal equilibrium, and exhibits general properties of glasses [1, 11, 13, 14]. This property has a striking consequence on the low shear rate flow of these systems: the stress σ tends to a finite nonzero value σ_c when shear rate goes to zero. This discontinuity of the shear stress *versus* shear rate is an experimental very common feature, but is very poorly understood. One of the most common physical explanations relies on the hypothesis that the system may locally store deformation. It relies on observations of the structures of concentrated dispersions that appear to be frozen in nonequilibrium positions: particular configurations of the dynamically arrested suspension store deformation energy. They do not correspond to configurations of minimal energy, and may thus store finite values of stress and strain. The system may thus be described as an heterogeneous field of stress and strain. Microscopic description of these fields does not exist, and, besides studies which aim at improving phenomenological models (such as the celebrated Herschel–Bulkley model [9]), a few attempts have been made to recover the rheological behavior of complex fluids from elementary physical processes. We consider here the model proposed by Hébraud and Lequeux, in which the system is divided in mesoscopic blocks whose size is large enough so that stress and strain tensors may be defined for each block, but small compared to the characteristic length scale of the stress field. A mesoscopic evolution equation of the stress of each block is then written as:

(i) At low shear, each particle keeps the same neighbors, and a block behaves as an Einstein elastic solid, in which the elasticity arises from interactions between neighboring particles.

(ii) Then, deformation induces local reorganization of the particles, at a given stress threshold σ_c. Above this threshold, the block flows as an Eyring fluid: the configuration reached by shearing the suspension relaxes with a characteristic time T_0 towards a completely relaxed state, where no stress is stored.

(iii) Lastly, coupling between the flow of neighboring blocks must be included. This is taken into account by the introduction of a diffusion term in the evolution equation, where it is assumed that the diffusion coefficient is proportional to the number of reorganizations per unit time.

In the model, each block carries a given shear stress σ (σ is a real number; it is in fact an extra-diagonal term of the stress tensor in convenient coordinates). The evolution of the blocks is described through a probability density $p(t,\sigma)$ which represents the distribution of stress in the assembly of blocks at time t. The equation for the probability density $p(t,\sigma)$ for a block to be under stress σ at time t is written as

$$\begin{cases} \partial_t p = -b(t)\partial_\sigma p + D(p(t))\partial^2_{\sigma\sigma} p - \dfrac{\mathbb{1}_{\mathbb{R}\setminus[-\sigma_c,\sigma_c]}(\sigma)}{T_0} p + \dfrac{D(p(t))}{\alpha}\delta_0(\sigma), \\ t \in (0;T), \quad \sigma \in \mathbb{R}, \\ p \geq 0, \\ p(0,\sigma) = p_0(\sigma), \end{cases} \tag{1.1}$$

where for $f \in L^1(\mathbb{R})$, we denote

$$D(f) = \frac{\alpha}{T_0}\int_{|\sigma|>\sigma_c} f(\sigma)d\sigma. \tag{1.2}$$

The equation satisfied by p in (1.1) is referred to as the SP equation in the following. In this equation, $\mathbb{1}_{\mathbb{R}\setminus[-\sigma_c,\sigma_c]}$ denotes the characteristic function of the open set $\mathbb{R}\setminus[-\sigma_c,\sigma_c]$ and δ_0 the Dirac delta function on $\mathbb{R}$. The three terms arising from the right-hand side of the HL equation model the three physical features described above. When a block is submitted to a shear rate $\dot\gamma(t)$, the stress of this block evolves with a variation rate $b(t) = G_0\dot\gamma(t)$, where G_0 is an elasticity constant. (In this study, the shear rate $\dot\gamma(t)$, and therefore the function $b(t)$, are assumed to be in $L^2_{\mathrm{loc}}(\mathbb{R}^+)$.) When the modulus of the stress overcomes the critical value σ_c, the block becomes unstable and may relax into a state with zero stress after a characteristic relaxation time T_0. This property is expressed by the last two terms in (1.1). This relaxation phenomenon induces a rearrangement of the other blocks and this is finally modelled through the diffusion term $D(p(t))\partial^2_{\sigma\sigma}p$. The diffusion coefficient $D(p(t))$ as given by (1.2) is assumed to be proportional to the density of blocks that rearrange during time T_0, by a proportional parameter α which depends on the microscopic properties of the sample and which is supposed to represent the "mechanical fragility" of the material. This nonlinear diffusion term is introduced to display the importance of collective effects in this kind of samples. For more details on the physical meaning of the model, we refer to the original article by Hébraud and Lequeux [8].

In all that follows, the parameters α, T_0 and σ_c are positive, and the initial data p_0 in (1.1) is a given probability density; that is,

$$p_0 \geq 0, \quad p_0 \in L^1(\mathbb{R}), \quad \int_{\mathbb{R}} p_0 = 1. \tag{1.3}$$

We will be looking for solutions $p = p(t,\sigma)$ in $C^0_t(L^1_\sigma \cap L^2_\sigma)$ such that σp belongs to $L^\infty_t(L^1_\sigma)$ of the nonlinear parabolic partial differential equation (1.1). The subscript σ refers to integration over $\mathbb{R}$ with respect to σ, whereas the subscript t refers to time integration on $[0,T]$ for any $T > 0$. Note that under a mean-field assumption the macroscopic stress in the material is given by

$$\tau(t) = \int_{\mathbb{R}} \sigma p(t,\sigma)d\sigma, \tag{1.4}$$

and therefore the above condition on σp ensures that the average stress is an essentially bounded function of time.

Actually in practice, the shear rate is not uniform in the flow (the shear creates elastic waves in the fluid), and in order to better describe the coupling between the macroscopic flow and the evolution of the microstructure we introduce and study in a second paper [2] a micro-macro model where the shear rate is a function of the velocity of the macroscopic flow. In this model p is also a function of the macroscopic space variables and the average stress defined by (1.4) is inserted into the macroscopic equation governing the velocity of the macroscopic flow (see also section 6 below).

In order to lighten the notation and without loss of generality we assume from now on that $\sigma_c = 1$ and $T_0 = 1$. This amounts to changing the time and stress scales.

The main difficulties one encounters in the mathematical analysis come from the nonlinearity in the diffusion term and even more from the fact that the parabolic equation may degenerate when the viscosity coefficient $D(p)$ vanishes, and this will be shown to appear only when $D(p_0) = 0$. This difficulty is illustrated on a simplified example just below and also in section 5 where we discuss the existence of stationary solutions in the case when the shear rate b is a constant.

Let us first of all look at the following simplified model which already includes the difficulties we are going to face to in the study of (1.1). We consider the equation

$$\begin{cases} \partial_t u = D(u(t))\partial^2_{\sigma\sigma} u, \\ u(0,\sigma) = \frac{1}{2}\mathbb{1}_{]-1,1[}(\sigma), \end{cases} \tag{1.5}$$

where $\mathbb{1}_{]-1,1[}$ is the characteristic function of the interval $]-1,1[$. The initial condition is purposely chosen in such a way that $D(u(t=0)) = 0$. The function $u = \frac{1}{2}\mathbb{1}_{]-1,1[}(\sigma)$ is a stationary solution to this equation and for this solution $D(u(t))$ is identically zero. But it is not the unique solution to (1.5) in $C^0_t(L^2_\sigma) \cap L^\infty_t(L^1_\sigma)$. It is indeed possible to construct a so-called *vanishing viscosity solution* for which $D(u(t)) > 0$ for all $t > 0$, and there are actually infinitely many solutions to this equation. (This statement is obtained as a corollary of Lemma 4.3 in section 4 below.)

As far as (1.1) is concerned, we show that, in the case when $D(p_0) = 0$ and $b \equiv 0$, we may have either a unique or infinitely many solutions, depending on the initial data (see Proposition 4.1 in section 4).

On the other hand, we are able to prove the following existence and uniqueness result in the nondegenerate case when $D(p_0) > 0$.

THEOREM 1.1. *Let the initial data p_0 satisfy the conditions*

$$p_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R}), \quad p_0 \geq 0, \int_{\mathbb{R}} p_0 = 1, \text{ and } \int_{\mathbb{R}} |\sigma| p_0 < +\infty, \tag{1.6}$$

and assume that

$$D(p_0) > 0.$$

Then, for every $T > 0$, there exists a unique solution p to (1.1) *in $L^\infty_t(L^1_\sigma \cap L^2_\sigma) \cap L^2_t(H^1_\sigma)$. Moreover, $p \in L^\infty_{t,\sigma} \cap C^0_t(L^1_\sigma \cap L^2_\sigma)$, $\int_{\mathbb{R}} p(t,\sigma)d\sigma = 1$ for all $t > 0$, $D(p) \in C^0_t$ and for every $T > 0$ there exists a positive constant $\nu(T)$ such that*

$$\min_{0 \leq t \leq T} D(p(t)) \geq \nu(T).$$

Besides $\sigma p \in L^\infty_t(L^1_\sigma)$ so that the average stress $\tau(t)$ is well-defined by (1.4) *in L^∞_t.*

The first step towards the existence proof of solutions to (1.1) will consist of the study of so-called vanishing viscosity approximations, which are the unique solutions to the following family of equations:

$$\begin{cases} \partial_t p_\varepsilon = -b(t)\partial_\sigma p_\varepsilon + (D(p_\varepsilon(t)) + \varepsilon)\partial^2_{\sigma\sigma} p_\varepsilon - \mathbb{1}_{\mathbb{R}\setminus[-1,1]} p_\varepsilon + \dfrac{D(p_\varepsilon(t))}{\alpha}\delta_0(\sigma), \\ p_\varepsilon \geq 0, \\ p_\varepsilon(0,\cdot) = p_0. \end{cases} \tag{1.7}$$

(Recall that we have rescaled the time and stress units to get $T_0 = 1$ and $\sigma_c = 1$.) Section 2 below is devoted to the proof of the following proposition.

PROPOSITION 1.2 (existence and uniqueness of vanishing viscosity approximations). *Let $T > 0$ be given. We assume that the initial data satisfies the same conditions* (1.6) *as in the statement of the theorem. Then, for every $T > 0$ and*

$0 < \varepsilon \leq 1$, there exists a unique solution p_ε to (1.7) *in $L^\infty_t(L^1_\sigma \cap L^2_\sigma) \cap L^2_t(H^1_\sigma)$. Moreover, $p_\varepsilon \in L^\infty_{t,\sigma} \cap C^0_t(L^1_\sigma \cap L^2_\sigma)$, $D(p_\varepsilon) \in C^0_t$,*

$$\int_{\mathbb{R}} p_\varepsilon = 1, \tag{1.8}$$

$$0 \leq p_\varepsilon \leq \|p_0\|_{L^\infty_\sigma} + \sqrt{\frac{\alpha}{\pi}}\sqrt{T}, \tag{1.9}$$

and for every $T > 0$, there exist positive constants $C_1(T, p_0)$, $C_2(T, p_0)$, and $C_3(T, p_0)$ which are independent of ε such that

$$\sup_{0 \leq t \leq T} \int_{\mathbb{R}} |\sigma| p_\varepsilon \leq C_1(T, p_0), \tag{1.10}$$

$$\sup_{0 \leq t \leq T} \int_{\mathbb{R}} p_\varepsilon^2 \leq C_2(T, p_0), \tag{1.11}$$

and

$$\int_0^T \big(\varepsilon + D(p_\varepsilon)\big) \int_{\mathbb{R}} |\partial_\sigma p_\varepsilon|^2 \leq C_3(T, p_0). \tag{1.12}$$

Theorem 1.1 is then proved in section 3 while the degenerate case is investigated in section 4. Lastly, the description of stationary solutions in the constant shear rate case is carried out in section 5.

2. The vanishing viscosity approximation. This section is devoted to the proof of Proposition 1.2. We begin with the following lemma.

LEMMA 2.1 (uniqueness). *Let p_0 satisfy* (1.3). *Then for every $T > 0$ and $0 < \varepsilon$, there exists at most one solution p_ε to* (1.7) *in $L^\infty_t(L^1_\sigma \cap L^2_\sigma) \cap L^2_t(H^1_\sigma)$. Moreover, $p_\varepsilon \in C^0_t(L^2_\sigma)$ (thus, the initial condition makes sense) and*

$$\int_{\mathbb{R}} p_\varepsilon = 1, \tag{2.1}$$

for almost every t in $[0, T]$.

Proof. We begin by proving that every solution to (1.7) in $L^\infty_t(L^1_\sigma \cap L^2_\sigma) \cap L^2_t(H^1_\sigma)$ satisfies (2.1). We fix $R > 1$ and we consider a cut-off C^2 function $\phi_R = \phi_R(\sigma)$ with compact support which is equal to 1 when $0 \leq |\sigma| \leq R$ and to 0 when $|\sigma| \geq 2R$ and such that

$$|\phi_R'| \leq \frac{C}{R}, \tag{2.2}$$

where here and below C denotes a positive constant that is independent of R. Notice that ϕ' is equal to 0 on $]-\infty, -2R]$, on $[-R, R]$ and on $[2R, +\infty[$.

Now, we multiply (1.7) by ϕ_R and integrate over $[0, t] \times \mathbb{R}$ to obtain

$$\begin{aligned}
&\int_{\mathbb{R}} p_\varepsilon(t)\phi_R - \int_{\mathbb{R}} p_0\phi_R \\
&\quad = -\int_0^t b(s) \int_{\mathbb{R}} \partial_\sigma p_\varepsilon(s)\phi_R - \int_0^t (D(p_\varepsilon(s)) + \varepsilon) \int_{\mathbb{R}} \partial_\sigma p_\varepsilon(s)\phi_R' \\
&\qquad - \int_0^t \int_{|\sigma|>1} p_\varepsilon(s)\phi_R + \frac{1}{\alpha}\int_0^t D(p_\varepsilon(s))\phi_R(0).
\end{aligned}$$

We bound the terms on the right-hand side from above as follows. First, we have

$$\begin{aligned}\left|\int_0^t b(s)\int_{\mathbb{R}}\partial_\sigma p_\varepsilon(s)\phi_R\right| &\le \int_0^t |b(s)|\int_{\mathbb{R}} p_\varepsilon(s)|\phi_R'|\\ &\le \frac{C}{R}\int_0^t |b(s)|\int_{R\le|\sigma|\le 2R} p_\varepsilon(s) \le \frac{C}{R},\end{aligned}$$

thanks to (2.2) and using that $p_\varepsilon \in L_t^\infty(L_\sigma^1)$ and $b \in L_t^1$. Next,

$$\begin{aligned}\int_0^t (D(p_\varepsilon)+\varepsilon)\left|\int_{\mathbb{R}}\partial_\sigma p_\varepsilon \phi'\right| &\le (\varepsilon + \alpha\|p_\varepsilon\|_{L_t^\infty(L_\sigma^1)})\int_0^t \|\partial_\sigma p_\varepsilon\|_{L_\sigma^2}\|\phi_R'\|_{L_\sigma^2}\\ &\le \frac{C\sqrt{t}}{R^{1/2}}\|\partial_\sigma p_\varepsilon\|_{L_{t,\sigma}^2} \le \frac{C}{R^{1/2}},\end{aligned}$$

thanks again to (2.2), the Cauchy–Schwarz inequality and since $\partial_\sigma p_\varepsilon$ is in $L_{t,\sigma}^2$. Finally,

$$\begin{aligned}0 \le \frac{1}{\alpha}\int_0^t D(p_\varepsilon) - \int_0^t\int_{|\sigma|>1} p_\varepsilon\phi_R &= \int_0^t\int_{|\sigma|>1} p_\varepsilon(1-\phi_R)\\ &\le \int_0^t\int_{|\sigma|>R} p_\varepsilon,\end{aligned}$$

and the right-hand side goes to 0 as R goes to infinity since p_ε is in $L_t^\infty(L_\sigma^1)$. All this together yields

$$\int_{\mathbb{R}} p_\varepsilon(t) = \lim_{R\to+\infty}\int_{\mathbb{R}} p_\varepsilon(t)\phi_R = \lim_{R\to+\infty}\int_{\mathbb{R}} p_0\phi_R = \int_{\mathbb{R}} p_0 = 1,$$

for almost every t in $[0,T]$. In particular, this implies that $D(p_\varepsilon)\le\alpha$.

Let us now argue by contradiction by assuming that there exist two solutions p_1 and p_2 to (1.7) corresponding to the same initial data p_0. By subtracting the equations satisfied by p_1 and p_2, respectively, we obtain

$$\begin{cases}\partial_t q = -b(t)\partial_\sigma q + D(q)\partial_{\sigma\sigma}^2 p_1 + (D(p_2)+\varepsilon)\partial_{\sigma\sigma}^2 q\\ \qquad -\mathbb{1}_{\mathbb{R}\setminus[-1,1]}q + \dfrac{D(q)}{\alpha}\delta_0(\sigma),\\ q(0,\sigma)=0,\end{cases}\tag{2.3}$$

where $q = p_1 - p_2$. We multiply (2.3) by q and integrate over $\mathbb{R}$ with respect to σ to obtain, after integrations by parts,

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}} q^2 + (D(p_2)+\varepsilon)\int_{\mathbb{R}}|\partial_\sigma q|^2 + \int_{|\sigma|>1} q^2\\ &\quad = \frac{D(q)}{\alpha}q(t,0) - D(q)\int_{\mathbb{R}}\partial_\sigma p_1\partial_\sigma q.\end{aligned}\tag{2.4}$$

We first remark that since $\int_{\mathbb{R}} p_1 = \int_{\mathbb{R}} p_2 = 1$ thanks to (2.1), we get

$$|D(q)| = \alpha\left|\int_{|\sigma|<1} q\right| \le \alpha\sqrt{2}\|q\|_{L_\sigma^2},$$

with the help of the Cauchy–Schwarz inequality. Next, using the Sobolev embedding of $H^1(\mathbb{R})$ into $L^\infty(\mathbb{R})$, we bound the terms on the right-hand side from above in the following way:

$$\begin{aligned}
&\left|\frac{D(q)}{\alpha}q(t,0) - D(q)\int_{\mathbb{R}} \partial_\sigma p_1 \partial_\sigma q\right| \\
&\quad \le \sqrt{2}\|q\|_{L^2_\sigma}\|q\|_{L^\infty_\sigma} + \sqrt{2}\alpha\|q\|_{L^2_\sigma}\int_{\mathbb{R}} |\partial_\sigma p_1 \partial_\sigma q| \\
&\quad \le \sqrt{2}\|q\|_{L^2_\sigma}\left(\|q\|^2_{L^2_\sigma} + \|\partial_\sigma q\|^2_{L^2_\sigma}\right)^{\frac{1}{2}} + \sqrt{2}\alpha\|q\|_{L^2_\sigma}\|\partial_\sigma p_1\|_{L^2_\sigma}\|\partial_\sigma q\|_{L^2_\sigma} \\
&\quad \le \frac{1}{\varepsilon}\|q\|^2_{L^2_\sigma} + \frac{\alpha^2}{\varepsilon}\|q\|^2_{L^2_\sigma}\|\partial_\sigma p_1\|^2_{L^2_\sigma} + \frac{\varepsilon}{2}\|q\|^2_{L^2_\sigma} + \varepsilon\|\partial_\sigma q\|^2_{L^2_\sigma}.
\end{aligned}$$

Therefore, comparing with (2.4) we deduce that

$$\frac{1}{2}\frac{d}{dt}\|q\|^2_{L^2_\sigma} \le \left(\frac{1}{\varepsilon} + \frac{\alpha^2}{\varepsilon}\|\partial_\sigma p_1\|^2_{L^2_\sigma} + \frac{\varepsilon}{2}\right)\|q\|^2_{L^2_\sigma}.$$

Finally, by applying the Gronwall lemma, we prove that $\|q\|^2_{L^2_\sigma} \le 0$, thus $q = 0$. The uniqueness of the solution follows. ∎

Remark 2.2. The same proof shows that if there exists a solution to (1.1) in $L^\infty_t(L^1_\sigma \cap L^2_\sigma) \cap L^2_t(H^1_\sigma)$ such that $\inf_{0\le t\le T} D(p(t)) > 0$, then it is unique in this space.

We now turn to the existence part in the statement of Proposition 1.2. From now on we fix a positive constant $\varepsilon \le 1$. The proof of Proposition 1.2 will be carried out by the Schauder fixed point theorem. For given positive constants $M(\ge \varepsilon)$ and R, we introduce $\mathcal{D}_{\varepsilon,M}$ and Y_R two closed convex subsets of, respectively, L^2_t and $L^2_{t,\sigma}$ as follows:

$$\begin{aligned}
\mathcal{D}_{\varepsilon,M} &= \left\{a \in L^2_t;\ \varepsilon \le a \le M\right\} \\
Y_R &= \left\{p \in L^2_{t,\sigma};\ p \ge 0,\ \sup_{0\le t\le T}\int_{\mathbb{R}} |\sigma| p \le R\right\}.
\end{aligned}$$

To simplify notation we denote

$$\begin{cases} \varphi_\eta(x) = \dfrac{1}{\sqrt{2\pi}\eta}\exp\left(-\dfrac{x^2}{2\eta^2}\right) & \text{if } \eta > 0, \\ \varphi_0 = \delta_0. \end{cases}$$

We first prove the following proposition.

Proposition 2.3. *Let $T > 0$ and let $p_0 \in L^2(\mathbb{R})$ such that $p_0 \ge 0$. Then, for every a in $\mathcal{D}_{\varepsilon,M}$ and q in Y_R, there exists a unique solution p in $L^\infty_t(L^2_\sigma) \cap L^2_t(H^1_\sigma)$ to*

$$\begin{cases} \partial_t p(t,\sigma) = -b(t)\partial_\sigma p(t,\sigma) + a(t)\partial^2_{\sigma\sigma} p(t,\sigma) \\ \qquad\qquad - \mathbb{1}_{\mathbb{R}\setminus[-1,1]}(\sigma)p(t,\sigma) + \dfrac{D(q)}{\alpha}\delta_0(\sigma), \\ p(0,\sigma) = p_0(\sigma). \end{cases} \tag{2.5}$$

Moreover, $p \in C^0_t(L^2_\sigma)$, p is nonnegative and

$$p_- \le p \le p_+, \tag{2.6}$$

with

$$p_-(t,\sigma) = e^{-t}\int_{-\infty}^{+\infty} p_0(\sigma')\varphi_{\sqrt{2\int_0^t a}}(\sigma-\sigma'-\chi(t))d\sigma' \tag{2.7}$$

and

$$\begin{aligned} p_+(t,\sigma) = &\int_{-\infty}^{+\infty} p_0(\sigma')\varphi_{\sqrt{2\int_0^t a}}(\sigma-\sigma'-\chi(t))d\sigma' \\ &+\frac{1}{\alpha}\int_0^t D(q(s))\varphi_{\sqrt{2\int_s^t a}}(\sigma-\chi(t)+\chi(s))ds, \end{aligned} \tag{2.8}$$

where $\chi(t) = \int_0^t b(s)ds$. *In addition,*

(i) *If* $p_0 \in L^\infty(\mathbb{R})$, *then* p *is in* $L^\infty_{t,\sigma}$ *and*

$$0 \le p \le \|p_0\|_{L^\infty} + \frac{R\sqrt{T}}{\sqrt{\pi}\sqrt{\varepsilon}}. \tag{2.9}$$

(ii) *If* $\int_{\mathbb{R}} |\sigma|p_0 < +\infty$ *(thus* $p_0 \in L^1(\mathbb{R})$*), then* $|\sigma|p \in L^\infty_t(L^1_\sigma)$. *More precisely, we have*

$$\begin{aligned} \sup_{0\le t\le T}\int_{\mathbb{R}}|\sigma|p \le &\int_{\mathbb{R}}|\sigma|p_0 + \sqrt{T}\|b\|_{L^2(0,T)}\|p_0\|_{L^1} + \frac{2R}{3}T^{3/2}\|b\|_{L^2(0,T)} \\ &+\frac{2}{\sqrt{\pi}}(MT)^{1/2}\|p_0\|_{L^1} + \frac{4R\sqrt{M}}{3\sqrt{\pi}}T^{3/2}. \end{aligned} \tag{2.10}$$

Moreover, $p \in C^0_t(L^1_\sigma)$ *and* $D(p) \in C^0_t$.

Proof. Let us first observe that for every q in Y_R, $D(q) \in L^\infty_t$ since

$$0 \le D(q(t)) \le \alpha\int_{|\sigma|>1}|\sigma|q \le \alpha R, \tag{2.11}$$

for almost every t in $[0,T]$. Therefore the source term $D(q(t))\delta_0(\sigma)$ in (2.5) is in $L^\infty_t(H^{-1}_\sigma)$ and the existence and uniqueness of a solution $p \in C^0_t(L^2_\sigma)\cap L^2_t(H^1_\sigma)$ to the system (2.5) is well known (see, for example, [4]). In particular, the initial condition makes sense. Owing to the fact that the source term is nonnegative, the proof that $p \ge 0$ is also standard (see again [4]).

We now check the pointwise inequality (2.6).

This is ensured by the maximum principle with observing that p_- and p_+ given, respectively, by (2.7) and (2.8), are the unique solutions to the systems

$$\begin{cases} \partial_t p_- = -b\partial_\sigma p_- + a\partial^2_{\sigma\sigma}p_- - p_-, \\ p_-(0,\sigma) = p_0(\sigma), \end{cases}$$

and

$$\begin{cases} \partial_t p_+ = -b\partial_\sigma p_+ + a\partial^2_{\sigma\sigma}p_+ + \dfrac{D(q)}{\alpha}\delta_0(\sigma), \\ p_+(0,\sigma) = p_0(\sigma), \end{cases}$$

respectively. We now turn to the proof of statement (i), and assume that p_0 belongs to $L^\infty(\mathbb{R})$. Then, using the two facts that for every $\nu > 0$, $\int_{\mathbb{R}} \varphi_\nu = 1$ and $\varphi_\nu \leq \frac{1}{\sqrt{2\pi\nu}}$, (2.9) is easily deduced from $p \leq p_+$ with the help of (2.11) and since $a \geq \varepsilon$.

Suppose now that $\int_{\mathbb{R}} |\sigma| p_0 < +\infty$. This together with the assumption $p_0 \in L^2(\mathbb{R})$, guarantees that $p_0 \in L^1(\mathbb{R})$ (see also below). Using (2.6) again, we now have

$$
\begin{aligned}
\int_{\mathbb{R}} |\sigma| p &\leq \int_{\mathbb{R}} |\sigma| p_+ \\
&\leq \int_{\mathbb{R}} \int_{\mathbb{R}} p_0(\sigma') |\sigma| \varphi_{\sqrt{2\int_0^t a}}(\sigma - \chi(t) - \sigma') d\sigma d\sigma' \\
&\quad + \frac{1}{\alpha} \int_0^t D(q(s)) \Big(\int_{\mathbb{R}} |\sigma| \varphi_{\sqrt{2\int_s^t a}}(\sigma - \chi(t) + \chi(s)) d\sigma \Big) ds \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} p_0(\sigma') |\sigma + \sigma' + \chi(t)| \varphi_{\sqrt{2\int_0^t a}}(\sigma) d\sigma d\sigma' \\
&\quad + \frac{1}{\alpha} \int_0^t D(q(s)) \Big(\int_{\mathbb{R}} |\sigma + (\chi(t) - \chi(s))| \varphi_{\sqrt{2\int_s^t a}}(\sigma) d\sigma \Big) ds \\
&\leq \int_{\mathbb{R}} |\sigma| p_0(\sigma) d\sigma + |\chi(t)| \|p_0\|_{L^1} + \frac{1}{\alpha} \int_0^t |\chi(t) - \chi(s)| D(q(s)) ds \\
&\quad + \frac{2}{\sqrt{\pi}} \Big(\int_0^t a \Big)^{1/2} \|p_0\|_{L^1} + \frac{2}{\alpha\sqrt{\pi}} \int_0^t D(q(s)) \Big(\int_s^t a \Big)^{1/2} ds,
\end{aligned}
\tag{2.12}
$$

since $\int_{\mathbb{R}} |\sigma| \varphi_\nu(\sigma) d\sigma = (2/\pi)^{1/2} \nu$ and $\int_{\mathbb{R}} \varphi_\nu = 1$. With the help of (2.11), and observing that $|\chi(t) - \chi(s)| \leq \sqrt{t-s} \|b\|_{L^2(0,T)}$, we then deduce (2.10).

We now use this bound to check that $p \in C_t^0(L^1_\sigma)$ and $D(p) \in C_t^0$. Indeed, for any t in $[0,T]$, any $A > 1$, and any sequence t_n in $[0,T]$ which converges to t, we have

$$
\begin{aligned}
\int_{\mathbb{R}} \big| p(t_n) - p(t) \big| &= \int_{|\sigma| \leq A} \big| p(t_n) - p(t) \big| + \int_{|\sigma| \geq A} \big| p(t_n) - p(t) \big| \\
&\leq \sqrt{2A} \Big(\int_{\mathbb{R}} \big| p(t_n) - p(t) \big|^2 \Big)^{1/2} + \frac{1}{A} \int_{\mathbb{R}} |\sigma| \big(|p(t_n)| + |p(t)| \big) \\
&\leq \sqrt{2A} \Big(\int_{\mathbb{R}} \big| p(t_n) - p(t) \big|^2 \Big)^{1/2} + \frac{2}{A} \sup_{0 \leq t \leq T} \int_{\mathbb{R}} |\sigma| |p(t)|.
\end{aligned}
\tag{2.13}
$$

For any fixed A, the first term on the right-hand side goes to 0 as n goes to infinity since $p \in C_t^0(L^2_\sigma)$ and then the second term is arbitrarily small as A goes to infinity. The same argument yields the continuity of $D(p(t))$ with respect to t. ∎

The following proposition aims at checking the required assumptions to apply the Schauder fixed point theorem.

PROPOSITION 2.4. *Let $T_f > 0$ be given. We assume that*

$$p_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R}), \quad p_0 \geq 0, \int_{\mathbb{R}} p_0 = 1, \text{ and } \int_{\mathbb{R}} |\sigma| p_0 < +\infty. \tag{2.14}$$

Let $0 < \varepsilon \leq 1$, $R = 1 + \int_{\mathbb{R}} |\sigma| p_0$, and $M = 1 + 2\alpha$. We define

$$T_c = \frac{9}{25} \left[\|b\|_{L^2(0,T_f)} + \frac{2\sqrt{1+2\alpha}}{\sqrt{\pi}} \right]^{-2}. \tag{2.15}$$

Then, for every $T \leq \min(\frac{1}{R}; T_c)$, the function $\mathcal{T} : (a; q) \mapsto (D(p) + \varepsilon; p)$, with p being the solution to the system (2.5), *maps $\mathcal{D}_{\varepsilon,M} \times Y_R$ into itself. Moreover, $\mathcal{T}$ is continuous and $\mathcal{T}(\mathcal{D}_{\varepsilon,M} \times Y_R)$ is relatively compact in $L^2(0,T) \times L^2_{t,\sigma}$.*

Proof. **Step 1.** *$\mathcal{T}$ is well-defined.* According to Proposition 2.3, p is in $C^0_t(L^1_\sigma)$ and $D(p) \in C^0_t$. We now prove that with our choice for M (which ensures that $\varepsilon + D(p_0) \leq 1 + \alpha \leq M$), $D(p) + \varepsilon \in \mathcal{D}_{\varepsilon,M}$. For this, we again use the inequality $p \leq p_+$, the definition (2.8) of p_+, the rough estimate $\int_{|\sigma|>1} \varphi_\nu \leq \int_{\mathbb{R}} \varphi_\nu = 1$ and (2.11) to obtain

$$\sup_{0 \leq t \leq T} D(p(t)) \leq \sup_{0 \leq t \leq T} D(p_+(t)) \leq \alpha + \alpha R T \leq 2\alpha,$$

for $T \leq \frac{1}{R}$. It only remains now to check that $\sup_{0 \leq t \leq T} \int_{\mathbb{R}} |\sigma| p \leq R$. We thus go back to (2.10) and observe that this condition holds provided

$$T \leq \max \left\{ t > 0; \|b\|_{L^2(0,T_f)} \sqrt{t} \left(1 + \frac{2R}{3} t \right) + \frac{2\sqrt{Mt}}{\sqrt{\pi}} + \frac{4R\sqrt{M} t^{3/2}}{3\sqrt{\pi}} \leq 1 \right\}.$$

Since we have already demanded that $t \leq T \leq \frac{1}{R}$, a sufficient condition is then

$$\sqrt{T} \left[\frac{5}{3} \|b\|_{L^2(0,T_f)} + \frac{10\sqrt{1+2\alpha}}{3\sqrt{\pi}} \right] \leq 1,$$

which reduces to $T \leq T_c$ with T_c given by (2.15).

Our next step will consist of establishing a priori bounds on p in $L^\infty_t(L^2_\sigma) \cap L^2_t(H^1_\sigma)$.

Step 2. *A priori bounds.* If we multiply (2.5) by p and integrate by parts over $\mathbb{R}$ with respect to σ we easily obtain

$$\frac{1}{2} \frac{d}{dt} \int_{\mathbb{R}} p^2 + a \int_{\mathbb{R}} |\partial_\sigma p|^2 \leq \frac{D(q)}{\alpha} p(t, 0).$$

Since from the Sobolev embedding of $H^1(\mathbb{R})$ into $L^\infty(\mathbb{R})$ and the bound (2.11) on $D(q)$, we get

$$\begin{aligned} \left| \frac{D(q)}{\alpha} p(t,0) \right| &\leq R \|p\|_{L^\infty_\sigma} \\ &\leq R \big(\|p\|^2_{L^2_\sigma} + \|\partial_\sigma p\|^2_{L^2_\sigma} \big)^{\frac{1}{2}} \\ &\leq \frac{R^2}{2\varepsilon} + \frac{\varepsilon}{2} \|p\|^2_{L^2_\sigma} + \frac{\varepsilon}{2} \|\partial_\sigma p\|^2_{L^2_\sigma}, \end{aligned}$$

we may write

$$\frac{1}{2} \frac{d}{dt} \|p\|^2_{L^2_\sigma} + \left(a - \frac{\varepsilon}{2} \right) \|\partial_\sigma p\|^2_{L^2_\sigma} \leq \frac{R^2}{2\varepsilon} + \frac{\varepsilon}{2} \|p\|^2_{L^2_\sigma}. \tag{2.16}$$

We recall that $a \geq \varepsilon$ and we apply the Gronwall lemma to obtain

$$\sup_{0\leq t\leq T} \|p\|^2_{L^2_\sigma} \leq e^{\varepsilon T}\left(\|p_0\|^2_{L^2_\sigma} + \frac{TR^2}{\varepsilon}\right). \tag{2.17}$$

We now return to (2.16) and integrate it over $[0;T]$ to obtain

$$\varepsilon\|\partial_\sigma p\|^2_{L^2_{t,\sigma}} \leq \|p_0\|^2_{L^2_\sigma}(1+\varepsilon T e^{\varepsilon T}) + \frac{TR^2}{\varepsilon}(1+\varepsilon T e^{\varepsilon T}). \tag{2.18}$$

Step 3. *The function $\mathcal{T}$ is continuous.* We consider a sequence $(a_n;q_n)$ in $\mathcal{D}_{\varepsilon,M}\times Y_R$ such that a_n converges to a strongly in L^2_t and q_n converges to q strongly in $L^2_{t,\sigma}$, and we denote $\mathcal{T}(a_n;q_n) = (D(p_n)+\varepsilon;p_n)$. We have to prove that p_n converges strongly to p in $L^2_{t,\sigma}$ and $D(p_n)$ converges to $D(p)$ strongly in L^2_t, with $(D(p)+\varepsilon;p) = \mathcal{T}(a;q)$.

In virtue of (2.17) and (2.18), the sequence p_n is bounded in $L^\infty_t(L^2_\sigma)\cap L^2_t(H^1_\sigma)$. Then, $\partial_\sigma p_n$ is bounded in $L^\infty_t(H^{-1}_\sigma)$ and $\partial^2_{\sigma\sigma}p_n$ is bounded in $L^2_t(H^{-1}_\sigma)$. Since $a_n\partial^2_{\sigma\sigma}p_n$ is bounded in $L^2_t(H^{-1}_\sigma)$, $b\in L^2_t$ and $D(q_n)\delta_0$ is bounded in $L^2_t(H^{-1}_\sigma)$, $\partial_t p_n$ is bounded in $L^2_t(H^{-1}_\sigma)$. This together with the fact that p_n is bounded in $L^2_t(H^1_\sigma)$ implies that, up to a subsequence, p_n converges strongly towards p in $L^2_t(L^2_{\mathrm{loc},\sigma})$ (the convergence being weak in $L^2_t(H^1_\sigma)$) thanks to a well-known compactness result [10]. In particular, p_n converges to p almost everywhere. Thus $p\geq 0$ and by Fatou's lemma, $\int_{\mathbb{R}}|\sigma|p\leq R$ almost everywhere on $[0;T]$. Hence p belongs to Y_R. We are going to show that the convergence is actually strong in $L^2_{t,\sigma}$.

In virtue of (2.9) in Proposition 2.3, we dispose of a uniform a priori bound on p_n in $L^\infty_{t,\sigma}$ (hence also on p). For the strong convergence in $L^2_{t,\sigma}$ we then argue as follows. For any fixed positive real number K, we have

$$\begin{aligned}\int_0^T\int_{\mathbb{R}}|p_n-p|^2 &\leq \int_0^T\int_{|\sigma|\leq K}|p_n-p|^2 + \int_0^T\int_{|\sigma|>K}|p_n-p|^2\\ &\leq \int_0^T\int_{|\sigma|\leq K}|p_n-p|^2 + (\|p_n\|_{L^\infty_{t,\sigma}} + \|p\|_{L^\infty_{t,\sigma}})\frac{2RT}{K},\end{aligned}$$

owing to the fact that p_n and p belong to a bounded subset of $Y_R\cap L^\infty_{t,\sigma}$. We then conclude by letting n, then K, go to infinity.

We now prove that $D(p_n)$ converges to $D(p)$ strongly in L^2_t. We shall actually prove that $D(p_n)$ converges to $D(p)$ strongly in L^1_t and then use the fact that $D(p_n)$ is bounded in L^∞_t, in virtue of (2.11) and because p_n lies in Y_R. Let us fix $K>1$. Then, we have

$$\begin{aligned}\frac{1}{\alpha}\int_0^T|D(p_n)-D(p)| &= \int_0^T\left|\int_{|\sigma|>1}(p_n-p)\right|\\ &\leq \int_0^T\int_{1<|\sigma|<K}|p_n-p| + \frac{1}{K}\int_0^T\int_{|\sigma|>K}|\sigma|(|p_n|+|p|)\\ &\leq \int_0^T\int_{1<|\sigma|<K}|p_n-p| + \frac{2RT}{K},\end{aligned} \tag{2.19}$$

because p and p_n belong to Y_R. Since p_n converges to p strongly in $L^1_t(L^1_{\mathrm{loc},\sigma})$, we conclude that $D(p_n)$ converges to $D(p)$ in L^1_t by letting n, then K, go to infinity in (2.19).

In order to pass to the limit in the equation satisfied by p_n (thereby proving that $(D(p)+\varepsilon;p) = \mathcal{T}(a;q)$), we now observe that the strong convergence of q_n to q in $L^2_{t,\sigma}$, together with the argument in (2.19) above shows that $D(q_n)$ converges to $D(q)$ strongly in L^2_t. It is then easily proved that p is a weak solution to (1.7) and since p is in $L^2_t(H^1_\sigma)$ it is the unique solution to (2.5) corresponding to a and q. In particular, the whole sequence p_n converges and not only a subsequence.

Step 4. *$\mathcal{T}(\mathcal{D}_\varepsilon \times Y_R)$ is relatively compact.* Let $(D(p_n)+\varepsilon;p_n) = \mathcal{T}(a_n;q_n)$ be a sequence in $\mathcal{T}(\mathcal{D}_{\varepsilon,M} \times Y_R)$. We have to prove that we may extract a subsequence which converges strongly in $L^2_t \times L^2_{t,\sigma}$. Exactly as for the proof of the continuity, the a priori estimates (2.17) and (2.18) ensure that the sequence p_n is bounded in $L^\infty_t(L^2_\sigma) \cap L^2_t(H^1_\sigma)$. Since $|\sigma|p_n$ is bounded $L^\infty_t(L^1_\sigma)$, we can mimic the argument in Step 3 above to deduce that up to a subsequence the sequence p_n converges to some p in Y_R strongly in $L^2_{t,\sigma}$ and that $D(p_n)$ converges to $D(p)$ strongly in L^2_t. □

We are now in position to conclude the proof of Proposition 1.2.

Let $T_f > 0$ and $0 < \varepsilon \le 1$ being given. We are going to prove the existence of a unique solution on $[0;T_f]$.

Being given an initial data p_0 which satisfies (1.6), existence of a solution p_ε is ensured from Proposition 2.4 by applying the Schauder fixed point theorem on "short" time interval $[0;T_1]$ with $T_1 = \min(\frac{1}{R_1}, T_c)$ and where $R_1 = 1 + \int_{\mathbb{R}} |\sigma| p_0$. This solution is uniquely defined in virtue of Lemma 2.1 and we know from (2.1) that $\int_{\mathbb{R}} p_\varepsilon(T_1) = 1$. Moreover, from Proposition 2.3 $p_\varepsilon(T_1) \in L^\infty_\sigma$ and by construction $\int_{\mathbb{R}} |\sigma| p_\varepsilon(T_1) \le R_1$. Therefore $p_\varepsilon(T_1)$ satisfies the same conditions (2.14) as p_0. Then, repeating the same argument we may build a solution to (1.7) with initial data $p_\varepsilon(T_1)$ on $[T_1;T_2]$ with $T_2 = \min(\frac{1}{R_2}, T_c)$, where $R_2 = R_1 + 1 = \int_{\mathbb{R}} |\sigma| p_0 + 2$. Thanks to the uniqueness result (Lemma 2.1), if we now glue this solution to p_ε at $t = T_1$ we obtain the unique solution to (1.7) on $[0;T_1+T_2]$. It is now clearly seen that for any integer $n \ge 1$ we may build a solution to (1.7) on $[0;\sum_{1\le k\le n} T_k]$ with $T_k = \min((k + \int_{\mathbb{R}} |\sigma| p_0)^{-1}; T_c)$. Since $\sum_{1\le k\le n} T_k$ obviously goes to $+\infty$ together with n, existence (and uniqueness) of the solution p_ε to (1.7) is obtained on every time interval.

For the proof of (1.9) we argue as for the proof of (2.9) in Proposition 2.3. Defining p^+_ε as in (2.8) with a replaced by $D(p_\varepsilon)+\varepsilon$ and $D(q)$ by $D(p_\varepsilon)$, we obtain

$$
\begin{aligned}
0 \le p_\varepsilon \le p^+_\varepsilon &\\
&\le \|p_0\|_{L^\infty} + \frac{1}{\alpha\sqrt{\pi}} \int_0^t \frac{D(p_\varepsilon(s))}{2\sqrt{\varepsilon + \int_s^t D(p_\varepsilon)}} ds \\
&\le \|p_0\|_{L^\infty} + \frac{1}{\alpha\sqrt{\pi}} \left[\sqrt{\varepsilon + \int_0^t D(p_\varepsilon)} - \sqrt{\varepsilon} \right] \\
&\le \|p_0\|_{L^\infty} + \frac{1}{\alpha\sqrt{\pi}} \sqrt{\int_0^t D(p_\varepsilon)} \\
&\le \|p_0\|_{L^\infty} + \frac{\sqrt{\alpha}\sqrt{T}}{\sqrt{\pi}}.
\end{aligned}
$$

Then

$$\int_{\mathbb{R}} p_\varepsilon^2 \le \|p_\varepsilon\|_{L^\infty_\sigma} \int_{\mathbb{R}} p_\varepsilon,$$

from which (1.11) follows gathering together (2.1) and (1.9) and, with the notation of the proposition,

$$C_2(T,p_0) = \|p_0\|_{L^\infty} + \frac{\sqrt{\alpha}\sqrt{T}}{\sqrt{\pi}}.$$

The proof of (1.10) follows the same lines as the proof of (2.12). Indeed, we again use the pointwise inequality $p_\varepsilon \le p_\varepsilon^+$ and replace $D(q)$ by $D(p_\varepsilon)(\le \alpha)$ and a by $D(p_\varepsilon) + \varepsilon(\le \alpha + 1)$ in (2.12) and use (2.14) to deduce

$$\sup_{0\le t\le T} \int_{\mathbb{R}} |\sigma| p_\varepsilon \le \int_{\mathbb{R}} |\sigma| p_0 + \sqrt{T}\left(\frac{2\sqrt{1+\alpha}}{\sqrt{\pi}} + \|b\|_{L^2(0,T)}\right) + \frac{2}{3}T^{3/2}\left(1 + \frac{2\sqrt{1+\alpha}}{\sqrt{\pi}}\right),$$

whence (1.10) with $C_1(T,p_0)$ being the quantity in the right-hand side of the above inequality.

In order to prove (1.12), we apply p_ε to (1.7) and we integrate by parts over $\mathbb{R}$ with respect to σ to obtain

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}} p_\varepsilon^2 + (D(p_\varepsilon)+\varepsilon)\int_{\mathbb{R}} |\partial_\sigma p_\varepsilon|^2 + \int_{|\sigma|>1} p_\varepsilon^2 = \frac{D(p_\varepsilon)}{\alpha} p_\varepsilon(t,0). \tag{2.20}$$

We use the L^∞ bound (1.9) to bound the right-hand side and we integrate (2.20) with respect to t over $[0;T]$ to deduce (1.12) with

$$C_3(T,p_0) = \|p_0\|_{L^\infty}\left(\frac{1}{2} + T\right) + \frac{\sqrt{\alpha}}{\sqrt{\pi}}T^{3/2},$$

using $\|p_0\|^2_{L^2_\sigma} \le \|p_0\|_{L^\infty}\int_{\mathbb{R}} p_0$.

3. The nondegenerate case: $\mathbf{D(p_0)}$ 0. The main result of this section corresponds to the statement of Theorem 1.1 and fully describes the issue of existence and uniqueness of solutions to the HL equation (1.1) in the nondegenerate case. It is summarized in the following proposition.

PROPOSITION 3.1. *Let p_0 satisfy* (1.6). *We assume that $D(p_0) > 0$. Then, the HL equation* (1.1) *has a unique solution p in $C^0_t(L^2_\sigma) \cap L^2_t(H^1_\sigma)$ and p is the limit (in $L^2_{t,\mathrm{loc}}(L^2_\sigma) \cap C^0_{t,\mathrm{loc}}(L^2_\sigma)$) of (p_ϵ) when ϵ goes to 0 where p_ε is the vanishing viscosity solution whose existence and uniqueness is ensured by Proposition* 1.2. *Moreover, $p \in L^\infty_{t,\sigma} \cap C^0_t(L^1_\sigma)$, $\sigma p \in L^\infty_t(L^1_\sigma)$ and $\int_{\mathbb{R}} p = 1$. Furthermore, $D(p) \in C^0_t$ and for every $T > 0$ there exists a positive constant $\nu(T)$ such that*

$$\min_{0\le t\le T} D(p(t)) \ge \nu(T). \tag{3.1}$$

We begin by proving the following lemma.

LEMMA 3.2. *We assume that p_0 satisfies* (1.6). *Then, if $D(p_0) > 0$, $D(p_\varepsilon)(t) > 0$ for every $t \in [0,T]$, with p_ε being the unique solution to* (1.7) *provided by Proposition* 1.2 *and, actually, for every $T > 0$ there exists a positive constant $\nu(T)$ such that*

$$\min_{0\le t\le T} D(p_\varepsilon(t)) \ge \nu(T), \tag{3.2}$$

for every $0 < \varepsilon \le 1$.

Remark 3.3. Note that this bound from below is independent of ε, but it comes out from the proof that it depends on p_0 and on the shear b.

Proof. The proof relies on the bound from below in (2.6) that we integrate over $|\sigma| > 1$ to obtain

$$\begin{aligned} D(p_\varepsilon(t)) &\geq \alpha \int_{|\sigma|>1} p_\varepsilon^- \\ &\geq \alpha e^{-t} \int_{\mathbb{R}} p_0(\sigma') \left(\int_{|\sigma|>1} \varphi_{\sqrt{2\int_0^t (D(p_\varepsilon)+\varepsilon)}} (\sigma - \sigma' - \chi(t)) d\sigma \right) d\sigma'. \end{aligned} \tag{3.3}$$

Let us define $K_\chi = [-1 - \chi(t), 1 - \chi(t)]$. The function $\sigma \mapsto \varphi_{\sqrt{2\int_0^t (D(p_\varepsilon)+\varepsilon)}}(\sigma - \sigma' - \chi(t))$ is a Gaussian probability density with mean $\sigma' + \chi(t)$ and squared width $2\int_0^t (D(p_\varepsilon) + \varepsilon)$. Therefore, for every $\sigma' \in \mathbb{R} \setminus K_\chi$, we have

$$\int_{|\sigma|>1} \varphi_{\sqrt{2\int_0^t (D(p_\varepsilon)+\varepsilon)}} (\sigma - \sigma' - \chi(t)) d\sigma \geq \frac{1}{2},$$

which implies (3.3) is

$$\geq \frac{\alpha}{2} e^{-T} \int_{\mathbb{R}\setminus K_\chi} p_0 = \frac{\alpha}{2} e^{-T} \int_{|\sigma+\chi(t)|>1} p_0.$$

In the zero shear case ($b \equiv 0$, thus $\chi \equiv 0$) the proof is over and

$$\min_{0\leq t\leq T} D(p(t)) \geq \frac{1}{2} e^{-T} D(p_0).$$

In the general case, a strictly positive bound from below is available as long as the support of p_0 is not contained in K_χ. We thus define

$$t^* = \inf \left\{ t > 0; \int_{|\sigma+\chi(t)|>1} p_0 = 0 \right\}. \tag{3.4}$$

Then $0 < t^*$ (t^* possibly even infinite), the support of p_0 is contained in $[-1-\chi(t^*), 1-\chi(t^*)[$, and for every $T < \frac{t^*}{2}$, (3.2) holds for some positive constant $\nu_1(T)$ defined by

$$\nu_1(T) = \frac{\alpha}{2} e^{-T} \min_{0\leq t\leq T} \int_{|\sigma+\chi(t)|>1} p_0. \tag{3.5}$$

It is worth emphasizing that this quantity is independent of ε. If $t^* = +\infty$, the proof is over and $\nu(T) = \nu_1(T)$ fits. Let us now examine the case when $t^* < +\infty$ and $T \geq \frac{t^*}{2}$.

We go back to (3.3), take t in $[\frac{t^*}{2}; T]$, and denote $x = \int_0^t (D(p_\varepsilon) + \varepsilon)$ for shortness.

Then

$$\begin{aligned}
D(p_\varepsilon(t)) &\geq \alpha e^{-T} \int_{-1-\chi(t^*)}^{1-\chi(t^*)} p_0(\sigma') \left(\int_{|\sigma|>1} \varphi_{\sqrt{2x}}(\sigma - \sigma' - \chi(t)) d\sigma \right) d\sigma' \\
&= \alpha e^{-T} \int_{-1-\chi(t^*)}^{1-\chi(t^*)} p_0(\sigma') \left(\int_{|\sigma|>1} \frac{e^{-(\sigma-\sigma'-\chi(t))^2/4x}}{2\sqrt{\pi}\sqrt{x}} d\sigma \right) d\sigma' \\
&= \frac{\alpha}{\sqrt{\pi}} e^{-T} \int_{-1-\chi(t^*)}^{1-\chi(t^*)} p_0(\sigma') \Bigg(\int_{-\infty}^{-1+\sigma'+\chi(t)} \frac{e^{-\sigma^2/4x}}{2\sqrt{x}} d\sigma \\
&\qquad\qquad + \int_{1+\sigma'+\chi(t)}^{+\infty} \frac{e^{-\sigma^2/4x}}{2\sqrt{x}} d\sigma \Bigg) d\sigma' \\
&= \frac{\alpha}{\sqrt{\pi}} e^{-T} \int_{-1-\chi(t^*)}^{1-\chi(t^*)} p_0(\sigma') \left(\int_{\frac{1+\sigma'+\chi(t)}{2\sqrt{x}}}^{+\infty} e^{-t^2} dt + \int_{\frac{1-\sigma'-\chi(t)}{2\sqrt{x}}}^{+\infty} e^{-t^2} dt \right) d\sigma' \\
&\geq \frac{\alpha}{\sqrt{\pi}} e^{-T} \left(\int_{-1-\chi(t^*)}^{1-\chi(t^*)} p_0(\sigma') d\sigma' \right) \Bigg(\int_{\frac{2-\chi(t^*)+\chi(t)}{\sqrt{2t^*\nu_1(t^*/2)}}}^{+\infty} e^{-t^2} dt \\
&\qquad\qquad + \int_{\frac{2+\chi(t^*)-\chi(t)}{\sqrt{2t^*\nu_1(t^*/2)}}}^{+\infty} e^{-t^2} dt \Bigg);
\end{aligned}$$

hence

$$D(p_\varepsilon(t)) \geq \frac{\alpha}{\sqrt{\pi}} e^{-T} \min_{t^*/2 \leq t \leq T} \left(\int_{\frac{2-\chi(t^*)+\chi(t)}{\sqrt{2t^*\nu_1(t^*/2)}}}^{+\infty} e^{-t^2} dt + \int_{\frac{2+\chi(t^*)-\chi(t)}{\sqrt{2t^*\nu_1(t^*/2)}}}^{+\infty} e^{-t^2} dt \right), \tag{3.6}$$

since $\int_{-1-\chi(t^*)}^{1-\chi(t^*)} p_0 = 1$ and $x \geq \int_0^{t^*/2} D(p_\varepsilon) \geq t^*\nu_1(t^*/2)/2$ thanks to (3.5). The proof of Lemma 3.2 then follows by defining

$$\nu(T) = \min(\nu_1(T); \nu_2(T)),$$

with $\nu_1(T)$ given by (3.5) and $\nu_2(T)$ being the positive quantity on the right-hand side of (3.6), that is

$$\nu_2(T) = \frac{\alpha}{\sqrt{\pi}} e^{-T} \min_{t^*/2 \leq t \leq T} \left(\int_{\frac{2-\chi(t^*)+\chi(t)}{\sqrt{2t^*\nu_1(t^*/2)}}}^{+\infty} e^{-t^2} dt + \int_{\frac{2+\chi(t^*)-\chi(t)}{2\sqrt{2t^*\nu_1(t^*/2)}}}^{+\infty} e^{-t^2} dt \right). \qquad \square$$

Proof of Proposition 3.1. We first go back to the proof of the bound (1.12) on $\partial_\sigma p_\varepsilon$, and more precisely we look at (2.20), and observe that in virtue of (3.1)

$$\nu(T) \int_0^T \int_{\mathbb{R}} |\partial_\sigma p_\varepsilon|^2 \leq C_3(T, p_0). \tag{3.7}$$

Now let ε_n denote any sequence in $[0,1]$ which goes to 0 as n goes to infinity. To shorten the notation we denote by p_n instead of p_{ε_n} the corresponding sequence of solutions to (1.7). With the above bound (3.7) on p_n and (1.11), we know that p_n is bounded in $L^2_t(H^1_\sigma)$ independently of n. Moreover, thanks to (2.1) and (1.9), p_n is bounded in $L^\infty_t(L^1_\sigma \cap L^\infty_\sigma)$ and we also dispose of a uniform bound on $\int_{\mathbb{R}} |\sigma| p_n$ in virtue of (1.10). Therefore arguing exactly as in the proof of Proposition 2.4 (Step 4) where we have proved that the mapping $\mathcal{T}$ is relatively compact in $L^2_t \times L^2_{t,\sigma}$ we show that p_n converges to some p strongly in $L^2_{t,\sigma}$ and $D(p_n)$ converges to $D(p)$ in L^2_t. Therefore, the nonlinear term $D(p_n)\partial^2_{\sigma\sigma}p_n$ converges to $D(p)\partial^2_{\sigma\sigma}p$ strongly in $L^1_t(H^{-2}_\sigma)$ (for instance). Then p is a weak solution to the initial problem (1.1) in $L^2_t(H^1_\sigma) \cap L^\infty_t(L^1_\sigma \cap L^\infty_\sigma)$, $\int_{\mathbb{R}} p = 1$ and $\int_{\mathbb{R}} |\sigma| p < +\infty$. Moreover,

$$\inf_{0 \le t \le T} D(p(t)) \ge \nu(T).$$

This nondegeneracy condition on the viscosity coefficient ensures that there is at most one solution to (1.1) in $L^2_t(H^1_\sigma) \cap L^\infty_t(L^2_\sigma)$ (this follows by an obvious adaptation of the proof of Lemma 2.1 to this case). Therefore the limiting function p is uniquely defined and does not depend on the sequence ε_n. Moreover, the whole sequence p_n converges to this unique limit and not only a subsequence. $\square$

As a conclusion of this subsection let us make the following comment which is a by product of Proposition 3.1. If p is a solution to (1.1) in $C^0_t(L^1_\sigma \cap L^2_\sigma)$, then as soon as $D(p(t))$ is positive for some time t it remains so afterwards since the solution can be continued in a unique way starting at time t.

4. The degenerate case: $\mathbf{D(p_0) = 0}$. Throughout this section we assume that p_0 satisfies (1.3) and that $D(p_0) = 0$. Therefore the support of p_0 is included in $[-1; +1]$. Assume that we dispose of a solution to (1.1) in $C^0_t(L^1_\sigma \cap L^2_\sigma)$. We may define $t_* \in \mathbb{R}^+ \cup \{+\infty\}$ by

$$t_* = \max\left\{t > 0; \int_0^t D(p) = 0\right\}. \tag{4.1}$$

According to the comment at the end of the previous section for every $t > t_*$, $D(p(t)) > 0$ while $D(p(t)) = 0$ for all t in $[0; t_*]$. On $[0; t_*[$, the HL equation (1.1) reads

$$\begin{cases} \partial_t p = -b(t)\partial_\sigma p, \\ p \ge 0, \\ p(0, \cdot) = p_0, \\ D(p(t)) = 0. \end{cases}$$

The above system reduces to

$$\begin{cases} p(t,\sigma) = p_0(\sigma - \chi(t)), \\ D(p(t)) = 0 \quad \text{for all } t \text{ in } [0; t_*]. \end{cases} \tag{4.2}$$

The second equation in (4.2) is compatible with the first one as long as

$$\int_{|\sigma + \chi(t)| > 1} p_0 = 0 \quad \text{for all } t \text{ in } [0; t_*].$$

Therefore there exists a maximal time interval $[0; T_c]$ on which the HL equation may reduce to a mere transport equation and this is for an intrinsic time T_c (possibly infinite) defined by

$$T_c = \inf\left\{t > 0; \int_{|\sigma + \chi(t)| > 1} p_0 > 0\right\}. \tag{4.3}$$

Note that T_c is completely determined by the data p_0 and b. If $T_c = +\infty$, the steady state $p(t,\sigma) = p_0(\sigma - \chi(t))$ is a solution of the HL equation for all time. We shall now exhibit circumstances under which it is not the unique solution. For convenience, we restrict ourselves to the case when $b \equiv 0$ (we then have obviously $T_c = +\infty$).

For $p_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ such that $p_0 \geq 0$, let us denote by F_{p_0} the function from $\mathbb{R}^+$ to $\mathbb{R}^+$ defined by $F_{p_0}(0) = D(p_0)$ and by

$$\text{for all } x > 0, \qquad F_{p_0}(x) = \alpha \int_{|\sigma|>1} \left(\int_{\mathbb{R}} p_0(\sigma')\varphi_{\sqrt{2x}}(\sigma - \sigma')d\sigma' \right) d\sigma.$$

Proposition 4.1. *Let p_0 satisfy* (1.6) *and be such that $D(p_0) = 0$, then*

(i) *If F_{p_0} satisfies*

$$\int_0^1 \frac{dx}{F_{p_0}(x)} = +\infty, \tag{4.4}$$

then $p(t,\sigma) = p_0(\sigma)$ is the unique solution to (1.1) *in $C^0_t(L^2_\sigma)$.*

(ii) *Otherwise,* (1.1) *has an infinite number of solutions in $C^0_t(L^2_\sigma)$. The set of solutions to* (1.1) *is made of the steady state $p(t,\sigma) = p_0(\sigma)$ and of the functions $(q_{t_0})_{t_0 \geq 0}$ defined by*

$$q_{t_0}(t,\sigma) = \left| \begin{array}{ll} p_0(\sigma), & \text{if } t \leq t_0 \\ q(t - t_0, \sigma), & \text{if } t > t_0, \end{array} \right.$$

where q is the unique solution to (1.1) *in $C^0_t(L^2_\sigma)$ such that $D(q) > 0$ on $]0,+\infty[$. Besides,*

$$p_\epsilon \underset{\epsilon \to 0}{\longrightarrow} q \quad \text{strongly in } L^2_{t,\mathrm{loc}}(L^2_\sigma). \tag{4.5}$$

Lemma 4.2. *Let $p_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ such that*

$$p_0 \geq 0, \qquad \int_{\mathbb{R}} p_0 = 1, \qquad D(p_0) = 0.$$

The function F_{p_0} is in $C^0([0,+\infty[) \cap C^\infty(]0,+\infty[)$, and is positive on $]0,+\infty[$. In addition, $F'_{p_0} > 0$ on $]0,+\infty[$.

Proof. It is easy to check that $F_{p_0} \in C^0([0,+\infty[) \cap C^\infty(]0,+\infty[)$, and that $F_{p_0} > 0$ on $]0,+\infty[$. Since $D(p_0) = 0$, the function p_0 is supported in $[-1,1]$. Thus, for any $x > 0$,

$$\begin{aligned} F_{p_0}(x) &= \alpha \int_{|\sigma|>1} \left(\int_{\mathbb{R}} p_0(\sigma')\varphi_{\sqrt{2x}}(\sigma - \sigma')d\sigma' \right) d\sigma \\ &= \alpha \int_{-1}^1 p_0(\sigma') \left(\int_{|\sigma|>1} \frac{e^{-(\sigma-\sigma')^2/4x}}{2\sqrt{\pi}\sqrt{x}} d\sigma \right) d\sigma' \\ &= \alpha \int_{-1}^1 p_0(\sigma') \left(\int_{-\infty}^{-1+\sigma'} \frac{e^{-\sigma^2/4x}}{2\sqrt{\pi}\sqrt{x}} d\sigma + \int_{1+\sigma'}^{+\infty} \frac{e^{-\sigma^2/4x}}{2\sqrt{\pi}\sqrt{x}} d\sigma \right) d\sigma' \\ &= \alpha \frac{1}{\sqrt{\pi}} \int_{-1}^1 p_0(\sigma') \left(\int_{\frac{1+\sigma'}{2\sqrt{x}}}^{+\infty} e^{-t^2} dt + \int_{\frac{1-\sigma'}{2\sqrt{x}}}^{+\infty} e^{-t^2} dt \right) d\sigma'. \end{aligned} \tag{4.6}$$

It follows that for any $x > 0$,

$$F'_{p_0}(x) = \alpha \frac{1}{\sqrt{\pi}} \int_{-1}^{1} p_0(\sigma') \left(\frac{1+\sigma'}{4x^{3/2}} e^{-\frac{(1+\sigma')^2}{4x}} + \frac{1-\sigma'}{4x^{3/2}} e^{-\frac{(1-\sigma')^2}{4x}} \right) d\sigma' > 0. \qquad \square$$

LEMMA 4.3. *Let $\gamma \geq 0$ and $p_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ such that*

$$p_0 \geq 0, \qquad \int_{\mathbb{R}} p_0 = 1, \qquad \int_{\mathbb{R}} |\sigma| p_0 < +\infty, \qquad D(p_0) = 0.$$

Let us consider the problem

$$\begin{cases} \partial_t w = D(w(t)) \partial^2_{\sigma\sigma} w - \gamma w, \\ w(0,\sigma) = p_0(\sigma). \end{cases} \tag{4.7}$$

(i) *If F_{p_0} satisfies* (4.4), *then $p(t,\sigma) = p_0(\sigma)$ is the unique solution to* (4.7) *in $C^0_t(L^2_\sigma)$.*

(ii) *Otherwise,* (4.7) *has an infinite number of solutions in $C^0_t(L^2_\sigma)$. The set of solutions to* (4.7) *is made of the steady state $w(t,\sigma) = p_0(\sigma)$ and of the functions $(v_{t_0})_{t_0 \geq 0}$ defined by*

$$v_{t_0}(t,\sigma) = \left| \begin{array}{ll} p_0(\sigma), & \text{if } t \leq t_0, \\ v(t-t_0,\sigma), & \text{if } t > t_0, \end{array} \right.$$

where v is the unique solution to (4.7) *in $C^0_t(L^2_\sigma)$ such that $D(v) > 0$ on $]0,+\infty[$.*

COROLLARY 4.4. *The initial data $p_0 = \frac{1}{2} \mathbb{1}_{]-1,1[}$ fulfills the assumptions of the above lemma and $\int_0^1 \frac{dx}{F_{p_0}(x)} < +\infty$. Therefore there are infinitely many solutions to* (1.5) *in the introduction.*

Proof of Corollary 4.4. The only point to be checked is that $\int_0^1 \frac{dx}{F_{p_0}(x)} < +\infty$. With the standard notation $\operatorname{erfc}(z) \equiv \int_z^{+\infty} e^{-t^2} dt$, and by using (4.6) and symmetry considerations, simple calculations yield

$$\begin{aligned} F_{p_0}(x) &= \frac{2\alpha\sqrt{x}}{\sqrt{\pi}} \int_0^{\frac{1}{\sqrt{x}}} \operatorname{erfc}(\sigma) d\sigma \\ &= \frac{2\alpha}{\sqrt{\pi}} \left[\operatorname{erfc}\left(\frac{1}{\sqrt{x}} \right) - \frac{1}{2}\sqrt{x} e^{-\frac{1}{x}} + \frac{1}{2}\sqrt{x} \right]. \end{aligned}$$

Since $\operatorname{erfc}(z) \sim \frac{1}{2} e^{-z^2}/z$ for z going to $+\infty$, $F_{p_0}(x) \sim \frac{\alpha}{\sqrt{\pi}} \sqrt{x}$ near 0 and the integrability of $1/F_{p_0}$ on $[0;1]$ follows. $\square$

Proof of Lemma 4.3. Let us consider a nonnegative continuous function D on $\mathbb{R}^+$. The unique solution in $C^0_t(L^2_\sigma)$ to the problem

$$\begin{cases} \partial_t w_D = D(t) \partial^2_{\sigma\sigma} w_D - \gamma w_D, \\ w_D(0,\sigma) = p_0(\sigma), \end{cases} \tag{4.8}$$

is given by

$$w_D(t,\sigma) = \left| \begin{array}{ll} e^{-\gamma t} p_0(\sigma), & \text{if } t \leq t^*, \\ e^{-\gamma t} \displaystyle\int_{\mathbb{R}} p_0(\sigma') \varphi_{\sqrt{2\int_0^t D(s)ds}}(\sigma - \sigma') d\sigma', & \text{if } t > t^*, \end{array} \right. \tag{4.9}$$

where $t^* = \inf\{t > 0; \int_0^t D > 0\}$. Any solution to (4.7) thus satisfies $w = w_{D(w)}$ and therefore

$$\begin{aligned}
D(w(t)) &= D(w_{D(w)}(t)) \\
&= \alpha \int_{|\sigma|>1} w_{D(w)}(t,\sigma) d\sigma \\
&= \alpha e^{-\gamma t} \int_{|\sigma|>1} \left(\int_{\mathbb{R}} p_0(\sigma') \varphi_{\sqrt{2\int_0^t D(w(s))ds}}(\sigma - \sigma') d\sigma' \right) d\sigma \\
&= e^{-\gamma t} F_{p_0} \left(\int_0^t D(w(s)) ds \right).
\end{aligned}$$

It follows that the function $D(w)$ is the solution in $C^0([0,+\infty[)$ to the nonlinear integral equation

$$y(t) = e^{-\gamma t} F_{p_0} \left(\int_0^t y(s) ds \right). \tag{4.10}$$

On the other hand, if $D \in C^0([0,+\infty[)$ is solution to (4.10) it is easy to check that the function w_D defined by (4.9) is solution to (4.8).

If condition (4.4) is fulfilled, (4.10) has a unique solution in $C^0([0,+\infty[)$ (the constant function equal to zero) and the steady state $w(t,\cdot) = p_0$ is therefore the unique solution to (4.7) in $C_t^0(L_\sigma^2)$; otherwise, the set of solutions to (4.10) is made of the steady state $w(t,\cdot) = p_0$ and of the family $(y_{t_0})_{t_0 \geq 0}$ with

$$y_{t_0}(t) = \left| \begin{array}{ll} 0, & \text{if } t \leq t_0, \\ z(t - t_0), & \text{if } t > t_0, \end{array} \right.$$

where the function z is defined on $[0, +\infty[$ by

$$\int_0^{z(t)} \frac{dx}{F(x)} = \begin{cases} \dfrac{1 - e^{-\gamma t}}{\gamma}, & \text{if } \gamma > 0, \\ t, & \text{otherwise.} \end{cases}$$

Statement (ii) is obtained by denoting by v the solution to (4.8) associated with the function $z(t)$. ∎

Proof of Proposition 4.1. The solution p_ϵ to (1.7) satisfies the inequalities

$$p_\epsilon^-(t,\sigma) \leq p_\epsilon(t,\sigma) \leq p_\epsilon^+(t,\sigma) \quad \text{almost everywhere,}$$

where p_ϵ^- and p_ϵ^+ are defined in $C_t^0(L_\sigma^2)$ by

$$\begin{cases} \partial_t p_\epsilon^- = (D(p_\epsilon(t)) + \epsilon)\, \partial_{\sigma\sigma}^2 p_\epsilon^- - p_\epsilon^-, \\ p_\epsilon^-(0,\sigma) = p_0(\sigma), \end{cases}$$

and

$$\begin{cases} \partial_t p_\epsilon^+ = (D(p_\epsilon(t)) + \epsilon) \partial_{\sigma\sigma}^2 p_\epsilon^+ + \dfrac{D(p_\epsilon)}{\alpha} \delta_0, \\ p_\epsilon^+(0,\sigma) = p_0(\sigma). \end{cases}$$

Therefore, on the one hand,

$$D(p_\epsilon(t)) \geq D(p_\epsilon^-(t)) = e^{-t} F_{p_0}\left(\int_0^t (D(p_\epsilon)+\epsilon)\right) \tag{4.11}$$

and, on the other hand,

$$\begin{aligned} D(p_\epsilon(t)) &\leq D(p_\epsilon^+(t)) \\ &= F_{p_0}\left(\int_0^t (D(p_\epsilon)+\epsilon)\right) + \int_0^t \frac{D(p_\epsilon)(s)}{\alpha}\left(\int_{|\sigma|>1} \varphi_{\sqrt{2\int_s^t (D(p_\epsilon)+\epsilon)}}\right) ds \\ &\leq F_{p_0}\left(\int_0^t (D(p_\epsilon)+\epsilon)\right) + \frac{1}{\alpha}\int_0^t D(p_\epsilon)(s) ds. \end{aligned}$$

If (4.4) is not fulfilled, using (4.11) and the property that F_{p_0} is strictly increasing on $[0, +\infty[$, we obtain that

$$D(p_\epsilon) \geq z(t),$$

where $z(t)$ is the function defined in the proof of Lemma 4.3. As for any $0 < t_0 \leq T$ there exists $\eta > 0$ such that $z(t) \geq \eta$ on $[t_0, T]$ the same reasoning as in the nondegenerate case leads to the conclusion that (p_ϵ) converges up to an extraction to p in $\mathcal{D}'(]0, +\infty[\times\mathbb{R})$ and in $L^2([t_0, T], L^2(\mathbb{R}))$ for any $0 < t_0 < T < +\infty$, p being a solution to (1.1) in $C^0(]0, +\infty[, L^2_\sigma)$ such that $D(p) > 0$ on $]0, +\infty[$. □

5. Steady states. Throughout this section the shear rate b is assumed to be a given constant and we are looking for solutions in $L^1(\mathbb{R})$ to the following system:

$$\begin{cases} -b\partial_\sigma p + D(p)\partial^2_{\sigma\sigma} p - \mathbb{1}_{\mathbb{R}\setminus[-1,1]} p + \dfrac{D(p)}{\alpha}\delta_0(\sigma) = 0 \quad \text{on } (0;T)\times\mathbb{R}, \\ p \geq 0, \quad \displaystyle\int_{\mathbb{R}} p = 1, \\ D(p) = \alpha \displaystyle\int_{|\sigma|>1} p(\sigma) d\sigma. \end{cases} \tag{5.1}$$

Our main results are summarized in the following proposition.

PROPOSITION 5.1 (existence of steady states).

(i) *If $b \equiv 0$, any probability density which is compactly supported in $[-1; +1]$ is a solution to (5.1) which satisfies $D(p) = 0$. If $\alpha \leq \frac{1}{2}$, these are the only stationary solutions (and there are infinitely many), whereas when $\alpha > \frac{1}{2}$ there exists a unique stationary solution corresponding to a positive value of D, which is explicitly given in (5.2) and (5.4) below. This solution is even and with exponential decay at infinity.*

(ii) *If $b \not\equiv 0$, for any $\alpha > 0$, there exists a unique stationary solution to (5.1), and it corresponds to a positive value for D, which is implicitly given in (5.5) and (5.6) below. This solution has exponential decay at infinity.*

Remark 5.2. The statement in the above proposition was already pointed out by Hébraud and Lequeux [8].

Proof. **The case when $\mathbf{b \equiv 0}$.** We first observe that any nonnegative function p which is normalized in $L^1(\mathbb{R})$ and with support in $[-1; +1]$ is a solution to the system (5.1) since in that case all terms in the equation satisfied by p in (5.1) cancel. We now examine the issue of existence of solutions of (5.1) such that $D(p) > 0$. For simplicity we denote $D = D(p)$. For given constant $D > 0$, it is very easy to calculate explicitly the solutions to (5.1) on each of the three regions $\sigma < -1$, $\sigma \in [-1; +1]$ and $\sigma > 1$. Using compatibility conditions on $\mathbb{R}$ and the fact that p has to be in $L^1(\mathbb{R})$, one obtains

$$p(\sigma) = \begin{cases} \frac{\sqrt{D}}{2\alpha} e^{(1+\sigma)/\sqrt{D}}, & \text{if } \sigma \leq -1, \\ \frac{1}{2\alpha}\sigma + \frac{\sqrt{D}+1}{2\alpha}, & \text{if } -1 \leq \sigma \leq 0, \\ -\frac{1}{2\alpha}\sigma + \frac{\sqrt{D}+1}{2\alpha}, & \text{if } 0 \leq \sigma \leq 1, \\ \frac{\sqrt{D}}{2\alpha} e^{(1-\sigma)/\sqrt{D}}, & \text{if } 1 \leq \sigma. \end{cases} \tag{5.2}$$

The compatibility condition $D = D(p)$ happens to then be automatically satisfied and the normalization constraint $\int_{\mathbb{R}} p = 1$ imposes that D solves

$$D + \sqrt{D} = \alpha - \frac{1}{2}. \tag{5.3}$$

Since $D \geq 0$, we immediately reach a contradiction when $\alpha < \frac{1}{2}$, whereas when $\alpha > \frac{1}{2}$ (5.3) admits a unique positive solution; namely

$$D = -\frac{1}{2} + \frac{\sqrt{4\alpha} - 1}{2}. \tag{5.4}$$

The case when $\mathbf{b} \not\equiv \mathbf{0}$. First of all, we observe that if $D = 0$ all terms in the equation satisfied by p in (5.1) but $b\partial_\sigma p$ vanish. Thus p has to be a nonzero constant which is in contradiction with $p \in L^1(\mathbb{R})$. So necessarily $D > 0$. For given positive constant D, we then solve (5.1) as above and obtain

$$p(\sigma) = \begin{cases} a_1 e^{\beta^+ \sigma}, & \text{if } \sigma \leq -1, \\ a_2 e^{\frac{b}{D}\sigma} + a_2 - \frac{D}{b\alpha}, & \text{if } -1 \leq \sigma \leq 0, \\ \left(a_2 - \frac{D}{b\alpha}\right) e^{\frac{b}{D}\sigma} + a_2, & \text{if } 0 \leq \sigma \leq 1, \\ a_1 e^{\beta^- \sigma}, & \text{if } 1 \leq \sigma, \end{cases} \tag{5.5}$$

with

$$\beta^\pm = \frac{b}{2D} \pm \frac{1}{2}\sqrt{\frac{b^2 + 4D}{D^2}},$$

$$a_1 = \frac{e^{\frac{1}{2}\sqrt{\frac{b^2}{D^2} + \frac{4}{D}}}}{\alpha\left(\beta^+ e^{b/2D} - \beta^- e^{-b/2D}\right)},$$

and

$$a_2 = \frac{D\beta^+ e^{b/2D}}{\alpha b(\beta^+ e^{b/2D} - \beta^- e^{-b/2D})}.$$

It is tedious but easy to check that this function always fulfills the self-consistency condition $D = D(p)$ and that the normalization condition $\int_{\mathbb{R}} p = 1$ reads

$$\frac{D}{b} \frac{(1 + \beta^+) + (\beta^- - 1)e^{-b/D}}{\beta^+ - \beta^- e^{-b/D}} + D = \alpha. \tag{5.6}$$

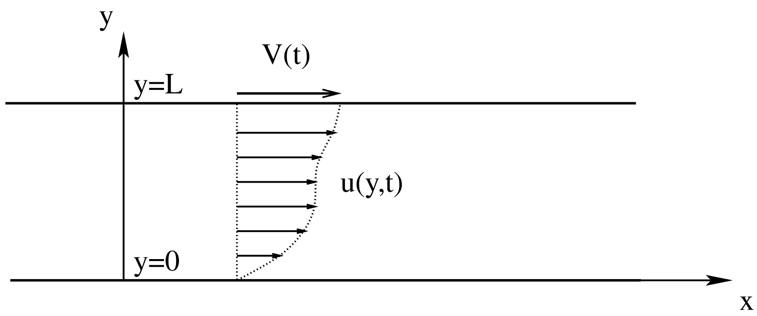


FIG. 1. *Planar Couette flow.*

For any $b > 0$ (the negative values of b are dealt with by replacing σ by $-\sigma$), the left-hand side of (5.6) is a continuous function which goes to $+\infty$ when D goes to infinity and goes to zero when D goes to 0. This already ensures the existence of at least one steady state for any $\alpha > 0$. Moreover, setting $z = b^2/D$ (for example) we may rewrite the left-hand side of (5.6) as

$$f(z) = \frac{b^2}{z} + \frac{2b^2}{z}\left[\frac{1 + \frac{1}{2b} z \coth(z/2b) + \frac{1}{2b}(z^2+4z)^{1/2}}{z + (z^2+4z)^{1/2}\coth(z/2b)}\right].$$

Next we check that the function f is monotone decreasing (thus, the left-hand side of (5.6) is increasing with respect to D), whence the uniqueness result. $\square$

6. Conclusion and future trends. Theorem 1.1 shows that the Hébraud–Lequeux model (1.1) is well-posed when the initial data p_0 is such that $D(p_0) > 0$. On the other hand, Proposition 4.1 claims that the model may have infinitely many solutions for certain p_0 such that $D(p_0) = 0$. This pathological behavior might be a flaw of the model but we think that is not the case. Indeed, our interpretation is that an initial data p_0 such that $D(p_0) = 0$ can be considered as admissible only if it is the long-time limit of a solution $p > 0$ of (1.1) with zero shear rate ($b = 0$); in view of both numerical simulations and heuristic arguments, we suspect that such a p_0 necessarily fulfills condition (4.4). We are currently studying the long-time asymptotics of the HL equation [3] and hope to be able to provide mathematical justifications of this statement in the near future.

To conclude, let us mention that the present article is the theoretical side of a joint project with rheologists aiming at better understanding the flows of complex fluids.

We are currently working on a multiscale model for planar Couette flows of concentrated suspensions (see Figure 1), in which the mesoscopic behavior of the suspension is described by the HL equation. It is indeed experimentally observed that the shear rate $b(t)$ is not homogeneous in space (in particular, the term $-b(t)\partial_\sigma p$ generates elastic waves in the fluid). In order to better describe the coupling between the macroscopic flow and the evolution of the mesostructure, we propose the following multiscale model:

$$
\begin{cases}
\rho\partial_t u(t,y) = \partial_y \tau(t,y) + \mu \partial_{yy} u(t,y), \\
\partial_t p(t,y,\sigma) = -G_0 \partial_y u(t,y) \partial_\sigma p(t,y,\sigma) + D(p(t,y)) \partial^2_{\sigma\sigma} p(t,y,\sigma) \\
\qquad\qquad -\dfrac{\mathbb{1}_{\mathbb{R}\setminus[-\sigma_c,\sigma_c]}(\sigma)}{T_0} p(t,y,\sigma) + \dfrac{1}{T_0}\left(\displaystyle\int_{|\sigma'|>\sigma_c} p(t,\sigma',y)\,d\sigma'\right)\delta_0(\sigma), \\
\tau(t,y) = \displaystyle\int_{\mathbb{R}} \sigma p(t,y,\sigma)\,d\sigma.
\end{cases}
\tag{6.1}
$$

In the above equations, $u(t,y)$ denotes the component along e_x of the velocity field (the flow being laminar and incompressible, the velocity field is of the form $\vec{u} = u(t,y)e_x$ and the pressure does not play any role), ρ is the volumic mass of the fluid and μ some nonnegative viscosity coefficient. This system is complemented by the no-slip boundary conditions

$$
\begin{cases}
u(t,0) = 0 & \text{for almost all } t, \\
u(t,L) = V(t) & \text{for almost all } t.
\end{cases}
$$

The theoretical study of this multiscale model will be the matter of another article. Numerical simulations of (6.1) have already been performed by one of us [6]. The results of these simulations will be compared in the near future with experimental data obtained with a NMR rheometer [12] (this very modern equipment allows measurements of local velocities in opaque fluids).

Acknowledgments. We would like to thank Philippe Coussot for pointing out the Hébraud–Lequeux equation to us. We also warmly thank Pascal Hébraud and Claude Le Bris for helpful discussions.

REFERENCES

[1] J.P. Bouchaud, *Weak ergodicity breaking and aging in disordered systems*, J. Phys. I, 2 (1992), pp. 1705–1713.

[2] E. Cancès, I. Catto, Y. Gati, and C. Le Bris, *On a multiscale model for non-Newtonian flows*, Multiscale Model. Simul., to appear.

[3] E. Cancès and C. Le Bris, preprint, IMA, 2005.

[4] M. Chipot, *Elements of Nonlinear Analysis*, Birkhäuser Verlag, Basel, 2000.

[5] R.S. Fall, J.R. Melrose, and R.C. Ball, *Kinetic theory of jamming in hard-sphere startup flows*, Phys. Rev. E, 55 (1997), pp. 7203–7211.

[6] Y. Gati, *Numerical simulation of micro-macro model of concentrated suspensions*, Internat. J. Numer. Methods Fluids, Special Issue: ICFD Conference on Numerical Methods for Fluid Dynamics, 47 (2005), pp. 1019–1025.

[7] J.-P. Hansen and I.R. McDonald, *Theory of Simple Liquids*, Academic Press, London, 1976.

[8] P. Hébraud and F. Lequeux, *Mode coupling theory for the pasty rheology of soft glassy materials*, Phys. Rev. Lett., 81 (1998), pp. 2934–2937.

[9] R.G. Larson, *The Structure and Rheology of Complex Fluids*, Oxford University Press, London, 1998.

[10] J.-L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites non Linéaires*, Dunod, Paris, 1969.

[11] C. Monthus and J.P. Bouchaud, *Models of traps and glass phenomenology*, J. Phys. A, 29 (1996), pp. 3847–3869.

[12] J.S. Raynaud, P. Moucheront, J.C. Baudez, F. Bertrand, J.P. Guilbaud, and Ph. Coussot, *Direct determination by NMR of the thixotropic and yielding behavior of suspensions*, J. Rheol., 46 (2002), pp. 709–732.

[13] P. Sollich, *Rheological constitutive equation for a model of soft glassy materials*, Phys. Rev. E, 58 (1998), pp. 738–759.

[14] P. Sollich, F. Lequeux, P. Hébraud, and M. Cates, *Rheological of soft glassy materials*, Phys. Rev. Lett., 78 (1997), pp. 2020–2023.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 83–102

A BIRKHOFF–LEWIS-TYPE THEOREM FOR SOME HAMILTONIAN PDES*

DARIO BAMBUSI† AND MASSIMILIANO BERTI‡

Abstract. In this paper we give an extension of the Birkhoff–Lewis theorem to some semilinear PDEs. Accordingly we prove existence of infinitely many periodic orbits with large period accumulating at the origin. Such periodic orbits bifurcate from resonant finite dimensional invariant tori of the fourth order normal form of the system. Besides standard nonresonance and nondegeneracy assumptions, our main result is obtained assuming a regularizing property of the nonlinearity. We apply our main theorem to a semilinear beam equation and to a nonlinear Schrödinger equation with smoothing nonlinearity.

Key words. infinite dimensional Hamiltonian systems, periodic solutions, Birkhoff normal form, variational methods, perturbation theory

AMS subject classifications. 35B10, 37K50, 37K55, 58E30

DOI. 10.1137/S0036141003436107

1. Introduction. In 1933, Birkhoff and Lewis [9] (see also [16], [17]) proved their celebrated theorem on existence of periodic orbits with large period close to elliptic equilibria of Hamiltonian systems.[1] Here we give a generalization of their result to some semilinear Hamiltonian PDEs.

The Birkhoff–Lewis procedure consists in putting the system in fourth order (Birkhoff) normal form, namely in the form

$$H = H_0 + G_4 + R_5, \qquad H_0 := \sum_{j=1}^{n} \omega_j \frac{p_j^2 + q_j^2}{2}, \tag{1}$$

where G_4 is a homogeneous polynomial of degree 2 in the actions $I_j := (p_j^2 + q_j^2)/2$ and R_5 is a remainder having a zero of fifth order at the origin. Then system (1) is a perturbation of the integrable system H_0+G_4. Under a nondegeneracy condition (that also plays a fundamental role in KAM theory) the action-to-frequency map of this integrable system is one-to-one, and therefore there exist infinitely many resonant tori on which the motion is periodic. The question is: Do some of these periodic orbits persist under the perturbation due to the term R_5? Birkhoff and Lewis used the implicit function theorem and a topological argument to prove that there exists a sequence of resonant tori accumulating at the origin with the property that at least two periodic orbits bifurcate from each one of them.

In order to extend this result to infinite dimensional systems describing Hamiltonian PDEs one meets two difficulties: the first is the generalization of Birkhoff normal form to PDEs and the second is the appearance of a small denominator problem.

*Received by the editors October 13, 2003; accepted for publication (in revised form) July 23, 2004; published electronically August 17, 2005.
http://www.siam.org/journals/sima/37-1/43610.html

†Dipartimento di Matematica, Via Saldini 50, 20133 Milano, Italy (bambusi@mat.unimi.it). This author's research was supported by GNFM.

‡SISSA, Via Beirut 2-4, 34014, Trieste, Italy (berti@sissa.it). This author's reasearch was supported by MIUR "Metodi Variazionali ed Equazioni Differenziali Nonlineari."

[1]Actually [9] considers the neighborhood of an elliptic, nonconstant, periodic orbit, but the scheme is essentially the same for elliptic equilibria.

Here we work in a way which is as straightforward as possible, so, instead of considering the standard Birkhoff normal form of the system, whose extension to PDEs is not completely understood at present,[2] we consider its "seminormal form," namely the kind of normal form employed to construct lower dimensional tori. Precisely, having fixed a positive n, we split the phase variables into two groups, namely the variables with index smaller than n and the variables with index larger than n. We will denote by $\hat{z}$ the whole set of variables with index larger than n. We construct a canonical transformation putting the system in the form

$$H_0 + \overline{G} + \hat{G} + K, \tag{2}$$

where $\overline{G}$ depends only on the actions, $\hat{G}$ is at least cubic in the variables $\hat{z}$ with index larger than n, and K has a zero of sixth order at the origin. The interest of such a seminormal form is that the normalized system $H_0 + \overline{G} + \hat{G}$ has the invariant $2n$-dimensional manifold $\hat{z} = 0$ which is filled by n-dimensional invariant tori.

Under a nondegeneracy condition, the frequencies of the flow in such tori cover an open subset of $\mathbf{R}^n$. We concentrate on the resonant tori filled by periodic orbits, and we prove that at least n geometrically distinct periodic orbits of each torus survive the perturbation due to the term K. Since the orbits bifurcate from lower dimensional tori, we have to impose a further nondegeneracy condition in order to avoid resonances between the frequency of the periodic orbit and the frequencies of the transversal oscillations.

The proof is based on a variational Lyapunov–Schmidt reduction similar to that employed in [6] and inspired by [1]. It turns out that in the present case the range equation involves small denominators. To solve the corresponding problem we use an approach similar to that of [2]. In particular we impose a strong condition on the small denominators and we show that, if the vector field of the nonlinearity is smoothing, then the range equation can be solved by the contraction mapping principle. Next, the kernel equation is solved by noting that it is the Euler–Lagrange equation of the action functional restricted to the solutions of the range equation. The restricted functional turns out to be defined on $\mathbf{T}^n$, and so existence and multiplicity of solutions (critical points) follows by the classical Lusternik–Schnirelmann theory.

Finally, we apply the general theorem to the nonlinear beam equation

$$u_{tt} + u_{xxxx} + mu = f(u), \tag{3}$$

with Dirichlet boundary conditions on a segment. We consider m as a parameter varying in the segment $[0, L]$, and we show that the assumptions of the abstract theorem are fulfilled provided one excludes from the interval a finite number of values of m. As a second application we will deal with a nonlinear Schrödinger equation with a smoothing nonlinearity of the type considered in [20].

We recall that families of periodic solutions to Hamiltonian PDEs have been constructed by many authors (see, e.g., [14], [21], [19], [12], [10], [2]). The main difference is that the periodic orbits of the above quoted papers are a continuation of the linear normal modes to the nonlinear system. In particular their period is close to one of the periods of the linearized system. Moreover (except in the resonant case; see [5], [7], [8]) each periodic solution involves only one of the linear oscillators.[3]

On the contrary, the periodic orbits constructed in the present paper are the shadows of resonant tori; they are a purely nonlinear phenomenon, have long period,

[2] See, however, the recent works [3], [4].

[3] In the sense that all the other ones have a much smaller amplitude of oscillation.

and moreover each periodic motion involves n linear oscillators that oscillate with amplitudes of the same order of magnitude.

2. Main result. Consider a real Hamiltonian system with real[4] Hamiltonian function

$$H(z,\overline{z}) = \sum_{j\geq 1} \omega_j z_j \overline{z}_j + P(z,\overline{z}) \equiv H_0 + P, \tag{4}$$

where P has a zero of third order at the origin and the symplectic structure is given by $\mathrm{i}\sum_j dz_j \wedge d\bar{z}_j$. Here z and $\bar{z}$ are considered as independent variables. Often we will write only the equation for z since the equation for $\bar{z}$ is obtained by complex conjugation. The formal Hamiltonian vector field of the system is $X_H(z,\overline{z}) := (\mathrm{i}\frac{\partial H}{\partial \overline{z}_j}, -\mathrm{i}\frac{\partial H}{\partial z_j})$, and therefore the equations of motion have the form

$$\dot{z}_j = \mathrm{i}\omega_j z_j + \mathrm{i}\frac{\partial P}{\partial \overline{z}_j}\ , \quad \dot{\overline{z}}_j = -\mathrm{i}\omega_j \overline{z}_j - \mathrm{i}\frac{\partial P}{\partial z_j}. \tag{5}$$

Define the complex Hilbert space

$$\mathcal{H}^{a,s}(\mathbf{C}) := \left\{ w = (w_1, w_2, \ldots) \in \mathbf{C}^\infty \ \middle|\ ||w||^2_{a,s} := \sum_{j\geq 1} |w_j|^2 j^{2s} e^{2ja} < \infty \right\}.$$

We fix $s \geq 0$ and $a \geq 0$ and will study the system in the phase space

$$\mathcal{P}_{a,s} := \mathcal{H}^{a,s}(\mathbf{C}) \times \mathcal{H}^{a,s}(\mathbf{C}) \ni (z,\bar{z}).$$

Fix any finite integer $n \geq 2$ and denote $\omega := (\omega_1, \ldots, \omega_n)$, $\Omega := (\omega_{n+1}, \omega_{n+2}, \ldots)$. We assume that

(A) The frequencies grow at least linearly at infinity; namely there exist $a > 0$ and $d_1 \geq 1$ such that

$$\omega_j \sim a j^{d_1}.$$

(NR) For any $k \in \mathbf{Z}^n$, $l \in \mathbf{Z}^\infty$ with $|l| \leq 2$ and $0 < |k| + |l| \leq 5$, one has

$$\omega \cdot k + \Omega \cdot l \neq 0. \tag{6}$$

(S) There exist a neighborhood of the origin $\mathcal{U} \subset \mathcal{P}_{a,s}$ and $d \geq 0$ such that $X_P \in C^\omega(\mathcal{U}, \mathcal{P}_{a,s+d})$; namely it is analytic.

Remark 2.1. In applications to PDEs, property (S) is usually a consequence of the smoothness of the Nemitsky operator defined by the nonlinear part of the equation. In order to ensure (S) one has usually to restrict to the case where the functions with Fourier coefficients in $\mathcal{P}_{a,s}$ form an algebra (with the product of convolution between sequences). This imposes some limitations on the choice of the indexes a, s.

Proposition 2.1. *Assume* (A), (NR), (S). *There exists a real analytic, symplectic change of variables $\mathcal{T}$ defined in some neighborhood $\mathcal{U}' \subset \mathcal{P}_{a,s}$ of the origin, transforming the Hamiltonian H in seminormal form up to order six, namely into*

$$H \circ \mathcal{T} \equiv \mathcal{H} = H_0 + \overline{G} + \hat{G} + K \tag{7}$$

[4] i.e., if $\overline{z}$ is actually the complex conjugate to z, then the Hamiltonian H takes real values.

with

$$\overline{G} = \frac{1}{2} \sum_{\min(i,j)\le n} \overline{G}_{ij} |z_i|^2 |z_j|^2,$$

$\overline{G}_{ij} = \overline{G}_{ji}$, $\hat{G} = O(||\hat{z}||^3_{a,s})$, *where* $\hat{z} := (z_{n+1}, z_{n+2}, \dots)$ *and* $K = O(||z||^6_{a,s})$. *Moreover*

$$X_{\overline{G}}, X_{\hat{G}}, X_K \in C^\omega(\mathcal{U}', \mathcal{P}_{a,s+d}), \quad \|z - \mathcal{T}(z)\|_{a,s+d} \le C\|z\|^2_{a,s}. \tag{8}$$

We defer the proof of this proposition to the appendix where we also give a formula for $\bar{G}$ (see (69)).

The interest of such a seminormal form is that the system obtained by neglecting the reminder K has the invariant manifold $\hat{z} = 0$ on which the system is integrable.

As a variant with respect to the standard finite dimensional procedure we have left the third order term $\hat{G}$ but normalized the system up to order six (instead of five). This is needed in Lemma 3.2.

We also remark that, assuming just the nonresonance condition (NR), one can not hope (in general) to transform the Hamiltonian H into the infinite dimensional analogue of the standard Birkhoff normal form.

We rewrite the Hamiltonian $\mathcal{H}$ in the form

$$\mathcal{H} := \omega \cdot I + \Omega \cdot Z + \frac{1}{2} AI \cdot I + (BI, Z) + \hat{G} + K, \tag{9}$$

where $I := (|z_1|^2, \dots, |z_n|^2)$, $Z := (|z_{n+1}|^2, |z_{n+2}|^2, \dots)$ are the actions, A is the $n \times n$ matrix

$$A = (\overline{G}_{ij})_{1\le i,j\le n}, \tag{10}$$

and B is the $\infty \times n$ matrix

$$B = (\overline{G}_{ij})_{1\le j\le n<i}. \tag{11}$$

Remark 2.2. Due to (8), (9), one has $|(BI)_j| \le C|I|j^{-d}$ for a suitable C. Indeed, since $X_{\overline{G}}$ maps $\mathcal{P}_{a,s}$ to $\mathcal{P}_{a,s+d}$, the operator $z_j \mapsto (BI)_j z_j$ maps $\mathcal{P}_{a,s}$ to $\mathcal{P}_{a,s+d}$, and therefore its eigenvalues $(BI)_j$ must fulfill the above property.

Introduce action angle variables for the first n modes by $z_j = |z_j| e^{\mathrm{i}\phi_j} = \sqrt{I_j} e^{\mathrm{i}\phi_j}$ for $j = 1, \dots, n$.

Perform the rescaling $I_j \to \eta^2 I_j$, $\phi_j \to \phi_j$ for $j = 1, \dots, n$, $z_j \to \eta z_j$, $\overline{z}_j \to \eta \overline{z}_j$ for $j \ge n+1$ and divide the Hamiltonian by η^2. We get

$$\mathcal{H}(I, \phi, \hat{z}, \bar{\hat{z}}) = \omega \cdot I + \Omega \cdot Z + \eta \hat{G}_\eta + \eta^2 \Big(\frac{1}{2} AI \cdot I + (BI, Z) \Big) + \eta^4 K_\eta, \tag{12}$$

where $\hat{G}_\eta = O(||\hat{z}||^3_{a,s})$ and $K_\eta(z) = O(||z||^6_{a,s})$. We will still denote by $\mathcal{P}_{a,s} \equiv \mathbf{R}^n \times \mathbf{T}^n \times \mathcal{H}^{a,s} \times \mathcal{H}^{a,s}$ the phase space.

We will find periodic solutions of the Hamiltonian system (12) close to periodic solutions of the integrable Hamiltonian system

$$\dot{I} = 0, \quad \dot{\phi} = \omega + \eta^2 (AI + B^T Z), \quad \dot{z}_j = \mathrm{i}\Big(\Omega_j + \eta^2 (BI)_j\Big) z_j, \quad j \ge n+1 \tag{13}$$

in which $\hat{G}_\eta$ and K_η are neglected. The manifold $\{\hat{z}=0\}$ is invariant for the Hamiltonian system (13) and is completely filled up by the invariant tori

$$\mathcal{T}(I_0) := \{I = I_0,\ \phi \in \mathbf{T}^n,\ \hat{z}=0\}$$

on which the motion is linear with frequencies

$$\widetilde{\omega} \equiv \widetilde{\omega}(I_0) := \omega + \eta^2 A I_0 .$$

Such a torus is linearly stable, and the frequencies of small oscillation about the torus $\mathcal{T}(I_0)$ are the "shifted elliptic frequencies," namely

$$\widetilde{\Omega}_j(I_0) := (\Omega + \eta^2 B I_0)_j . \tag{14}$$

If all the $\widetilde{\omega}$'s are integer multiples of a single frequency, namely if

$$\widetilde{\omega} := \omega + \eta^2 A I_0 = \frac{1}{T} 2\pi k \in \frac{1}{T} 2\pi \mathbf{Z}^n , \tag{15}$$

then $\mathcal{T}(I_0)$ is a *completely resonant* torus, supporting the family of T-periodic motions

$$\mathcal{P} := \big\{ I(t) = I_0, \quad \phi(t) = \phi_0 + \widetilde{\omega} t, \quad \hat{z}(t) = 0 \big\}. \tag{16}$$

The whole family $\mathcal{P}$ will not persist in the dynamics of the complete Hamiltonian system (12). We will show that, under suitable assumptions, at least n geometrically distinct T-periodic solutions persist. More precisely, we will show that this happens for η small enough and for any choice of I_0 and T with

$$\|I_0\| \le C, \qquad \frac{1}{\eta^2} \le T \le \frac{2}{\eta^2}, \tag{17}$$

where C is independent of η, fulfilling the following.

(H1) Equation (15) holds.

(H2) There exist $\delta > 0$ and $\tau < d$ such that

$$|\widetilde{\Omega}_j T - 2\pi l| \ge \frac{\delta}{j^\tau} \quad \forall l \in \mathbf{Z},\ \forall j \ge n+1. \tag{18}$$

PROPOSITION 2.2. *Fix* $\tau > 1$. *Assume* (A), $\det A \neq 0$, *and*

$$\hat{\Omega}_j := \big(\Omega - B A^{-1} \omega\big)_j \neq 0 \qquad \forall j \ge n+1; \tag{19}$$

then, for any $\eta > 0$ *and almost any* T *fulfilling* (17) *there exists* I_0 *such that* (H1,H2) *hold.*

Proof. Fix η; we define $I_0 := I_0(T)$ as a function of T so that (15) is identically satisfied. Then we find T so that the nonresonance property (18) holds. Fix η and define

$$I_0 := I_0(T) := \frac{2\pi}{\eta^2 T} A^{-1} \Big(\Big[\frac{\omega T}{2\pi} \Big] - \frac{\omega T}{2\pi} \Big), \tag{20}$$

$$k := k(T) = \Big[\frac{\omega T}{2\pi} \Big], \tag{21}$$

where $[(x_1, \dots, x_n)] := ([x_1], \dots, [x_n])$ and $[x] \in \mathbf{Z}$ denotes the integer part of $x \in \mathbf{R}$. With the choice (20), (21), $\omega T + T\eta^2 A I_0 = 2\pi k$, and I_0 is of order 1 since $T\eta^2 \ge 1$.

We come to the nonresonance property (18). To study it, we remark that the function $T \to [\omega T/2\pi]$ is piecewise constant. Hence, for any $T_0 \in (\eta^{-2}, 2\eta^{-2})$ there exists an interval $\mathcal{I}_0 = (T_0 - a, T_0 + b) \subset [\eta^{-2}, 2\eta^{-2}]$ such that $[\omega T/2\pi] := k_0$ is constant for $T \in \mathcal{I}_0$. Moreover the union of such intervals covers the whole set of values in which we are interested. We will construct a subset of full measure of $\mathcal{I}_0$, in which condition (H2) is fulfilled.

So, for fixed j, l consider the set

$$B_{jl}(\tau, \delta) := \left\{ T \in \mathcal{I}_0 \ : \ |\widetilde{\Omega}_j T - 2\pi l| < \frac{\delta}{j^\tau} \right\}. \tag{22}$$

Remark that

$$\widetilde{\Omega}_j T = \hat{\Omega}_j T + \left(2\pi B A^{-1} \left[\frac{\omega T}{2\pi} \right] \right)_j ,$$

so that, in $\mathcal{I}_0$

$$\frac{d}{dT} \left(\widetilde{\Omega}_j T - 2\pi l \right) = \hat{\Omega}_j .$$

By (A) and Remark 2.2, there exists C such that

$$\left| \widetilde{\Omega}_j \right| \geq C j^{d_1} \ , \quad \left| \hat{\Omega}_j \right| \geq C j^{d_1} . \tag{23}$$

Then B_{jl} is an interval with length $|B_{jl}|$ controlled by

$$|B_{jl}| < 2 \frac{\delta}{C j^{\tau + d_1}} . \tag{24}$$

Fix j and estimate the number of l for which the set B_{jl} is (possibly) nonempty. First remark that, due to (A), one has that, as T varies in $\mathcal{I}_0$, the quantity $\widetilde{\Omega}_j T$ varies in a segment of length smaller than $C j^{d_1}$, with a suitable C. This means that there are at most $C j^{d_1}$ values of l which fall in such an interval (with redefined C). So, one has

$$\left| \bigcup_l B_{jl} \right| \leq \frac{C\delta}{j^\tau} . \tag{25}$$

Thus, provided $\tau > 1$ as we assumed, one has that

$$\left| \bigcup_{jl} B_{jl} \right| \leq C\delta . \tag{26}$$

By this estimate, the intersection over δ of such sets has zero measure, which is the thesis. ∎

THEOREM 2.3. *Consider the system* (9)*; let T and I_0 fulfill* (17) *and* (H1), (H2). *Then, provided η is small enough, there exist n geometrically distinct periodic orbits of the Hamiltonian system $\mathcal{H}$ (cf.* (9)*) with period T which are η^2 close in $\mathcal{P}_{a,s}$ to the torus $\mathcal{T}(I_0)$.*

Going back to the original system, one has the following corollary.

COROLLARY 2.4. *Consider the Hamiltonian system* (5) *and fix a positive n. Assume that* (A), (NR), (S) *hold, that* $\det A \neq 0$ *(cf.* (10)*), and that $\hat{\Omega}_j \neq 0$ for all $j \geq n+1$ (cf.* (19)*). Finally assume $d > 1$.*

Then, for any positive $\eta \ll 1$ there exist at least n distinct periodic orbits $z^{(1)}(t), \dots, z^{(n)}(t)$ with the following properties:

- $\|z^{(l)}(t)\|_{a,s} \le C\eta$ *for* $l = 1, \dots, n$ *and* $t \in \mathbf{R}$*;*
- $\|\Pi_{>n} z^{(l)}(t)\|_{a,s} \le C\eta^2$ *for* $l = 1, \dots, n$ *and* $t \in \mathbf{R}$*; here* $\Pi_{>n}$ *is the projector on the modes with index larger than* n*;*
- *the period* T *of* $z^{(l)}$ *does not depend on* l *and fulfills* $\eta^{-2} \le T \le 2\eta^{-2}$*.*

Remark 2.3. By Theorem 2.3 the results of Corollary 2.4 remain true even if the Hamiltonian system (5) does not fulfill one of the assumptions (A), (NR) but nevertheless one is able to transform it to the form (7).

Remark 2.4. If the integer numbers $(k_1, \dots, k_n) = \widetilde{\omega} T/2\pi$ are relatively prime, then T is the minimal period of the periodic solutions $z^{(l)}$. Indeed $z^{(l)}$ are T-periodic functions close to the functions defined in (16) which have minimal period T.

3. Proof of Theorem 2.3. Since the problem is Hamiltonian, any periodic solution of the system is a critical point of the action functional

$$S(I, \phi, \hat{z}, \hat{\bar{z}}) = \int_0^T \left(I \cdot \dot{\phi} + \mathrm{i} \sum_{j \ge n+1} z_j \dot{\bar{z}}_j - \mathcal{H}(I, \phi, \hat{z}, \hat{\bar{z}}) \right) dt \tag{27}$$

in the space of T-periodic, $\mathcal{P}_{a,s}$-valued functions. Here $\mathcal{H}$ is given by (12).

We look for a periodic solution $\zeta := (\phi, I, \hat{z}, \hat{\bar{z}})$ of the form

$$\phi(t) = \phi_0 + \widetilde{\omega} t + \psi(t), \quad I(t) = I_0 + J(t), \tag{28}$$

where $(\psi, J, \hat{z}, \hat{\bar{z}})$ are periodic functions of period T taking values in the covering space $\mathbf{R}^n \times \mathbf{R}^n \times \mathcal{H}^{a,s} \times \mathcal{H}^{a,s}$ of $\mathcal{P}_{a,s}$ (that for simplicity will still be denoted by $\mathcal{P}_{a,s}$). Hence $(\psi,\ J,\ \hat{z})$ must satisfy (in what follows for simplicity of notation we will only consider the equation for $\hat{z}$)

$$\nabla_\phi S(\zeta) = 0 \iff \dot{J} = R_\phi(\zeta), \tag{29}$$

$$\nabla_I S(\zeta) = 0 \iff \dot{\psi} - \eta^2 A\, J = R_I(\zeta), \tag{30}$$

$$\nabla_{\bar{z}_j} S(\zeta) = 0 \iff \dot{z}_j - \mathrm{i}\widetilde{\Omega}_j z_j = (R_{\bar{z}})_j(\zeta), \tag{31}$$

where

$$\left\{ \begin{array}{lcc} R_\phi(\zeta) & := & -\eta^4 \partial_\phi K_\eta(I_0 + J, \phi_0 + \widetilde{\omega} t + \psi, \hat{z}) - \eta \partial_\phi \hat{G}_\eta, \\ R_I(\zeta) & := & \eta^2 B^T Z + \eta^4 \partial_I K_\eta(I_0 + J, \phi_0 + \widetilde{\omega} t + \psi, \hat{z}) + \eta \partial_I \hat{G}_\eta, \\ (R_{\bar{z}})_j(\zeta) & = & \mathrm{i}\eta^2 (BJ)_j z_j + \mathrm{i}\eta \partial_{\bar{z}_j} \hat{G}_\eta + \mathrm{i}\eta^4 \partial_{\bar{z}_j} K_\eta(I_0 + J, \phi_0 + \widetilde{\omega} t + \psi, \hat{z}). \end{array} \right. \tag{32}$$

Remark that, since one expects J, ψ, and $\hat{z}$ to be small (they will turn out to be of order η^2), and by Proposition 2.1 one has $\partial_{\bar{z}} \hat{G}_\eta(\zeta) = O(\|\hat{z}\|^2_{a,s})$, $\partial_\phi \hat{G}_\eta(\zeta) = O(\|\hat{z}\|^3_{a,s})$, $\partial_I \hat{G}_\eta(\zeta) = O(\|\hat{z}\|^3_{a,s})$ the r.h.s. of (29), (30), (31) remain small even when they are multiplied by $T = O(\eta^{-2})$ (see Lemma 3.2).

Define the Hilbert space $H^1_P((0,T); \mathcal{P}_{a,s})$ of the T-periodic $\mathcal{P}_{a,s}$-valued periodic functions of class H^1. In order to simplify notations we will denote this space by $H^1_{P,s}$.

Denote for $\zeta = (\psi, J, w) \in H^1_{P,s}$,

$$|J|^2_{L^2,T} := \frac{1}{T} \int_0^T |J|^2 \ dt, \quad |\psi|^2_{L^2,T} := \frac{1}{T} \int_0^T |\psi|^2 \ dt, \tag{33}$$

$$||w||^2_{L^2,T,a,s} := \frac{1}{T} \int_0^T ||w(t)||^2_{a,s} \ dt, \tag{34}$$

$$\|\zeta\|_{L^2,T,a,s} := |J|_{L^2,T} + |\psi|_{L^2,T} + ||w||_{L^2,T,a,s}. \tag{35}$$

We will endow $H^1_P((0,T);\mathcal{P}_{a,s}) \equiv H^1_{P,s}$ with the norm

$$\|\zeta\|_{T,a,s} := \|\zeta\|_{L^2,T,a,s} + T\|\dot{\zeta}\|_{L^2,T,a,s}. \tag{36}$$

Remark 3.1. With this choice one has

$$\|\zeta(t)\|_{\mathcal{P}_{a,s}} \le C\|\zeta\|_{T,a,s} \quad \forall t \in \mathbf{R}$$

with a constant independent of T. Therefore, with this choice of the norm, the space $H^1_{P,s}$ is a "Banach algebra," and the T,a,s norm of the product of any component of a vector ζ with any component of a vector ζ' is bounded by $C\|\zeta\|_{T,a,s}\|\zeta'\|_{T,a,s}$ *with a constant* C *independent of* T.

We will consider the system (29), (30), (31) as a functional equation in $H^1_{P,s}$.

Remark 3.2. As a consequence of (8) and Remark 3.1, the map $\zeta \mapsto R(\zeta) := (R_\phi(\zeta), R_I(\zeta), R_{\overline{z}}(\zeta))$ is a C^∞ map from $H^1_{P,s}$ to $H^1_{P,s+d}$.

We are going to use the method of Lyapunov–Schmidt decomposition in order to solve (29), (30), (31). To this end remark that the kernel of the linear operator $\mathcal{L}$ at the l.h.s. of (29), (30), (31) is given by $(\phi,0,0)$ with constant $\phi \in \mathbf{T}^n$. The range of $\mathcal{L}$ is the space of the functions $\zeta = (\psi, J, \hat{z})$ with $\psi(t)$ having zero mean value. So, there is a natural decomposition of $H^1_{P,s}$ into Range+Kernel. Explicitly, we write

$$\zeta = (\phi + \psi, J, \hat{z}) = (\psi, J, \hat{z}) + (\phi, 0, 0) \equiv \zeta_R + \phi$$

with ψ having zero mean value and ϕ being constant. Then we fix ϕ, take the projection of the system (29), (30), (31) on the range, and solve it. The solution is a function $\zeta_R(\eta,\phi)$. Finally, we insert this function in the variational principle in order to find critical points of S.

3.1. The range equation. The range equation has the form

$$\begin{cases} \dot{J} & = & R_\phi(\zeta) - \langle R_\phi(\zeta)\rangle, \\ \dot{\psi} - \eta^2 A\, J & = & R_I(\zeta), \\ \dot{z}_j - \mathrm{i}\widetilde{\Omega}_j z_j & = & (R_{\overline{z}})_j(\zeta), \end{cases} \tag{37}$$

where $\langle R_\phi(\zeta)\rangle := (1/T)\int_0^T R_\phi(\zeta)dt$. We look for its solution in the range, namely in the space

$$\overline{H}^1_{P,s} \subset H^1_{P,s}$$

of the functions $\zeta_R \equiv (\psi, J, \hat{z})$ with ψ having zero average.

First of all we analyze the linear problem defined by the l.h.s. of (37). A "small denominator problem" appears since inverting this linear system the denominators $\widetilde{\Omega}_j T - 2\pi l$, $j \ge n+1$, $l \in \mathbf{Z}$ are present. So, define the linear operator

$$\mathcal{L}(\psi, J, \hat{z}) \equiv \mathcal{L}\zeta_R := (\dot{J}, \dot{\psi} - \eta^2 AJ, \dot{w}_j - \mathrm{i}\widetilde{\Omega}_j w_j)$$

and study

$$\mathcal{L}\zeta_R = (\widetilde{\psi}, \widetilde{J}, \widetilde{w}) \tag{38}$$

with $(\widetilde{\psi}, \widetilde{J}, \widetilde{w}) \in \overline{H}^1_{P,s+\tau}$ given.

LEMMA 3.1. *Assume* (H2). *If*

$$\widetilde{\zeta}_R \equiv (\widetilde{\psi}, \widetilde{J}, \widetilde{w}) \in \overline{H}^1_{P,s+\tau}, \qquad \text{i.e., in } H^1_{P,s+\tau} \text{ with } \int_0^T \widetilde{\psi}(t)\, dt = 0,$$

then (38) *has a unique solution,*

$$\zeta_R \equiv (\psi, J, w) \in \overline{H}^1_{P,s}.$$

Moreover for $T \in (\eta^{-2}, 2\eta^{-2})$ and a constant $C := C(\delta)$

$$\|\zeta_R\|_{T,a,s} \le \frac{C}{\eta^2} \|\widetilde{\zeta}_R\|_{T,a,s+\tau}.$$

Proof. Since A is symmetric and invertible it has an orthonormal basis of eigenvectors $e_1, \dots, e_n$ with eigenvalues $\lambda_1, \dots, \lambda_n$. In these coordinates $J(t) = \sum_{k=1}^n J_k(t) e_k$, $\psi(t) = \sum_{k=1}^n \psi_k(t) e_k$, and the solution ζ_R of (38) with $\psi_0 = 0$ has Fourier coefficients

$$J_{kl} = \frac{T\widetilde{\psi}_{kl}}{\mathrm{i}2\pi l} \quad \text{for} \quad l \neq 0, \qquad J_{k0} = -\frac{\widetilde{J}_{k0}}{\eta^2 \lambda_k},$$

$$\psi_{kl} = T\frac{\widetilde{J}_{kl} + \eta^2 J_{kl}\lambda_k}{\mathrm{i}2\pi l} \quad \text{for} \quad l \neq 0,$$

and, for $j \ge n+1$,

$$w_{jl} := \frac{T\widetilde{w}_{jl}}{\mathrm{i}(2\pi l - \widetilde{\Omega}_j T)}.$$

We then find

$$\begin{aligned} |J|^2_{L^2,T} = \sum_{kl} J^2_{kl} &= \sum_k \left(\frac{\widetilde{J}_{k0}}{\eta^2\lambda_k}\right)^2 + \sum_{k,l\neq 0} \left(\frac{T\widetilde{\psi}_{kl}}{\mathrm{i}2\pi l}\right)^2 \\ &\le \frac{C}{\eta^4} |\widetilde{J}|^2_{L^2,T} + CT^2 |\widetilde{\psi}|^2_{L^2,T}. \end{aligned} \tag{39}$$

A similar estimate for $|\psi|_{L^2,T}$ holds. Moreover

$$|\dot{\psi}|_{L^2,T} \le \eta^2 |J|_{L^2,T} + |\widetilde{\psi}|_{L^2,T} \le C(|\widetilde{J}|_{L^2,T} + |\widetilde{\psi}|_{L^2,T}), \tag{40}$$

using (39). Finally, the solution $w = (w_j)_{j \ge n+1}$ of (38) is

$$w_j(t) = \sum_{l\in\mathbf{Z}} \frac{T\widetilde{w}_{jl}}{\mathrm{i}(2\pi l - \widetilde{\Omega}_j T)} e^{\mathrm{i}(2\pi/T)lt},$$

where $\widetilde{w}_j(t) = \sum_{l\in\mathbf{Z}} \widetilde{w}_{jl} e^{\mathrm{i}(2\pi/T)lt}$. From (H2) we get

$$||w||_{L^2,T,a,s} \le C\frac{T}{\delta} ||\widetilde{w}||_{L^2,T,a,s+\tau}, \qquad ||\dot{w}||_{L^2,T,a,s} \le C\frac{T}{\delta} ||\dot{\widetilde{w}}||_{L^2,T,a,s+\tau}. \tag{41}$$

By (39), (40), and (41) the last estimate of the lemma follows. ∎

Thus $\mathcal{L}^{-1}$ defines a linear bounded operator $L : \overline{H}^1_{P,s+\tau} \to \overline{H}^1_{P,s}$.

In order to find a solution $\zeta_R = (\psi, J, \hat{z})$ of the range equation it is sufficient to find a fixed point of

$$\zeta_R = \Phi(\zeta_R) := L\big(N(\zeta_R; \phi)\big) \tag{42}$$

in the space $\overline{H}^1_{P,s}$, where $N := N(\zeta_R; \phi)$ denotes the r.h.s. of (37).

LEMMA 3.2. *Assume $d > \tau$. Then there exists a constant C sufficiently large such that $\forall \eta \ll 1$ the map Φ is a contraction of a ball of radius $C\eta^2$.*

Proof. Consider a $\zeta_R \in \overline{H}^1_{P,s}$ with $\|\zeta_R\|_{T,a,s} \le \rho$ with some positive (small) ρ. Since $H^1_{P,s}$ is an algebra with constants independent of T (cf. Remark 3.1), one has, by (32),

$$\|N(\zeta_R)\|_{T,a,s+d} \le C(\eta^4 + \eta\rho^2)$$

with a suitable C. Therefore, by Lemma 3.1 one has

$$\|\Phi(\zeta_R)\|_{T,a,s} \le \|\Phi(\zeta_R)\|_{T,a,s+d-\tau} \le C\left(\eta^2 + \frac{\rho^2}{\eta}\right),$$

which is smaller than ρ, provided $C(\eta^2 + \rho^2/\eta) < \rho$, which is implied, e.g., by $\rho = 2C\eta^2$ and η small enough.

Similarly one estimates the Lipschitz constant of Φ by the norm of its differential. Such a differential is bounded in a ball of radius ρ by $C(\eta^2 + \rho/\eta)$, from which the thesis follows. ∎

COROLLARY 3.3. *There exists a unique smooth function $\mathbf{T}^n \ni \phi \mapsto \zeta_R(\phi, \eta) \in \overline{H}^1_{P,s}$ solving* (37) *and fulfilling*

$$\|\zeta_R(\phi, \eta)\|_{T,a,s} \le C\eta^2.$$

3.2. The kernel equation. The geometric interpretation of the construction of the previous subsection is that we have found a submanifold $\mathcal{T}^n \equiv \{\zeta_{\phi_0} := (\phi_0 + \widetilde{\omega}t, I_0, 0) + \zeta_R(\phi_0, \eta),\ \phi_0 \in \mathbf{T}^n\} \subset H^1_{P,s}$, diffeomorphic to an n-dimensional torus, on which the partial derivatives of the action functional S with respect to the variables ζ_R vanish. We claim that at a critical point of S restricted to $\mathcal{T}^n$, all the partial derivatives of the complete functional S vanish and therefore that such a point is critical also for the nonrestricted functional.

Indeed, let $S_n : \mathbf{T}^n \to \mathbf{R}$ be the functional defined by $S_n(\phi_0) := S(\zeta_{\phi_0})\ \forall \phi_0 \in \mathbf{T}^n$.

LEMMA 3.4. *If $\phi_0 \in \mathbf{T}^n$ is a critical point of $S_n : \mathbf{T}^n \to \mathbf{R}$, then ζ_{ϕ_0} is a critical point of the nonrestricted functional S.*

Proof. Since $\zeta_R(\phi_0, \eta) = (\psi_{\phi_0}, J_{\phi_0}, \hat{z}_{\phi_0}) \in \overline{H}^1_{P,s}$ solves (37), then ζ_{ϕ_0} satisfies $\forall \phi_0 \in \mathbf{T}^n$,

$$\nabla_\phi S(\zeta_{\phi_0}) = \langle R_\phi(\zeta_{\phi_0})\rangle, \qquad \nabla_I S(\zeta_{\phi_0}) = 0, \qquad \nabla_{\overline{z}_j} S(\zeta_{\phi_0}) = 0 \tag{43}$$

(∇S denote the L^2-gradients). By (43) and since $\int_0^T \partial_{\phi_0}\psi_{\phi_0}(t)\, dt = 0\ \forall \phi_0 \in \mathbf{T}^n$,

$$\partial_{\phi_0} S_n(\phi_0) := (\nabla_\phi S(\zeta_{\phi_0}), \partial_{\phi_0}\zeta_{\phi_0})_{L^2} = T\langle R_\phi(\zeta_{\phi_0})\rangle.$$

Therefore, if $\phi_0 \in \mathbf{T}^n$ is a critical point of S_n, then $\langle R_\phi(\zeta_{\phi_0})\rangle = 0$, and at $\zeta_{\phi_0} \in \mathcal{T}^n$ all the partial derivatives of the complete functional S vanish. ∎

By standard Lusternik–Schnirelmann theory there exist at least n geometrically distinct T-periodic solutions, i.e., solutions not obtained from each other simply by time-translations. Indeed, restrict S_n to the plane $E := [\widetilde{\omega}]^{\perp}$ orthogonal to the periodic flow $\widetilde{\omega} = (1/T)2\pi k$ with $k \in \mathbf{Z}^n$. The set $\mathbf{Z}^n \cap E$ is a lattice of E, and hence S_n defines a functional $S_{n|\Gamma}$ on the quotient space $\Gamma := E/(\mathbf{Z}^n \cap E) \sim \mathbf{T}^{n-1}$.

Due to the invariance of S_n with respect to the time shift, a critical point of $S_{n|\Gamma}$ is also a critical point of $S_n : \mathbf{T}^n \to \mathbf{R}$. By the Lusternik–Schnirelmann category theory since $\mathrm{cat}\Gamma = \mathrm{cat}\mathbf{T}^{n-1} = n$, we can define the n min-max critical values $c_1 \leq c_2 \leq \cdots \leq c_n$ for $S_{n|\Gamma}$. If the critical levels c_i are all distinct, the corresponding T-periodic solutions are geometrically distinct, since their actions c_i are all different. On the other hand, if some min-max critical level c_i coincides, then, by the Lusternik–Schnirelmann theory, $S_{n|\Gamma}$ possesses infinitely many critical points. However *not* all the corresponding T-periodic solutions are necessarily geometrically distinct, since two different critical points could belong to the same orbit. Nevertheless, since a periodic solution can cross Γ at most a finite number of times, the existence of infinitely many geometrically distinct orbits follows. For further details, see [6].

This concludes the proof of Theorem 2.3.

4. Applications.

4.1. The nonlinear beam equation. Consider the beam equation

$$u_{tt} + u_{xxxx} + mu = f(u) \tag{44}$$

subject to hinged boundary conditions

$$u(0,t) = u(\pi,t) = u_{xx}(0,t) = u_{xx}(\pi,t) = 0, \tag{45}$$

where the nonlinearity $f(u)$ is a real analytic odd function of the form

$$f(u) = au^3 + \sum_{k \geq 5} f_k u^k, \qquad a \neq 0.$$

The beam equation (44) is a Hamiltonian PDE with associated Hamiltonian

$$H = \int_0^{\pi} \frac{u_t^2}{2} + \frac{u_{xx}^2}{2} + \frac{mu^2}{2} - g(u)\ dx,$$

where $g(u) := \int_0^u f(s)\ ds$ is a primitive of f.

Write the system in first order form

$$\begin{cases} \dot{u} &= \quad v, \\ \dot{v} &= -u_{xxxx} - mu + f(u). \end{cases} \tag{46}$$

The standard phase space[5] for (46) is $\mathcal{F}_s := H_C^s \times H_C^{s-2} \ni (u,v)$, where H_C^s is the space of the functions which extend to skew symmetric H^s periodic functions over $[-\pi,\pi]$. Note that $H_C^s = \{u(x) = \sum_{j\geq 1} u_j \sin(jx) \mid \sum_{j \geq 1} |u_j|^2 j^{2s} < +\infty\}$. It is then immediate to realize that, due to the regularity and skew symmetry of the vector field of the nonlinear part, f defines a smoothing operator, namely a smooth map from $\mathcal{F}_s$ to $\mathcal{F}_{s+2}$, provided $s \geq 1$.

[5] An equivalent definition makes use of the so-called compatibility conditions required for the smoothness of solutions of second order equations with Dirichlet boundary conditions; see, e.g., [11, Theorem X.8].

Here we are also interested in spaces of analytic functions, namely functions whose Fourier coefficients belong to $\mathcal{H}^{a,s}$ with some positive a. It is easy to see that the smoothing property of the nonlinearity holds also for these spaces.

Introduce coordinates $q = (q_1, q_2, \dots)$, $p = (p_1, p_2, \dots)$ through the relations

$$u(x) = \sum_{j \geq 1} \frac{q_j}{\sqrt{\omega_j}} \phi_j(x), \qquad v(x) = \sum_{j \geq 1} p_j \sqrt{\omega_j} \phi_j(x),$$

where $\phi_j(x) = \sqrt{2/\pi} \sin(jx)$ and

$$\omega_j^2 = j^4 + m. \tag{47}$$

Remark also that

$$\omega_j \sim j^2.$$

Passing to complex coordinates

$$z_j := \frac{q_j + \mathrm{i} p_j}{\sqrt{2}}, \qquad \bar{z}_j := \frac{q_j - \mathrm{i} p_j}{\sqrt{2}},$$

the Hamiltonian takes the form (4), and the nonlinearity fulfills (S) with $s \geq 1$ a suitable a, depending on the anayticity strip of f, and $d = 2$ (for more details, see [19], [13]).

In order to verify the nonresonance property we use m as a parameter belonging to the set $[0, L]$ with an arbitrary L.

LEMMA 4.1. *There exists a finite set $\Delta \subset [0, L]$ such that, if $m \in [0, L] \backslash \Delta$, then condition* (NR) *holds.*

Proof. First remark that, due to the growth property of the frequencies, there is at most a finite number of vectors $l \in \mathbf{Z}^2$ at which $\omega \cdot k + \Omega \cdot l$ is small. It follows that, having fixed an arbitrary constant C, there is at most a finite set of k's and l's over which $|\omega \cdot k + \Omega \cdot l| < C$. Denote by $\mathcal{S}$ such a set.

For $(k, l) \in \mathcal{S}$ consider

$$f_{kl}(m) = \omega(m) \cdot k + \Omega(m) \cdot l;$$

since f_{kl} is an analytic function, it has only isolated zeros. So at most finitely many of them fall in $[0, L]$. The set Δ is the union over $k, l \in \mathcal{S}$ of such points. Fix $m \in [0, L] \backslash \Delta$. $\square$

Then one can put the system in seminormal form. The explicit computation was essentially done in [15] (see also [19], [13]), obtaining that the matrices A and B are given by

$$A = \frac{6}{\pi} \begin{pmatrix} \frac{3}{\omega_1^2} & \frac{4}{\omega_1 \omega_2} & \cdots & \frac{4}{\omega_1 \omega_n} \\ \frac{4}{\omega_2 \omega_1} & \frac{3}{\omega_2^2} & \cdots & \frac{4}{\omega_2 \omega_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{4}{\omega_1 \omega_n} & \frac{4}{\omega_n \omega_2} & \cdots & \frac{3}{\omega_n^2} \end{pmatrix}, \; B = \frac{6}{\pi} \begin{pmatrix} \frac{4}{\omega_{n+1} \omega_1} & \cdots & \frac{4}{\omega_{n+1} \omega_n} \\ \frac{4}{\omega_{n+2} \omega_1} & \cdots & \frac{4}{\omega_{n+2} \omega_n} \\ \vdots & \vdots & \vdots \end{pmatrix}. \tag{48}$$

Remark that, defining the matrices

$$S_1 := \mathrm{diag}(\omega_1, \dots, \omega_n) \qquad \text{and} \qquad S_2 := \mathrm{diag}(\omega_{n+1}, \omega_{n+2}, \dots),$$

one can write $A = \frac{6}{\pi} S_1^{-1} \widetilde{A} S_1^{-1}$, $B = \frac{6}{\pi} S_2^{-1} \widetilde{B} S_1^{-1}$ with

$$\widetilde{A} = \begin{pmatrix} 3 & 4 & \dots & 4 \\ 4 & 3 & \dots & 4 \\ \dots & \dots & \dots & \dots \\ 4 & 4 & \dots & 3 \end{pmatrix}, \quad \widetilde{B} = \begin{pmatrix} 4 & \dots & 4 \\ 4 & \dots & 4 \\ \vdots & \vdots & \vdots \end{pmatrix}. \tag{49}$$

With these expressions at hand it is immediate to verify that $\det A \neq 0$. For what pertains $\hat{\Omega}_j$ (cf. (19)) by exactly the same argument in the proof of Lemma 4.1 one has that they are different from zero except for at most finitely many values of $m \in [0, L]$.

Thus, provided m does not belong to a finite subset of $[0, L]$, Theorem 2.4 and its Corollary 2.4 apply.

4.2. A nonlinear Schrödinger equation. Consider the space H^s_C as in the previous section. Following Pöschel [20] we define a smoothing operator as follows. Fix a sequence $\{\rho_j\}_{j\geq 1}$ with the property

$$\forall j \geq 1, \qquad \rho_j \neq 0 \;\text{ and }\; |\rho_j| \leq C j^{-d/2}, \quad d > 1. \tag{50}$$

Consider the even, 2π-periodic, real function $\rho(x) := \sum_j \rho_j \cos(jx)$ and define

$$\Gamma : H^s_C \to H^{s+d/2}_C, \quad \Gamma u := \rho * u, \tag{51}$$

where the star denotes convolution (it is defined first by extending the function u to an odd 2π-periodic function).

Remark 4.1. It is easy to see that, expanding u in Fourier series

$$u(x) = \sum_{j\geq 1} z_j \sqrt{\frac{2}{\pi}} \sin(jx),$$

the jth Fourier coefficient of Γu is proportional to $\rho_j z_j$.

Consider the Hamiltonian system with Hamiltonian function

$$H(u, \overline{u}) = \int_0^\pi |u_x|^2 + F\left(|\Gamma u|^2\right)\, dx \tag{52}$$

with F an analytic function having a zero of order 2 at the origin, i.e., $F''(0) \neq 0$.

The equations of motion are

$$-\mathrm{i} u_t = u_{xx} + \Gamma\left(F'\left(|\Gamma u|^2\right)\Gamma u\right). \tag{53}$$

Inserting the Fourier expansion of u, the Hamiltonian H takes the form (4),

$$H(z, \overline{z}) = \sum_{j\geq 1} \omega_j z_j \overline{z}_j + P_4(z, \overline{z}) + \text{higher order terms of degree at least 6},$$

with $\omega_j = j^2$,

$$P_4 = a \sum_{i_1, i_2, i_3, i_4 \in \mathbf{N}} F_{i_1 i_2 i_3 i_4} z_{i_1} z_{i_2} \overline{z}_{i_3} \overline{z}_{i_4}, \qquad a \neq 0, \tag{54}$$

$$F_{i_1 i_2 i_3 i_4} = \rho_{i_1}\rho_{i_2}\rho_{i_3}\rho_{i_4} \int_0^\pi \sin(i_1 x)\, \sin(i_2 x)\, \sin(i_3 x)\, \sin(i_4 x) dx. \tag{55}$$

Then the Hamiltonian vector field fulfills (S) with d given by (50).

It results in $F_{i_1 i_2 i_3 i_4} = 0$ unless $i_1 \pm i_2 \pm i_3 \pm i_4 = 0$ for some choice of the signs. Thus only a codimension 1 set of coefficients $F_{i_1 i_2 i_3 i_4}$ is actually different from zero, and the sum in (54) extends only over $i_1 \pm i_2 \pm i_3 \pm i_4 = 0$.

The nonresonance assumption (NR) is here violated. So one could expect the seminormal form (9) of Proposition 2.1 not to hold for this system. Indeed one could only expect to transform H into a "resonant normal form." Nevertheless, it turns out that such resonant normal form *depends on the actions only*, and so the Hamiltonian H can still be written in the form (9). Actually, even a stronger result holds (as in [15]): the Hamiltonian H can be brought into the infinite dimensional analogue of the classical Birkhoff normal form. More precisely, we have the following proposition.

PROPOSITION 4.2. *There exists a real analytic, symplectic change of variables $\mathcal{T}$ defined in some neighborhood $\mathcal{U}' \subset \mathcal{P}_{a,s}$ of the origin, transforming the Hamiltonian H into*

$$H \circ \mathcal{T} \equiv \mathcal{H} = H_0 + \overline{G} + K \tag{56}$$

with

$$\overline{G} = \frac{1}{2} \sum_{i,j \geq 1} \overline{G}_{ij} |z_i|^2 |z_j|^2, \qquad \overline{G}_{ij} = \alpha \rho_i^2 \rho_j^2 (4 - \delta_{ij}), \qquad \alpha \neq 0, \tag{57}$$

and $K = O(||z||^6_{a,s})$. Moreover

$$X_{\overline{G}}, X_{\hat{G}}, X_K \in C^\omega(\mathcal{U}', \mathcal{P}_{a,s+d}), \quad \|z - \mathcal{T}(z)\|_{a,s+d} \leq C \|z\|^2_{a,s}. \tag{58}$$

The proof follows section 3 of [15] and, for the reader's convenience, we reproduce it at the end of the appendix. The key ingredient is that the relevant divisors in the normalizing transformation are uniformly bounded away from 0 since they are nonvanishing integers.

Clearly the Birkhoff normal form Hamiltonian $\mathcal{H}$ given in (56) can be written also in the seminormal form (9). By (57), also in this case the matrices A and B (cf. (10)–(11)) have the structure $A = \alpha S_1 \widetilde{A} S_1$, $B = 2\alpha S_2 \widetilde{B} S_1$ with matrices $\widetilde{A}$ and $\widetilde{B}$ still given by (49), $S_1 := \operatorname{diag}(\rho_1^2, \ldots, \rho_n^2)$ and $S_2 := \operatorname{diag}(\rho_{n+1}^2, \rho_{n+2}^2, \ldots)$. So the determinant of A is still different from zero. The frequencies $\hat{\Omega}_j$ (cf. (19)) now have the structure

$$\hat{\Omega}_j(\rho) = j^2 - \rho_j^2 a(\rho) \quad \forall j \geq n+1,$$

where a is a function of $\rho_1^2, \ldots, \rho_n^2$. So, except for exceptional choices of $\{\rho_j\}_{j \geq n+1}$, the nondegeneracy conditions are fulfilled, Theorem 2.3 applies to the Hamiltonian system generated by $\mathcal{H}$ in (56), and (see Remark 2.3) Corollary 2.4 applies to (53).

5. Appendix: Proof of the normal form propositions.

Proof of Proposition. 2.1. The idea is to proceed as in the proof of the standard Birkhoff normal form theorem, i.e., by successive elimination of the nonresonant monomials. As a variant with respect to the standard procedure one does not eliminate terms which are at least cubic in the variables $\hat{z}$. Remark that the estimates involved in the proofs are much more complicated than in the finite dimensional case.

To start with, expand P in a Taylor series up to order five: $P = P_3 + P_4 + P_5$+higher order terms. Then we begin by looking for the transformation simplifying P_3. So write

$$P_3 = P_3^1 + \hat{G}_3(z)$$

with

$$\hat{G}_3(z) = O(\|\hat{z}\|^3),$$

and P_3^1 is composed by the first three terms of the Taylor expansion of P_3 in the variables $\hat{z}$ only (so it contains only terms of degree 0, 1, and 2 in such variables). We use the Lie transform to eliminate from P_3^1 all the nonresonant terms; i.e., we make a canonical transformation which is the time 1 flow Φ^1 of an auxiliary Hamiltonian system with a Hamiltonian function χ of degree 3. By considering the Taylor expansion of Φ^1 at zero, one has

$$H \circ \Phi^1 = H_0 + P_3^1 + \{\chi, H_0\} + O(\|z\|^4) + O(\|\hat{z}\|^3). \tag{59}$$

One wants to determine χ so that

$$\overline{G}_3 := P_3^1 + \{\chi, H_0\}$$

is a function of the actions $|z_j|^2$ only. Since $\overline{G}_3$ has to be a function of the actions only and moreover it is a polynomial of degree 3, it must vanish. To this end we proceed as usual in the theory of Birkhoff normal form.

Denote by $x = (x_1, \ldots, x_n) \equiv (z_1, \ldots, z_n)$ the first n variables and take χ to be a homogeneous polynomial of degree 3. Write

$$\chi = \sum_{|j_1|+|j_2|+|j_3|+|j_4|=3} \chi_{j_1 j_2 j_3 j_4} x^{j_1} \overline{x}^{j_2} \hat{z}^{j_3} \overline{\hat{z}}^{j_4} \tag{60}$$

with multi-indexes j_1, j_2, j_3, j_4. For a multi-index $j_l \equiv (j_{l,1}, \ldots, j_{l,n})$ we used the notation $|j_l| := |j_{l,1}| + \cdots + |j_{l,n}|$ and $x^{j_l} := x_1^{j_{l,1}}, \ldots, x_n^{j_{l,n}}$, and similarly for a multi-index with infinitely many components. So, one has

$$\{\chi, H_0\} = \sum_{|j_1|+|j_2|+|j_3|+|j_4|=3} \mathrm{i}\left(\omega \cdot (j_1 - j_2) + \Omega \cdot (j_3 - j_4)\right) \chi_{j_1 j_2 j_3 j_4} x^{j_1} \overline{x}^{j_2} \hat{z}^{j_3} \overline{\hat{z}}^{j_4}.$$

Write now

$$P_3^1 = \sum_{|j_1|+|j_2|+|j_3|+|j_4|=3} P_{j_1 j_2 j_3 j_4} x^{j_1} \overline{x}^{j_2} \hat{z}^{j_3} \overline{\hat{z}}^{j_4} \tag{61}$$

and remark that the indexes are here subjected to the further limitation $|j_3|+|j_4| \leq 2$. So, in order to have $\overline{G}_3 := P_3^1 + \{\chi, H_0\} = 0$, one is led to the choice

$$\chi_{j_1 j_2 j_3 j_4} := \frac{-P_{j_1 j_2 j_3 j_4}}{\mathrm{i}\left(\omega \cdot (j_1 - j_2) + \Omega \cdot (j_3 - j_4)\right)}, \quad j_1 - j_2 + j_3 - j_4 \neq 0, \tag{62}$$

and zero otherwise.

Since $|j_1| + |j_2| + |j_3| + |j_4| = 3$, then $0 < |j_1 - j_2| + |j_3 - j_4| \leq 5$ and so, due to assumption (NR) (recall also $|j_3| + |j_4| \leq 2$), the denominators appearing in (62) are all different from zero. Moreover, due to the growth of the frequencies ω_j (assumption (A)), they are actually bounded away from zero. Then in order to conclude the proof (at least for what concerns the elimination of the third order part) one has to ensure that the function χ is well defined and that it has a smooth Hamiltonian vector field. The terms of χ of different degree in $\hat{z}$ have to be treated in a different way, so we will denote by χ_0, χ_1, χ_2 the homogeneous parts of degree 0, 1, and 2, respectively, with respect to the variables $\hat{z}$.

We need a few lemmas.

LEMMA 5.1. *Let $\mathbf{R}^n \ni x \mapsto f(x) \in \ell^2$ be a homogeneous bounded polynomial of degree r. Write*

$$f(x) = \sum_{j\in\mathbf{N}^n,|j|=r} \sum_{k\geq 1} f_{jk} x^j e_k,$$

where e_k is the standard basis of ℓ^2. Let $\{\rho_{j,k}\}_{j\in\mathbf{N}^n}^{k\geq 1}$ be a sequence with the property $|\rho_{jk}| \geq C$, and define a function g by

$$g(x) = \sum_{j\in\mathbf{N}^n,|j|=r} \sum_{k\geq 1} \frac{f_{jk}}{\rho_{jk}} x^j e_k. \tag{63}$$

Then there exists $\overline{C}$ such that $\|g(x)\| \leq \overline{C}\|x\|^r$.

Proof. Write $g(x) = \sum_j g_j x^j$ and remark that the cardinality of the set over which the sum is carried out is finite. We estimate each of the vectors g_j's. Therefore, one has

$$\|g_j\|^2 = \sum_k \left(\frac{f_{kj}}{\rho_{kj}}\right)^2 \leq \frac{1}{C^2} \sum_k f_{kj}^2 = \frac{1}{C_0^2}\|f_j\|^2.$$

Now the norms of the vectors f_j are bounded, and therefore the thesis follows. ∎

Remark 5.1. By the same proof, the same result holds if the space ℓ^2 is substituted by the spaces $\mathcal{H}^{a,s}$.

LEMMA 5.2. *Let $\mathbf{R}^n \times \ell^2 \ni (x,z) \mapsto f(x,z) \in \mathbf{R}$ be a homogeneous bounded polynomial of degree r in x, linear and bounded in z. Write*

$$f(x,z) = \sum_{\substack{k\geq 1\\ j\in\mathbf{N}^n,|j|=r}} f_{jk} x^j z_k.$$

Let$\{\rho_{j,k}\}_{j\in\mathbf{N}^n}^{k\geq 1}$ be as above, and define a function g by

$$g(x,z) = \sum_{\substack{k\geq 1\\ j\in\mathbf{N}^n,|j|=r}} \frac{f_{jk}}{\rho_{jk}} x^j z_k. \tag{64}$$

Then there exists C such that $|g(x,z)| \leq C\|x\|^r\|z\|$.

Proof. Just write $g(x,z) = \sum_j g_j(z) x^j$. Fix j and study the linear functional $g_j(z)$; one has

$$|g_j(z)| = \left|\sum_{k\geq 1} f_{jk} \frac{z_k}{\rho_{jk}}\right| \leq \|f_j\| \left\|\frac{z}{\rho}\right\|,$$

where f_j is defined in analogy to g_j, its norm is the norm as a linear functional, and z/ρ is the vector of ℓ^2 with kth component equal to z_k/ρ_{jk}. From this inequality, summing over j, the thesis follows. ∎

In order to estimate the vector field of χ_2 we will need the following lemma.

LEMMA 5.3. [Lemma A. 1 of [18]]. *If $A = (A_{kl})$ is a bounded linear operator on ℓ^2, then also $B = (B_{kl})$ with*

$$B_{kl} := \frac{|A_{kl}|}{1 + |k - l|} \tag{65}$$

is a bounded linear operator on ℓ^2.

For the proof we refer to [18].

LEMMA 5.4. *Let $\mathbf{R}^n \times \ell^2 \ni (x, z) \mapsto f(x, z) \in \ell^2$ be a homogeneous bounded polynomial of degree r in x linear and bounded in z. Write*

$$f(x, z) = \sum_{\substack{k,l \geq 1 \\ j \in \mathbf{N}^n, |j| = r}} f_{jkl} x^j z_k e_l.$$

Let $\{\rho_{j,k,l}\}_{j \in \mathbf{N}^n}^{k,l \geq 1}$ be a sequence fulfilling,

$$|\rho_{jkl}| \geq C_1 (1 + |k - l|) \tag{66}$$

and define a function g by

$$g(x, z) = \sum_{\substack{k,l \geq 1 \\ j \in \mathbf{N}^n, |j| = r}} \frac{f_{jkl}}{\rho_{jkl}} x^j z_k. \tag{67}$$

Then there exists C such that $\|g(x, z)\| \leq C \|x\|^r \|z\|$.

Proof. Write $g(x, z) = \sum_j g_j(z) x^j$. Fix j and apply Lemma 5.3 to such operators, obtaining the result. □

Remark 5.2. An identical statement holds for functions from $\mathbf{R} \times \mathcal{H}_{a,s}$ to $\mathcal{H}_{a,s+d}$. To obtain the proof just remark that the boundedness of a linear operator $B = (B_{kl})$ (g_j in the proof) as an operator from $\mathcal{H}^{a,s}$ to $\mathcal{H}^{a,s+d}$ is equivalent to the boundedness of $\widetilde{B} := (v_k B_{kl} s_l)$ as an operator from ℓ^2 to itself, where v_k, s_l are suitable weights.

With the above lemmas at hand it easy to estimate the vector field of χ. We treat explicitly only χ_1.

LEMMA 5.5. *Let χ_1 be the component linear in $\hat{z}$ and $\overline{\hat{z}}$ of the function χ defined by* (62)*. Then there exists a constant C such that its vector field is bounded by*

$$\|X_{\chi_1}(z, \overline{z})\|_{a,s+d} \leq C \|z\|_{a,s}^2.$$

Proof. Write χ_1 as follows:

$$\chi_1(x, \overline{x}, \hat{z}, \overline{\hat{z}}) = \langle \chi_{01}(x, \overline{x}); \hat{z} \rangle_{\ell^2} + \left\langle \chi_{10}(x, \overline{x}); \overline{\hat{z}} \right\rangle_{\ell^2}.$$

Consider the first term. Separating the $x, \overline{x}$, and $\overline{\hat{z}}$ components, its vector field is given by

$$\left(\mathrm{i} \left\langle \frac{\partial \chi_{01}}{\partial \overline{x}}; \hat{z} \right\rangle_{\ell^2}, -\mathrm{i} \left\langle \frac{\partial \chi_{01}}{\partial x}; \hat{z} \right\rangle_{\ell^2}, -\mathrm{i} \chi_{01}(x, \overline{x}) \right).$$

Explicitly χ_{01} is given by

$$\sum_{\substack{|j_1| + |j_2| = 2 \\ l \geq n+1}} \frac{-P_{j_1 j_2 e_l}}{\mathrm{i}(\omega \cdot (j_1 - j_2) + \Omega_l)} x^{j_1} \overline{x}^{j_2} e_l.$$

It follows that each of the x (and $\overline{x}$) components of the vector field has the structure considered in Lemma 5.2, which therefore gives the estimate of such part of the vector field. Concerning the $\overline{\hat{z}}$ component, Lemma 5.1 applies and gives the result. The remaining components can be treated exactly in the same way. ∎

The estimate of the vector fields of χ_0 and χ_2 are obtained in a similar way. In order to apply Lemma 5.4 to the estimate of the vector field of χ_2 one has just to remark that from (A) and (NR) one has the estimate

$$|\omega \cdot k + \Omega_j - \Omega_l| \geq C(1 + |j - l|).$$

Thus we have the following proposition.

PROPOSITION 5.6. *The vector field of the function* χ *defined by* (62) *fulfills the inequality*

$$\|X_\chi(z, \overline{z})\|_{a,s+d} \leq C\|z\|^2_{a,s}.$$

Then by standard existence and uniqueness theory one has that such vector fields define a unique smooth time 1 flow in a neighborhood of the origin both in $\mathcal{P}_{a,s}$ and in $\mathcal{P}_{a,s+d}$. It follows that the transformation is well defined. Transforming the vector field of H, one gets a vector field having the same smoothness properties of the original one. Moreover the transformed Hamiltonian will have the form

$$\widetilde{H} := H \circ \Phi^1 = H_0 + \hat{G}_3 + \widetilde{P}_4 + \widetilde{P}_5 + \cdots, \tag{68}$$

where $\widetilde{P}_j$ is a homogeneous polynomial of degree j. In particular it turns out that

$$\widetilde{P}_4 = P_4 + \{\chi, P_3^1\} + \frac{1}{2}\{\chi, \{\chi, H_0\}\} = P_4 + \frac{1}{2}\{\chi, P_3^1\} + \left\{\chi, \hat{G}_3\right\}$$

since $P_3 = P_3^1 + \hat{G}_3$ and, by the definition of χ, $P_3^1 + \{\chi, H_0\} = 0$.

Thus one can iterate the construction and eliminate the unwanted terms of degree 4. Define $\widetilde{P}_4^1$ (in analogy to P_3^1), setting $\widetilde{P}_4 = \widetilde{P}_4^1 + \hat{G}_4(z)$, where $\hat{G}_4(z) = O(\|\hat{z}\|^3)$ and $\widetilde{P}_4^1$ is composed of the first three terms of the Taylor expansion of $\widetilde{P}_4$ in the variables $\hat{z}$ only. Next we perform a new canonical transformation which is the time 1 flow of another auxiliary Hamiltonian $\widetilde{\chi}$ of degree 4, such that $\widetilde{P}_4^1 + \{\widetilde{\chi}, H_0\}$ is a function of the actions $|z_j|^2$ only. Remark that in this case a nonvanishing normalized part of the Hamiltonian exists (in general) since $\omega \cdot (j_1 - j_2) + \Omega \cdot (j_3 - j_4) = 0 \ \forall \ j_1 = j_2$, $j_3 = j_4$, $|j_1| + |j_2| + |j_3| + |j_4| = 4$. Since, by (NR), $\omega \cdot (j_1 - j_2) + \Omega \cdot (j_3 - j_4) \neq 0$ for all the remaining $|j_1| + |j_2| + |j_3| + |j_4| = 4$, $|j_3| + |j_4| \leq 2$, the normalized part $\overline{G}_4$ is explicitly given by

$$\overline{G}_4 := \sum_{|j_1|+|j_3|=2, |j_3|\leq 1} \widetilde{P}^1_{j_1 j_1 j_3 j_3} |x|^{2j_1} |\hat{z}|^{2j_3}, \tag{69}$$

where $\widetilde{P}^1_{j_1 j_2 j_3 j_4}$ are the coefficients of $\widetilde{P}_4^1$ and $|x|^2 := x\overline{x}$, $|\hat{z}|^2 := \hat{z}\overline{\hat{z}}$. $\overline{G}_4$ is the function $\overline{G}$ introduced in (7). The regularity and estimates for the canonical transformation generated by $\widetilde{\chi}$ are obtained as before.

Finally, using again (NR), one iterates the construction with an auxiliary Hamiltonian of degree 5, eliminating all the terms of order 5 and concluding the proof of Proposition 2.1 (getting no further contributions to the function $\overline{G}$ which is of order 4).

Remark 5.3. If the Hamiltonian contains only monomials of even degree then the terms of order 3 and 5 are not present in the Hamiltonian. Thus by just *one* symplectic change of coordinates it is possible to eliminate the nonnormalized terms of order 4, and the remaining higher order terms are yet of order 6.

Proof of Proposition 4.2. The proof follows Lemma 4 of [15] (and Proposition 2.1).

The Hamiltonian function H contains only monomials of even degree, and the 4th order term of the nonlinearity is given explicitly by (54). Therefore, let us define the auxiliary Hamiltonian χ of degree 4,

$$\chi := \sum_{i_1 \pm i_2 \pm i_3 \pm i_4 = 0, \{i_1, i_2\} \neq \{i_3, i_4\}} \frac{-aF_{i_1 i_2 i_3 i_4}}{\mathrm{i}(i_1^2 - i_2^2 + i_3^2 - i_4^2)} z_{i_1} z_{i_2} \overline{z}_{i_3} \overline{z}_{i_4}.$$

By Lemma 5 of [15], if $i_1 \pm i_2 \pm i_3 \pm i_4 = 0$ and the nonordered pair $\{i_1, i_2\} \neq \{i_3, i_4\}$, then $i_1^2 - i_2^2 + i_3^2 - i_4^2 \neq 0$, and so χ is well defined.

The fourth order term of the transformed Hamiltonian via the time 1 flow map generated by χ is given by

$$\begin{aligned} P_4 + \{\chi, H_0\} &= \sum_{i_1 \pm i_2 \pm i_3 \pm i_4 = 0} \Big(aF_{i_1 i_2 i_3 i_4} - \mathrm{i}(i_1^2 - i_2^2 + i_3^2 - i_4^2) \chi_{i_1 i_2 i_3 i_4} \Big) \\ &= \sum_{\{i_1, i_2\} = \{i_3, i_4\}} aF_{i_1 i_2 i_3 i_4} z_{i_1} z_{i_2} \overline{z}_{i_3} \overline{z}_{i_4} \\ &= \frac{\alpha}{2} \sum_{i,j \geq 1} \rho_i^2 \rho_j^2 (4 - \delta_{ij}) |z_i|^2 |z_j|^2 := \overline{G} \end{aligned}$$

recalling (55) and since $F_{ijij} = a\rho_i^2 \rho_j^2 \int_0^\pi \sin^2(ix) \sin^2(jx) dx = (8a/\pi) \rho_i^2 \rho_j^2 (2 + \delta_{ij})$.

The estimates for the vector field generated by χ and the corresponding time 1 flow map can be carried out as in Proposition 2.1.

Finally, note that the remaining terms, which constitute the higher order term K, are yet of order 6 or more (see Remark 5.3).

Acknowledgments. The authors thank L. Biasco, P. Bolle, and M. Procesi for interesting discussions.

REFERENCES

[1] A. Ambrosetti, V. Coti-Zelati, and I. Ekeland, *Symmetry breaking in Hamiltonian systems*, J. Differential Equations, 67 (1987), pp. 165–184.

[2] D. Bambusi, *Lyapunov center theorems for some nonlinear PDE's: A simple proof*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 29 (2000), pp. 823–837.

[3] D. Bambusi, *Birkhoff normal form for some nonlinear PDEs*, Comm. Math. Phys., 234 (2003), pp. 253–285.

[4] D. Bambusi and B. Grebert, *Forme normale pour NLS en dimension quelconque*, C.R. Acad. Sci. Paris Ser. I Math., 337 (2003), pp. 409–414.

[5] D. Bambusi and S. Paleari, *Families of periodic solutions for resonant PDE's*, J. Nonlinear Science, 11 (2001), pp. 69–87.

[6] M. Berti, L. Biasco, and E. Valdinoci, *Periodic orbits close to elliptic tori and applications to the three-body problem*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 3 (2004), pp. 87–138.

[7] M. Berti and P. Bolle, *Periodic solutions of nonlinear wave equations with general nonlinearities*, Comm. Math. Phys., 243 (2003), pp. 315–328.

[8] M. Berti and P. Bolle, *Multiplicity of periodic solutions of nonlinear wave equations*, Nonlinear Anal., 56 (2004), pp. 1001–1046.

[9] G. D. Birkhoff and D. C. Lewis, *On the periodic motions near a given periodic motion of a dynamical system*, Ann. Mat. Pura Appl., (4), 12 (1933), pp. 117–133.

[10] J. Bourgain, *Construction of periodic solutions of nonlinear wave equations in higher dimension*, Geom. Funct. Anal., 5 (1995), pp. 629–639.
[11] H. Brezis, *Analyse fonctionelle*, Masson, Paris, 1993.
[12] W. Craig and C. E. Wayne, *Newton's method and periodic solutions of nonlinear wave equations*, Comm. Pure Appl. Math., 46 (1993), pp. 1409–1498.
[13] J. Geng and J. You, *KAM tori of Hamiltonian perturbations of* $1D$ *linear beam equations*, J. Math. Anal. Appl., 277 (2003), pp. 104–121.
[14] S. B. Kuksin, *Nearly integrable infinite-dimensional Hamiltonian systems*, Springer-Verlag, Berlin, 1993.
[15] S. B. Kuksin and J. Pöschel, *Invariant Cantor manifolds of quasi-periodic oscillations for a nonlinear Schrödinger equation*, Ann. of Math., 143 (1996), pp. 149–179.
[16] D. C. Lewis, *Sulle oscillazioni periodiche d'un sistema dinamico*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl., 19 (1934), pp. 234–237.
[17] J. Moser, *Proof of a generalized form of a fixed point theorem due to G. D. Birkhoff*, in Geometry and Topology (Proc. III Lat. Am. Sch. Math., Inst. Mat. Pura Aplicada CNP, Rio de Janeiro, 1976), Lecture Notes in Math., 597, Springer-Verlag, Berlin, 1977, pp. 464–494. .
[18] J. Pöschel, *A KAM-theorem for some nonlinear partial differential equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 23 (1996), pp. 119–148.
[19] J. Pöschel, *Quasi-periodic solutions for a nonlinear wave equation*, Comment. Math. Helv., 71 (1996), pp. 269–296.
[20] J. Pöschel, *On the construction of almost periodic solutions for a nonlinear Schrödinger equation*, Ergodic Theory Dynam. Systems, 22 (2002), pp. 1537–1549.
[21] C. E. Wayne, *Periodic and quasi-periodic solutions of nonlinear wave equations via KAM theory*, Commun. Math. Phys., 127 (1990), pp. 479–528.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 103–129

SCALAR CONSERVATION LAWS WITH MIXED LOCAL AND NONLOCAL DIFFUSION-DISPERSION TERMS*

CHRISTIAN ROHDE†

Abstract. We consider a nonlinear scalar conservation law that is regularized by a local viscous term and a *nonlocal* dispersive term. This nonstandard regularization is motivated by phase transition problems that take into account long range interactions close to the interface.

We identify a parameter regime such that this mixed-type regularization provides a new example that is able to drive *nonclassical undercompressive shock waves* in the limit of vanishing regularization parameter. In view of the applications this shows that nonlocal regularizations can be used to model dynamical phase transition processes.

In the next step we establish the existence and uniqueness of classical solutions for the Cauchy problem in multiple space dimensions. In the main part of the paper we then deduce appropriate a priori estimates to analyze the sharp-interface limit for vanishing regularization parameter with the method of compensated compactness in one space dimension and, using measure-valued solutions, in multiple space dimensions. It is shown that the limits exist and are weak solutions of the corresponding Cauchy problem for the hyperbolic conservation law.

Key words. nonlocal free energy, diffusion-dispersion, sharp-interface limit, undercompressive shock waves

AMS subject classifications. 35L65, 35L67

DOI. 10.1137/S0036141004443300

1. Introduction.

In this paper we are concerned with weak solutions of the Cauchy problem for the scalar conservation law

$$u_t(\mathbf{x},t) + \operatorname{div}(\mathbf{f}(u(\mathbf{x},t))) = 0 \quad ((\mathbf{x},t) \in \mathbb{R}^d \times \mathbb{R}_{>0},\ d \in \mathbb{N}). \tag{1.1}$$

The unknown function $u : \mathbb{R}^d \times \mathbb{R}_{>0} \to \mathbb{R}$ is a function of the d-dimensional space variable $\mathbf{x} = (x_1, \dots, x_d)^T$ and time $t > 0$. By $\mathbf{f} = (f_1, \dots, f_d)^T : \mathbb{R} \to \mathbb{R}^d$ we denote the vector of flux functions. It is well known that the Cauchy problem for equations of type (1.1) cannot have global smooth solutions for arbitrary initial datum. Prototype discontinuous solutions that satisfy the Cauchy problem in a weak sense are planar shock waves. Let us consider for simplicity the one-dimensional case $d = 1$ with a scalar flux function. If we choose the flux function as a convex function the only physically relevant type of shock waves are so-called Laxian (or classical) shock waves. If the flux function is not assumed to be convex also a *nonclassical type of shock wave* can arise: the undercompressive shock wave. While a Laxian shock wave on the level of (1.1) is completely compressive, i.e., the characteristics run into the discontinuity, for an undercompressive shock only one of the characteristics crosses the shock line. In recent years many important examples of problems have been identified that involve undercompressive shock waves: combustion fronts, precursors in thin film flow, kinematic waves in suspensions, and, most notably for this paper, phase transitions. In applications, the conservation law (1.1) is equipped with dissipative terms that lead to a regularization of the weak solutions of (1.1). Let us denote the

*Received by the editors April 26, 2004; accepted for publication (in revised form) August 30, 2004; published electronically August 17, 2005.
http://www.siam.org/journals/sima/37-1/44330.html

†Albert-Ludwigs-Universität Freiburg, Institut für Mathematik, Hermann-Herder Str. 10, D-79104 Freiburg, Germany (chris@mathematik.uni-freiburg.de).

regularized problem by

$$u_t^\varepsilon(\mathbf{x},t) + \operatorname{div}(\mathbf{f}(u^\varepsilon(\mathbf{x},t))) = R[\varepsilon; u^\varepsilon](\mathbf{x},t) \quad ((\mathbf{x},t) \in \mathbb{R}^d \times \mathbb{R}_{>0}), \tag{1.2}$$

where $\varepsilon > 0$ is a small regularization parameter.

It is an interesting question to identify physically relevant spatial regularization operators R such that

(i) the Cauchy problem for (1.2) has a global smooth solution u^ε in $\mathbb{R}^d \times \mathbb{R}_{>0}$,

(ii) the sequence $\{u^\varepsilon\}_{\varepsilon>0}$ converges for $\varepsilon \to 0$ almost everywhere to a limit function $u \in L^1_{loc}(\mathbb{R}^d \times \mathbb{R}_{>0})$ which is a weak solution of (1.1),

(iii) the function u contains undercompressive shock waves with reasonable physical interpretation.

For $w \in C^2(\mathbb{R}^d)$ consider the simplest possible local dissipation term

$$R[\varepsilon; w](\mathbf{x}) := \varepsilon \Delta w(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d.$$

The fundamental work of Kruzkov shows that tasks (i) and (ii) hold true for this choice of R ([24]). However, the regularization does rule out undercompressive shocks in the limit $\varepsilon \to 0$ which are known to contradict the Kruzkov entropy condition. Another picture results if one takes into account local diffusion and dispersion. For $w \in C^3(\mathbb{R}^d)$ and $\gamma > 0$ let

$$R[\varepsilon; w](\mathbf{x}) := \varepsilon \Delta w(\mathbf{x}) + \gamma \varepsilon^2 \sum_{j=1}^{d} \Delta w_{x_j}(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d. \tag{1.3}$$

This regularization is motivated from phase transition problems like, e.g., the dynamics of liquid-vapor transitions governed by the Navier–Stokes–Korteweg equations ([5, 18, 2, 11, 17, 34, 14]). There the physical density ρ^ε takes the rôle of u^ε. While the second order term in (1.3) stands for effects of viscosity, the third order term models the effects of surface tension forces close to the phase boundary. It can be directly referred to the free energy functional

$$E^\varepsilon_{local}[\rho^\varepsilon] = \int_{\mathbb{R}^d} W(\rho^\varepsilon(\mathbf{x})) + \gamma \varepsilon^2 \frac{|\nabla \rho^\varepsilon(\mathbf{x})|^2}{2} \, d\mathbf{x}. \tag{1.4}$$

Here W is the free energy function having double well structure. Minima of E_{local} describe static equilibrium solutions.

On the level of the model problem (1.2) with R from (1.3) it has been shown that undercompressive shock waves can occur in the limit of vanishing regularization. We refer to the papers [4, 19, 28, 22, 21, 32, 20, 23] and moreover to the monograph [27]. A special feature is that the limit depends on the number γ describing the ratio between diffusion and dispersion: shock waves are regularization-sensitive. The occurrence of undercompressive shocks has also been analyzed for local nonlinear and fourth order regularizations ([7]).

If one looks to the derivation of the free energy functional E^ε_{local} one observes that it is a simplification (see even the first paper on diffuse interface theories by Van der Waals: [36]). On a more basic level one has to consider the nonlocal energy

$$E^\varepsilon_{global}[\rho^\varepsilon] = \int_{\mathbb{R}^d} W(\rho^\varepsilon(\mathbf{x})) + \frac{\gamma}{4} \int_{\mathbb{R}^d} \phi_\varepsilon(\mathbf{x}-\mathbf{y})(\rho^\varepsilon(\mathbf{x}) - \rho^\varepsilon(\mathbf{y}))^2 \, d\mathbf{y} d\mathbf{x}. \tag{1.5}$$

For background on nonlocal energies in the framework of phase transition problems, we refer to [25, 1, 15, 30, 31] and references therein. To pass to dynamical models usually the action principle is used. This approach can be performed for local and nonlocal energies. In the first case for E_{local} equation (1.2) with choice (1.3) would be the most basic model problem. In the nonlocal case it means that we have to choose for $w \in C^2(\mathbb{R}^d)$

$$R[\varepsilon; w](\mathbf{x}) := \varepsilon \Delta w(\mathbf{x}) + \gamma \sum_{j=1}^{d} \left[[\phi_\varepsilon * w_{x_j}](\mathbf{x}) - w_{x_j}(\mathbf{x}) \right], \quad \mathbf{x} \in \mathbb{R}^d. \tag{1.6}$$

Here we have used for an arbitrary function $v \in L^1_{loc}(\mathbb{R}^d, \mathbb{R})$ the convolution $\phi_\varepsilon * v$ given by

$$[\phi_\varepsilon * v](\mathbf{x}) = \int_{\mathbb{R}^d} \phi_\varepsilon(\mathbf{x} - \mathbf{y}) v(\mathbf{y})\, d\mathbf{y}.$$

The kernel function ϕ_ε is defined by

$$\phi_\varepsilon(\mathbf{x}) = \frac{1}{\varepsilon^d} \phi\left(\frac{\mathbf{x}}{\varepsilon}\right). \tag{1.7}$$

ϕ is a smooth even nonnegative function with $\operatorname{supp}(\phi) \subset [-1, 1]^d$. Moreover let $\int_{\mathbb{R}^d} \phi(\mathbf{x})\, d\mathbf{x} = 1$. The kernel function ϕ_ε models the long-range interaction.

To show that (1.3) is in fact a simplification of (1.6), assume for simplicity $d = 1$ and that w is a sufficiently smooth function. By Taylor expansion of w around $x = x_1$ we obtain

$$[\phi_\varepsilon * w](x) - w(x) = \frac{1}{\varepsilon} \int_{\mathbb{R}} \phi\left(\frac{x - y}{\varepsilon}\right) (w(y) - w(x))\, dy = \sum_{i=1}^{\infty} c_{2i} \varepsilon^{2i} w^{(2i)}(x),$$

where $c_{2i} = \int_{\mathbb{R}} y^{2i} \phi(y)\, dy/(2i)!$. Neglecting the terms with $i > 1$ we obtain the third order derivative in (1.3) with $d = 1$.

In this paper we start the analysis of the three issues (i), (ii), (iii) for the new regularization (1.6). First we focus on planar wave solutions for (1.2), (1.6) in section 2. We show in particular that as the local model (1.2), (1.3) also the nonlocal model admits traveling wave solutions which correspond to undercompressive shock waves. Nonlocal dispersion is strong enough to drive these nonclassical shocks. Physically speaking the model is able to permit phase transitions. We observe also a striking sensitive dependence on the ratio parameter γ as in the local case.

In section 3 we establish the existence and uniqueness of smooth solutions for the Cauchy problem to (1.2), (1.6) in multiple space dimensions. Essentially we state that the L^2-norm is a Lyapunov functional for the flow associated with the evolution equation (1.2). Furthermore we remark that the existence is achieved using techniques for second order parabolic problems and not (as would be necessary in the case of the local regularization (1.3)) with the much more complicated techniques for third order equations. The main result is given in Theorem 3.2.

Finally in section 4 we study the limit $\varepsilon \to 0$ for general Cauchy problems. For the spatially one-dimensional case it turns out that one can obtain a priori estimates with respect to the L^2- and L^4-norm that are uniform in the parameter ε. These two estimates have also been identified in the local case. The dissipation estimates on spatial derivatives which come along with the estimates on the solution itself are

not as strong as in the local case. However, the obtained uniform bounds are strong enough to apply the method of compensated compactness. We show in the proof of Theorem 4.5 that the limit process can be performed in a (strong) L^p-norm and the limit function is moreover a weak solution of (1.1). We conclude with a study of the limit in the multidimensional case in section 5 based on measure-valued solutions.
In all sections we compare the results and used techniques with those for the local model (1.2), (1.3).

2. Traveling-wave analysis for the scalar model problem. In this section we seek special solutions u^ε of (1.2) with regularization given by (1.6): planar waves. It suffices to consider the one-dimensional version with $f := f_1 \in C^2(\mathbb{R})$ and $x := x_1$:

$$u_t^\varepsilon + f(u^\varepsilon)_x = \alpha\varepsilon u_{xx}^\varepsilon + \big[\phi_\varepsilon * u^\varepsilon - u^\varepsilon\big]_x \quad \text{in } \mathbb{R}\times(0,\infty). \tag{2.1}$$

The kernel ϕ_ε is defined as in (1.7) with $d = 1$. Note that we have introduced the factor $\alpha \geq 0$ in (2.1) while γ from (1.6) has been set to be 1. Shifting the weight for the ratio between local diffusion and nonlocal dispersion to the diffusion term turns out to be more convenient in this section. To include the purely dispersive case we allow $\alpha = 0$.

2.1. Shock waves for the limit problem. We start the investigations with a review on different types of shock wave solutions for the limit problem which we obtain by neglecting the terms on the right-hand side of (2.1):

$$u_t + f(u)_x = 0 \quad \text{in } \mathbb{R}\times(0,\infty). \tag{2.2}$$

Let some states $u_- \in \mathbb{R}$ and $u_+ \in \mathbb{R}$ with $u_- \neq u_+$ be given and define the speed $s \in \mathbb{R}$ by the ratio

$$s = \frac{f(u_-) - f(u_+)}{u_- - u_+}. \tag{2.3}$$

Then the discontinuous function $U : \mathbb{R}\times(0,\infty) \to \mathbb{R}$ given by

$$U(x,t) := \begin{cases} u_- : x - st \leq 0 \\ u_+ : x - st > 0, \end{cases} \tag{2.4}$$

is a weak solution of (2.2). We call it a **shock wave** provided $f'(u_\pm) \neq s$. If the states u_- and u_+ satisfy the condition

$$f'(u_-) > s > f'(u_+), \tag{2.5}$$

the function U is called a **compressive (or Laxian) shock wave**. If either

$$s < f'(u_-),\ s < f'(u_+) \text{ or } s > f'(u_-),\ s > f'(u_+) \tag{2.6}$$

holds the associated shock wave U is called **undercompressive**. An entropy solution of (2.2) is a function $u \in L^\infty_{loc}(\mathbb{R}\times(0,\infty))$ such that the entropy inequality

$$\int_0^T \int_{\mathbb{R}} \eta(u(x,t))\psi_t(x,t) + q(u(x,t))\psi_x(x,t)\,dxdt \geq 0 \tag{2.7}$$

holds for all entropy pairs (η, q) and all $\psi \in C_0^\infty(\mathbb{R}\times(0,\infty))$. Here an entropy pair (η, q) consists of the entropy function $\eta \in C^2(\mathbb{R})$, supposed to satisfy $\eta'' \geq 0$, and the entropy flux $q \in C^2(\mathbb{R})$, supposed to satisfy $\eta' f' = q'$. We remark that the discontinuous solution U is an entropy solution if and only if the inequality

$$D_{(\eta,q)}(u_-, u_+) := -s(\eta(u_+) - \eta(u_-)) + q(u_+) - q(u_-) \leq 0 \tag{2.8}$$

is satisfied for all entropy pairs. It is well known that for convex fluxes a shock wave is an entropy solution of (2.2) if and only if it is compressive ([8]). However, the condition (2.7) is quite restrictive since in the systems case often only one entropy pair is available.

Let us now consider (2.2) for the nonconvex cubic flux $f(u) = u^3$ and choose—for simplicity—the state $u_- \in (0, \infty)$ as an arbitrary but fixed number. Then we get from (2.3) the equation

$$s = u_-^2 + u_- u_+ + u_+^2. \tag{2.9}$$

Conditions (2.5) and (2.7) are satisfied for $u_+ \in (-u_-/2, u_-)$. We neglect the characteristic case $u_+ = u_-/2$ which does not correspond to a shock wave in the sense of a strict inequality as required in (2.5) or (2.6). For $u_+ \in (-\infty, -u_-/2)$, we observe that only one of the characteristics runs into the propagating shock wave. The first case in (2.6) applies and the associated shock wave U is undercompressive. Straightforward calculations reveal the following statements.

(i) Let u_+ be in $(-u_-, -u_-/2)$. For the entropy pair $(\eta(u), q(u)) = (u^2/2, 3u^4/4)$ we have

$$D_{(\eta,q)}(u_-, u_+) = \frac{1}{4}(u_+ - u_-)^2(u_+^2 - u_-^2) < 0.$$

(ii) Let u_+ be in $(-u_-, -\tau u_-)$ for $\tau \in (1/2, 1)$ and let $l \in \mathbb{N} \setminus \{2\}$ be even. For the entropy pair

$$(\eta_l(u), q_l(u)) = \begin{cases} (u^l/l, 3u^{l+2}/(l+2)) : u < 0, \\ \qquad (0, 0) \qquad\qquad : u \geq 0, \end{cases}$$

we have

$$\begin{aligned} D_{(\eta_l,q_l)}(u_-, u_+) &= \eta_l(u_+)\Big(2\frac{l-1}{l+2}u_+^2 - u_- u_+ - u_-^2\Big) \\ &> 2\eta_l(u_+)u_-^2\Big(\frac{l-1}{l+2}\tau^2 + \frac{1}{2}\tau - \frac{1}{2}\Big). \end{aligned}$$

Thus for each $\tau \in (1/2, 1)$ there is an $l_0 \in \mathbb{N}$ such that for $l > l_0$ we have

$$D_{(\eta_l,q_l)}(u_-, u_+) > 0.$$

(iii) Let u_+ with $u_+ < -u_-$ be given. For the entropy pair from (i) we have

$$D_{(\eta,q)}(u_-, u_+) > 0. \tag{2.10}$$

Thus U is for $u_+ < -u_-/2$ not an entropy solution anymore, but if u_+ is not too small, i.e., not less or equal than $-u_-$, an entropy inequality holds for the entropy pair with quadratic entropy. Exactly this entropy is important for our regularized problem (cf. Theorem 3.2 and estimate (3.5) below).

These considerations can be done for any nonconvex flux function. For a more extensive study we refer to [26]. The possible occurrence of undercompressive shock waves is the reason why (1.1) can be seen as a simple model for phase transitions. Note that phase transitions in realistic models can take the form of undercompressive waves (cf. [17] for instance).

2.2. Nonlocal diffusion-dispersion profiles. Let $u_-, u_+ \in \mathbb{R}$, compute s through (2.3), and define the shock wave U by (2.4). The function $v \in C^2(\mathbb{R})$ is called a **nonlocal diffusion-dispersion profile for the shock wave** U if it solves

(2.11)
$$\begin{aligned} \alpha\dot{v}(\xi) + ([\phi * v](\xi) - v(\xi)) &= f(v(\xi)) - sv(\xi) - (f(u_-) - su_-) \quad (\xi \in \mathbb{R}), \\ v(\pm\infty) &= u_\pm. \end{aligned}$$

If there is a nonlocal diffusion-dispersion profile for some shock wave U we can define the traveling-wave function $u^\varepsilon \in C^2(\mathbb{R} \times (0,\infty))$ through

$$u^\varepsilon(x,t) = v\left(\frac{x - st}{\varepsilon}\right) \quad ((x,t) \in \mathbb{R} \times \mathbb{R}_{>0}). \tag{2.12}$$

We see that u^ε is a classical solution of (2.1). Moreover we have for $(x,t) \in \mathbb{R} \times (0,\infty)$ with $x - st \neq 0$

$$\lim_{\varepsilon \to 0} u^\varepsilon(x,t) = U(x,t) = \begin{cases} u_- : x - st < 0, \\ u_+ : x - st > 0. \end{cases} \tag{2.13}$$

It is the issue of this section to show that for our nonlocal equation there are traveling waves such that the weak solution U represents an undercompressive shock wave.

NOTE 2.1. *The problem* (2.11) *itself does not depend on ε. This is a consequence of the scaling with respect to ε between the local diffusive and nonlocal dispersive term. For other scalings, e.g. $\gamma = \gamma(\varepsilon)$, the problem* (2.11) *can depend on ε. We shall see below that the chosen scaling leads to nonlocal diffusion-dispersion profiles for shock waves that are undercompressive shock waves. We conjecture that all other scalings lead to either classical behavior excluding undercompressive waves or purely dispersive behavior without any distinguished wave structure.*

The problem (2.11) belongs to the class of ordinary but nonlocal boundary value problems. These kinds of problems have been studied by [3, 12, 9], for instance. Our analysis relies on the result by Bates et al. ([3]) which we recall now.

THEOREM 2.2. *For $r \in \mathbb{N}$ let the function $\phi \in C^r(\mathbb{R},[0,\infty)) \cap W^{r,1}(\mathbb{R})$ be even and satisfy*

$$\int_{\mathbb{R}} \phi(x)\,dx = 1, \quad \int_{\mathbb{R}} \phi(x)|x|dx < \infty. \tag{2.14}$$

Let $u_-, u_+ \in \mathbb{R}$ and $F \in C^r(\mathbb{R})$ be given. For the unknowns $v : \mathbb{R} \to \mathbb{R}$ and $\alpha \in \mathbb{R}$ consider the problem

$$\begin{aligned} \alpha\dot{v}(\xi) + ([\phi * v](\xi) - v(\xi)) &= F(v(\xi)) \quad (\xi \in \mathbb{R}), \\ v(\pm\infty) &= u_\pm. \end{aligned} \tag{2.15}$$

We suppose that the states $u_\pm \in \mathbb{R}$ and the function F satisfy

$$\begin{aligned} &\text{(i)} \quad u_- > u_+, \\ &\text{(ii)} \quad F(u_\pm) = 0,\ F'(u_\pm) > 0, \\ &\text{(iii)} \quad \exists!\ u_0 \in (u_+, u_-)\ :\ F(u_0) = 0, \\ &\text{(iv)} \quad u \in [u_+, u_-] \Rightarrow F'(u) + 1 > 0. \end{aligned} \tag{2.16}$$

Then exactly one of the following statements holds true.

(a) *There is a monotonely decreasing function* $v \in C^{r+1}(\mathbb{R})$ *and a unique* $\alpha \in \mathbb{R} \setminus \{0\}$ *such that* (2.15) *holds.*
In this case we have

$$\alpha = H(u_-, u_+) \left(\int_{-\infty}^{\infty} (\dot{v}(\xi))^2 \, d\xi \right)^{-1}, \quad H(u_-, u_+) := \int_{u_-}^{u_+} F(u) \, du. \tag{2.17}$$

(b) *There is a monotonely decreasing function* $v \in C^r(\mathbb{R})$ *such that* (2.11) *holds for* $\alpha = 0$.
In this case we have

$$H(u_-, u_+) = 0.$$

In both cases the function v *is unique up to translation.*

Proof. See Theorems 2.7, 3.1, and 4.1 in [3]. ∎

The differential equation in problem (2.15) should be seen as the nonlocal counterpart to

$$\alpha \dot{v}(\xi) + \ddot{v}(\xi) = F(v(\xi)) \quad (\xi \in \mathbb{R}). \tag{2.18}$$

Then it is clear that condition (ii) from Theorem 2.2 ensures that u_- and u_+ are hyperbolic rest points of saddle type. With (iii) we just simplify the problem to exclude having more than the (necessary) three equilibria. The local problem (2.18) is a classical bistable equation for which the existence and bifurcation of traveling waves (local diffusion-dispersion profiles) is completely understood. Condition (iv) is not a necessary condition for the existence of orbits (cf. [3, Theorem 3.1]). However, if it is violated, the relation $H(u_-, u_+)$ in case (b) is substituted by a more complicated relation. Since here we only want to show that the nonlocal regularization is strong enough to drive nonclassical undercompressive waves, we do not pursue this issue. We return to our problem (2.11) and define

$$F = F(v) := f(v) - sv - (f(u_-) - su_-) \tag{2.19}$$

for $v \in \mathbb{R}$. A necessary structural assumption on f to construct undercompressive shock waves is the *nonconvexity* of f as we have seen in section 2.1. We choose again

$$f(u) = u^3. \tag{2.20}$$

Furthermore we require that the states u_-, u_+ satisfy

$$u_- > 0 > -\frac{u_-}{2} > u_+. \tag{2.21}$$

Using (2.9) we see that $s < f'(u_-)$ and $s < f'(u_+)$ holds. Thus the function U defined by (2.4) is an undercompressive shock wave. The choice $u_+ = -u_-/2$ leads to the excluded characteristic case $f'(u_+) = s$. We now prove the existence of a nonlocal diffusion-dispersion profile for U.

PROPOSITION 2.3. *We consider the problem* (2.11) *with cubic flux* (2.20) *and choose* $u_-, u_+ \in \mathbb{R}$ *such that* (2.21) *holds. We suppose that the kernel function* ϕ *is even, in the set* $C^1(\mathbb{R}, [0, \infty)) \cap W^{1,1}(\mathbb{R})$, *and satisfies*

$$\int_{\mathbb{R}} \phi(x) \, dx = 1, \quad \int_{\mathbb{R}} \phi(x)|x| dx < \infty. \tag{2.22}$$

Moreover, if the states u_-, u_+ *fulfill*

$$(u_+ + u_-)^2 - u_- u_+ < 1, \tag{2.23}$$

we have

(i) *for* $u_+ \in (-u_-, -u_-/2)$ *that there is a function* $v \in C^2(\mathbb{R})$ *and a unique* $\alpha > 0$ *solving* (2.11),

(ii) *for* $u_+ = -u_-$ *that there is a function* $v \in C^1(\mathbb{R})$ *solving* (2.11) *with* $\alpha = 0$.

The functions v *are unique up to translation.*

Proof. With the assumptions of the proposition we can apply Theorem 2.2 with $r = 1$. We have to verify the conditions (i),…,(iv) in (2.16). Condition (i) is satisfied due to (2.21). The Rankine–Hugoniot relation (2.3) and (2.9) imply $F(u_\pm) = 0$. For the derivative $F'(v) = 3v^2 - s$ we obtain with the upper bound for u_+ in (2.21) the estimates

$$F'(u_-) = 2u_-^2 - u_- u_+ - u_+^2 > u_-^2 - u_- u_+ > 0,$$
$$F'(u_+) = 2u_+^2 - u_- u_+ - u_-^2 > \tfrac{1}{2}u_-^2 + \tfrac{1}{2}u_-^2 - u_-^2 = 0.$$

Thus condition (ii) in Theorem 2.2 holds. Since F from (2.19) is cubic the condition (iii) is a consequence of $F'(u_\pm) > 0$. The function $F'(u) + 1$ has a minimum in $u_{min} = 0 \in [u_+, u_-]$ with

$$F(u_{min}) = 1 - (u_+ + u_-)^2 + u_- u_+.$$

This is condition (iv) due to (2.23).

We can apply the theorem and compute in our case for H defined in (2.17) the value

$$\begin{aligned} H(u_-, u_+) &= \int_{u_-}^{u_+} u^3 - su - (u_-^3 - su_-)\, du \\ &= -\frac{1}{4}\big(u_+^4 - 2u_+^3 u_- + 2u_+ u_-^3 - u_-^4\big) \\ &= -\frac{1}{4}(u_+ + u_-)(u_+ - u_-)^3. \end{aligned}$$

For fixed u_- the function $H(u_-, .)$ vanishes in $u_+ = -u_-$ and is positive in $(-u_-, -u_-/2)$. This implies the two statements of the proposition since exactly one of the two cases (a), (b) in Theorem 2.2 must hold true. ∎

Some remarks are in order.

NOTE 2.4.

(i) *The condition* (2.23) *can be satisfied for sufficiently small values of* u_-. *It holds for instance in the case* $u_- = 1$. *Together with the discussion in section* 2.1 *on undercompressivity we have shown that nonlocal dispersive terms can drive nonclassical shock waves of this type. We stress that condition* (2.23) *is used to verify* (iv) *in Theorem* 2.2 *which is not a necessary assumption for the existence of solutions.*

(ii) *Using the local term* (1.3) *in* (1.2) *leads to the study of* (2.18) *for traveling waves. As long as condition* (2.23) *holds we get exactly the same result on the existence of nonlocal diffusion-dispersion profile as in the case of the local dispersion term. Moreover the coefficient* α *for given* u_- *and* u_+ *is the same in the local and nonlocal case. We refer to* [21, 19]. *In the nonlocal case the results do not rely on the form of the kernel* ϕ_ε *as long as the assumptions in* (2.14) *are satisfied. This freedom in modelling is not present for the local regularization.*

(iii) *According to Proposition* 2.3 *the case* $u_+ = -u_-$ *can be solved only with* $\alpha = 0$ *(or equivalently* $\gamma = \infty$*) . This is of no interest to us since we want to have a positive parameter in front of the diffusion term in the original regularization* (1.6)*. However, it is interesting to see that this situation is detected by the analysis. The choice* $u_+ = -u_-$ *in the definition of the shock wave* U *is the extreme nonclassical one. If one takes* u_+ *strictly less than* $-u_-$ *the entropy condition* (2.7) *does not hold for the quadratic entropy function (cf.* (2.10)*). The overall importance of this entropy becomes clear in sections* 3, 4, *and* 5*. It is a Lyapunov functional for solutions* $u^\varepsilon(.,t)$ *of the Cauchy problem to* (1.2) *with* (1.6)*.*

(iv) *Theorem* 2.2 *is applied in* [3] *to a Ginzburg–Landau-type problem. However, the meaning of the parameter* α *is completely different. While here it is the diffusion parameter it plays the rôle of the traveling-wave speed in* [3]*. In our case the wave speed parameter is given a priori by* (2.3)*.*

3. Existence of classical solutions for the scalar model problem. In this section we consider (1.2) for the regularization (1.6), that is,

$$u_t^\varepsilon + \operatorname{div}(\mathbf{f}(u^\varepsilon)) = \varepsilon\Delta u^\varepsilon + \gamma\sum_{j=1}^{d}\left[\phi_\varepsilon * u_{x_j}^\varepsilon - u_{x_j}^\varepsilon\right] \quad \text{in } \mathbb{R}^d \times (0,T), \tag{3.1}$$

with initial condition

$$u^\varepsilon(.,0) = u_0 \text{ in } \mathbb{R}^d. \tag{3.2}$$

The time $T > 0$ is an arbitrary but fixed number and we recall the definition of ϕ_ε from (1.7). Under appropriate conditions on the initial function u_0, the flux $\mathbf{f}$, and the function ϕ specified below, we prove the existence of classical solutions for the Cauchy problem (3.1), (3.2).

3.1. Some notations and assumptions. For a multi-index $\alpha = (\alpha_1, \dots, \alpha_d)^T \in \mathbb{N}_0^d$, we define the differential operator

$$D^\alpha := \frac{\partial^{\alpha_1 + \cdots \alpha_d}}{\partial_{x_1}^{\alpha_1}\cdots\partial_{x_d}^{\alpha_d}}.$$

For the partial derivative with respect to the space coordinate x_j, $j = 1, \dots, d$, we also use the symbol ∂_{x_j}. The open set Ω_R, $R \in [0,T]$ is defined by $\Omega_R = \mathbb{R}^d \times (0,R)$. The following assumption on the flux functions and the function ϕ hold throughout the following two sections on the scalar model problem if nothing else is stated explicitly.

ASSUMPTION 3.1.

(i) *There is a number* $k > d/2 + 1$ *such that* $f_1, \dots, f_d \in C^k(\mathbb{R})$*. Furthermore, for each* $r > 0$ *let* $L_{\mathbf{f}}(r) > 0$ *be a number such that*

$$\sum_{j=1}^{d}|f_j'(u)| \le L_{\mathbf{f}}(r) \qquad (|u| \le r).$$

(ii) *The function* $\phi \in C^\infty(\mathbb{R}^d)$ *is even, bounded, nonnegative, and satisfies*

$$\int_{\mathbb{R}^d}\phi(\mathbf{x})\,d\mathbf{x} = 1, \quad \operatorname{supp}(\phi) \subset [-1,1]^d.$$

We introduce the set $K_0(r)$ for $r > 0$ by

$$K_0(r) = \{w \in L^\infty(\mathbb{R}^d) \,|\, \|w\|_{L^\infty(\mathbb{R}^d)} \leq r\}.$$

For $R > 0$ we denote by $C_1^2(\Omega_R)$ the set of real-valued functions on $\bar{\Omega}_R$ such that all spatial derivatives up to order two and the first order time derivative exist and are continuous in Ω_R. A function $u^\varepsilon \in C([0,R]; L^\infty(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)) \cap C_1^2(\Omega_R)$ is called a **classical solution of** (3.1)**,** (3.2) **in** Ω_R if it satisfies (3.1) in Ω_R and (3.2) almost everywhere in $\mathbb{R}^d$.

For $\varepsilon > 0$ we also introduce the operator $q^\varepsilon : L^\infty(\mathbb{R}^d) \to L^\infty(\mathbb{R}^d)$ with

$$q^\varepsilon[w] = \gamma(\phi_\varepsilon * w - w) \quad (w \in L^\infty(\mathbb{R}^d)).$$

We note that q^ε satisfies for almost all $\mathbf{x} \in \mathbb{R}^d$ the simple estimate

$$|q^\varepsilon[w](\mathbf{x})| \leq 2\gamma \|w\|_{L^\infty(\mathbb{R}^d)}. \tag{3.3}$$

3.2. The main result. The main result in this section is the following theorem.

THEOREM 3.2 (global existence). *Let $r_\infty > 0$ be given and suppose that Assumption* 3.1 *holds. Then there exists a number $r_2 = r_2(r_\infty) > 0$ such that for all*

$$u_0 \in K_0(r_\infty) \cap L^2(\mathbb{R}^d) \tag{3.4}$$

with

$$\|u_0\|_{L^2(\mathbb{R}^d)} \leq r_2$$

there is a classical solution u^ε of (3.1), (3.2) *in Ω_T that is unique in the set of classical solutions. We have for $t \in [0,T]$*

$$\frac{1}{2}\|u^\varepsilon(.,t)\|^2_{L^2(\mathbb{R}^d)} + \varepsilon \sum_{j=1}^d \|\partial_{x_j} u^\varepsilon\|^2_{L^2(\Omega_t)} = \frac{1}{2}\|u_0\|^2_{L^2(\mathbb{R}^d)}. \tag{3.5}$$

The proof of the theorem will be presented in section 3.4 below and relies on a local-in-time existence result presented in section 3.3. We add some comments. The smallness condition on the L^2-norm of u_0 is a consequence of the fact that (3.1) is not equipped with a maximum principle if $\gamma > 0$ holds. This cannot be expected in the framework of local or nonlocal dispersive regularizations.

In this section the dependence of results and estimates on ε does not play a role so that we take $\varepsilon > 0$ as arbitrary but fixed. However, we note for later use that (3.5) provides an ε-independent estimate on the $L^\infty([0,T]; L^2(\mathbb{R}^d))$-norm of u^ε and for $\sqrt{\varepsilon}\|\nabla u^\varepsilon\|_{L^2(\mathbb{R}^d \times [0,T])}$.

3.3. Local existence. At first we provide a local-in-time existence and uniqueness result for (3.1), (3.2) which we then extend to arbitrary times. For both steps we use a method which has been developed for parabolic systems that do not allow for an a priori L^∞-bound ([33, Chapter 6.2]). It is one of the advantages of regularization (1.6) in comparison to (1.3) that the more flexible second order theory can still be applied while the local counterpart requires dealing with a more complicated third order differential operator.

We make use of the scaled heat kernel

$$K_\varepsilon(\mathbf{x},t) = (4\pi\varepsilon t)^{-d/2} \exp\Big(-\frac{|\mathbf{x}|^2}{4\pi\varepsilon t}\Big) \quad ((\mathbf{x},t) \in \Omega_T).$$

For $t > 0$ we collect the following properties for K_ε:

(3.6)

$$\int_{\mathbb{R}^d} K_\varepsilon(\mathbf{x},t)\,d\mathbf{x} = 1,$$

$$\exists\, C_\varepsilon > 0: \; \sum_{j=1}^{d} \|\partial_{x_j} K_\varepsilon(.,t) * w\|_{L^p(\mathbb{R}^d)} \leq \frac{C_\varepsilon}{\sqrt{t}} \|w\|_{L^p(\mathbb{R}^d)}, \quad w \in L^p(\mathbb{R}^d),\; p = 2, \infty.$$

Now let $r_\infty \in \mathbb{R}$ and $\bar{r} \in \mathbb{R}$ be numbers such that

$$\bar{r} > r_\infty > 0. \tag{3.7}$$

Note that these numbers $\bar{r}$ and r_∞ are arbitrary but fixed throughout the section. Finally we also fix a number $\bar{t} \in (0,T)$ such that

$$r_\infty \;+ 2C_\varepsilon(L_{\mathbf{f}}(\bar{r}) + 2\gamma d)\bar{r}\sqrt{\bar{t}} < \bar{r} \tag{3.8}$$

and

$$2C_\varepsilon(L_{\mathbf{f}}(\bar{r}) + 2\gamma d)\sqrt{\bar{t}} < 1 \tag{3.9}$$

hold.

Lemma 3.3. *Suppose that Assumption* 3.1 *holds and let* $u_0 \in K_0(r_\infty) \cap L^2(\mathbb{R}^d)$. *Then there exists a unique fixed point* $u^\varepsilon \in L^\infty(\Omega_{\bar{t}}) \cap L^\infty(0,\bar{t};L^2(\mathbb{R}^d))$ *of the mapping* $L_\varepsilon : L^\infty(\Omega_{\bar{t}}) \to L^\infty(\Omega_{\bar{t}})$ *which is defined for* $v \in L^\infty(\Omega_{\bar{t}})$ *and* $\varepsilon > 0$ *by*

(3.10)

$$\begin{aligned}(L_\varepsilon v)(\mathbf{x},t) \;\;=\;\; & \big[K_\varepsilon(.,t) * u_0\big](\mathbf{x}) + \int_0^t \sum_{j=1}^{d} \big[\partial_{x_j} K_\varepsilon(.,t-s) * f_j(v(.,s))\big](\mathbf{x})\,ds \\ & - \int_0^t \sum_{j=1}^{d} \big[\partial_{x_j} K_\varepsilon(.,t-s) * q^\varepsilon[v(.,s)]\big](\mathbf{x})\,ds \quad ((\mathbf{x},t) \in \Omega_{\bar{t}}).\end{aligned}$$

For the function u^ε *there is a constant* $C_0 > 0$ *such that we have*

$$\|u^\varepsilon\|_{L^\infty(\Omega_{\bar{t}})} + \|u^\varepsilon\|_{L^\infty(0,\bar{t};L^2(\mathbb{R}^d))} \;\;\leq\;\; C_0. \tag{3.11}$$

The number C_0 *depends on* $r_\infty, \varepsilon,\, f_1,\ldots,f_d, \phi$ *and* $\|u_0\|_{L^2(\mathbb{R}^d)}$.

Proof. At first we show that under condition (3.8) for $\bar{t}$ the inclusion $L_\varepsilon(K(\bar{r})) \subset K(\bar{r})$ holds. Here we denote $K(\bar{r}) = \{v \in L^\infty(\Omega_{\bar{t}}) \,|\, \|v\|_{L^\infty(\Omega_{\bar{t}})} \leq \bar{r}\}$ and $\bar{r}$ is the number from (3.7). Using the definition of L_ε and the properties of the kernel in (3.6) we deduce for $t \in [0,\bar{t}\,]$

$$\begin{aligned}&\|(L_\varepsilon v)(.,t)\|_{L^\infty(\mathbb{R}^d)} \leq \|u_0\|_{L^\infty(\mathbb{R}^d)} \\ &\quad + (L_{\mathbf{f}}(\bar{r}) + 2\gamma d)\|v\|_{L^\infty(\Omega_{\bar{t}})} \int_0^t \int_{\mathbb{R}^d} \sum_{j=1}^{d} \big|\partial_{x_j} K_\varepsilon(\mathbf{x},t-s)\big|\,d\mathbf{x}ds.\end{aligned}$$

We have used Assumption 3.1 and the estimate (3.3) and proceed with the estimate on $\partial_{x_j} K_\varepsilon$ in (3.6) and observe for $t \in [0,\bar{t}\,]$

$$\|(L_\varepsilon v)(.,t)\|_{L^\infty(\mathbb{R}^d)} \leq r_\infty \;+ 2C_\varepsilon(L_{\mathbf{f}}(\bar{r}) + 2\gamma d)\|v\|_{L^\infty(\Omega_{\bar{t}})}\sqrt{t}. \tag{3.12}$$

From (3.8) we get $L_\varepsilon(K(\bar r)) \subset K(\bar r)$ since $\bar r > r_\infty$. With the same arguments we obtain for $v_1, v_2 \in L^\infty(\Omega_{\bar t})$ and $t \in [0,\bar t\,]$

$$\|(L_\varepsilon v_1)(.,t) - (L_\varepsilon v_2)(.,t)\|_{L^\infty(\mathbb{R}^d)} \le 2C_\varepsilon(L_{\mathbf{f}}(\bar r) + 2\gamma d)\sqrt{t}\|v_1 - v_2\|_{L^\infty(\Omega_{\bar t})}. \tag{3.13}$$

Since $\bar t$ satisfies (3.9) we see that L_ε is a contraction in $L^\infty(\Omega_{\bar t})$ and has a unique fixed point u^ε in this space. With the arguments above one obtains the similar estimates in $L^\infty(0,\bar t; L^2(\mathbb{R}^d))$:

(3.14)
$$\begin{aligned} \|(L_\varepsilon v)(.,t)\|_{L^2(\mathbb{R}^d)} &\le \|u_0\|_{L^2(\mathbb{R}^d)} + 2C_\varepsilon(L_{\mathbf{f}}(\bar r) + 2\gamma d)\sqrt{t}\|v\|_{L^\infty(0,\bar t;L^2(\mathbb{R}^d))}, \\ \|(L_\varepsilon v_1)(.,t) - (L_\varepsilon v_2)(.,t)\|_{L^2(\mathbb{R}^d)} &\le 2C_\varepsilon(L_{\mathbf{f}}(\bar r) + 2\gamma d)\sqrt{t}\|v_1 - v_2\|_{L^\infty(0,\bar t;L^2(\mathbb{R}^d))}. \end{aligned}$$

Consider now the sequence $\{u^\varepsilon_m\}_{m\in\mathbb{N}}$ defined by

$$u^\varepsilon_{m+1}(\mathbf{x},t) := (L_\varepsilon u^\varepsilon_m)(\mathbf{x},t), \quad u^\varepsilon_1(\mathbf{x},t) = u_0(\mathbf{x}) \quad ((\mathbf{x},t) \in \Omega_{\bar t}). \tag{3.15}$$

This sequence is a Cauchy sequence in the Banach space $L^\infty(0,\bar t; L^2(\mathbb{R}^d))$ by (3.14). A subsequence must converge also in $L^\infty(0,\bar t; L^\infty(\mathbb{R}^d))$ by (3.13). The limit is the unique fixed point u^ε of L_ε. The estimate (3.11) follows from (3.12) and (3.14) with $v = u^\varepsilon$ and taking into account $2C_\varepsilon(L_{\mathbf{f}}(\bar r) + 2\gamma d)\sqrt{\bar t} < 1$. □

The fixed point u^ε of L_ε is a weak solution of (3.1) that satisfies (3.2). The implicit representation formula (3.10) shows $u^\varepsilon \in C^2_1(\Omega_{\bar t})$ (with $u^\varepsilon(.,0) := u_0$, $u^\varepsilon(.,\bar t) := \lim_{s\to\bar t, s<t} u^\varepsilon(.,s)$). For $p = 2,\infty$ we observe $u^\varepsilon \in C([0,\bar t\,]; L^p(\mathbb{R}^d))$ by means of (3.10), the L^p-bound for u^ε and the continuity of convolutions with the heat kernel. Thus u^ε is a classical solution of (3.1), (3.2). We now estimate the Sobolev norms of derivatives of u^ε. Due to the regularizing behavior of the heat kernel we obtain estimates on the Sobolev norm of each order that do not depend on the regularity of the initial function.

Lemma 3.4 (estimates for derivatives). *Suppose that Assumption* 3.1 *holds and let $u_0 \in K_0(r_\infty) \cap L^2(\mathbb{R}^d)$ with r_∞ from* (3.7) *be given. Then the fixed point u^ε of L_ε satisfies $u^\varepsilon \in L^\infty(0,\bar t; W^{k,2}(\mathbb{R}^d) \cap W^{k,\infty}(\mathbb{R}^d))$ with $k \in \mathbb{N}$ given by Assumption* 3.1*. Moreover for all $t^* \in (0,\bar t\,]$ there is a continuous function $\mathcal{C}^\alpha : [0,\infty) \to [0,\infty)$ such that for all $\alpha \in \mathbb{N}^d_0$, $|\alpha| \in \{1,\dots,k\}$ and $t \in (t^*,\bar t\,]$ we have*

(3.16)
$$\|(t-t^*)^{\frac12} D^\alpha u^\varepsilon(.,t)\|_{L^p(\mathbb{R}^d)} \le \mathcal{C}^\alpha\Big(\|u_0\|_{L^\infty(\mathbb{R}^d)} + \|u_0\|_{L^2(\mathbb{R}^d)} + \frac{1}{\sqrt{t^*}}\Big) \quad (p = 2,\infty).$$

The function $\mathcal{C}$ depends in particular on k and $\bar t$ but not on u_0.

Proof. For $m \in \mathbb{N}$, we define the iterates u^ε_m as in (3.15). Let $t^* \in (0,\bar t\,]$. For a continuous function $\mathcal{C}^\alpha : [0,\infty) \to [0,\infty)$ independent of m we shall prove below the following uniform estimate:

(3.17)
$$\|(t-t^*)^{\frac12} D^\alpha u^\varepsilon_m(.,t)\|_{L^p(\mathbb{R}^d)} \le \mathcal{C}^\alpha\Big(\|u_0\|_{L^\infty(\mathbb{R}^d)} + \|u_0\|_{L^2(\mathbb{R}^d)} + \frac{1}{\sqrt{t^*}}\Big) \quad (p = 2,\infty).$$

Here $\alpha \in \mathbb{N}^d_0$ is a multi-index with $|\alpha| \le k$. From Lemma 3.3 we have $u^\varepsilon_m \to u^\varepsilon$ in $L^\infty(\Omega_{\bar t})$. By (3.17) there is then for each $\alpha \in \mathbb{N}^d_0$, $|\alpha| \le k$, a function $v^\varepsilon_\alpha \in L^\infty(\mathbb{R}^d \times (t^*,\bar t))$ such that (for a subsequence) $D^\alpha u^\varepsilon_m \overset{*}{\rightharpoonup} v^\varepsilon_\alpha$. We conclude

$u^\varepsilon \in L^\infty(t^*,\bar t;W^{k,\infty}(\mathbb{R}^d))$ and $D^\alpha u^\varepsilon = v_\alpha$. Since t^* is arbitrary we get $u^\varepsilon \in L^\infty(0,\bar t;W^{k,\infty}(\mathbb{R}^d))$. A similar argument applies to show $u^\varepsilon \in L^\infty(0,\bar t;W^{k,2}(\mathbb{R}^d))$. The estimate (3.16) follows from (3.17) by taking the limit $m\to\infty$. It remains to prove (3.17). For $v \in W^{2,k}(\mathbb{R}^d)\cap W^{\infty,k}(\mathbb{R}^d)$ and $\beta\in\mathbb{N}_0^d$, $|\beta|\in\{1,\ldots,k\}$ we deduce from $f_j\in C^k(\mathbb{R}^d)$ that there is a continuous monotone increasing function $\mathcal{A}^\beta:[0,\infty)\to[0,\infty)$ independent of v such that

(3.18)
$$\sum_{j=1}^d \|D^\beta(f_j(v)) - f_j'(v)D^\beta v\|_{L^p(\mathbb{R}^d)} \le \begin{cases} \mathcal{A}^\beta\big(\|v\|_{W^{2,|\beta|-1}(\mathbb{R}^d)} + \|v\|_{W^{\infty,|\beta|-1}(\mathbb{R}^d)}\big) : p=2, \\ \mathcal{A}^\beta\big(\|v\|_{W^{\infty,|\beta|-1}(\mathbb{R}^d)}\big) \qquad : p=\infty. \end{cases}$$

Note that $D^\beta f_j(v) - f_j'(v)D^\beta v$ contains only derivatives of v up to order $|\beta|-1$ but there are derivatives of the fluxes up to order $|\beta|$.

We return to the proof of (3.17). The case $|\alpha|=0$ has been proven in Lemma 3.3. We start with $|\alpha|=1$ and obtain from the iteration formula (3.10) with (3.6) for $i=1,\ldots,d$, $p=2,\infty$, and $t\in(0,\bar t\,]$,

$$\begin{aligned} &\|\partial_{x_i} u^\varepsilon_{m+1}(.,t)\|_{L^p(\mathbb{R}^d)} \\ &\le \|\partial_{x_i} K_\varepsilon(.,t)*u_0\|_{L^p(\mathbb{R}^d)} + \sum_{j=1}^d \int_0^t \|\partial_{x_j} K_\varepsilon(.,t-s)*\partial_{x_i}(f_j(u^\varepsilon_m(.,s)))\|_{L^p(\mathbb{R}^d)}\,ds \\ &\quad + \sum_{j=1}^d \int_0^t \|\partial_{x_j} K_\varepsilon(.,t-s)*\partial_{x_i}(q^\varepsilon[u^\varepsilon_m(.,s)])\|_{L^p(\mathbb{R}^d)}\,ds \\ &\le \frac{C_\varepsilon}{\sqrt t}\|u_0\|_{L^p(\mathbb{R}^d)} + 2C_\varepsilon(L_{\mathbf f}(\bar r)+2\gamma d)\sqrt t\|\partial_{x_i} u^\varepsilon_m\|_{L^p(0,\bar t;L^p(\mathbb{R}^d))}. \end{aligned}$$

Multiplication with $(t-t^*)^{1/2}$ implies (3.16) for $|\alpha|=1$. We proceed by induction with respect to α and choose $p=\infty$ first. For $t^*\in(0,\bar t\,]$, $l\in\{1,\ldots,k\}$, $i\in\{1,\ldots,d\}$, $\alpha = le_i$, e_i the ith unit vector and $t\in(t^*,\bar t\,]$ we can write

$$\begin{aligned} &\|D^\alpha u^\varepsilon_{m+1}(.,t)\|_{L^\infty(\mathbb{R}^d)} \\ &\le \|\partial_{x_i} K_\varepsilon(.,t-t^*)*D^{\alpha-e_i}u^\varepsilon_{m+1}(.,t^*)\|_{L^\infty(\mathbb{R}^d)} \\ &\quad + \int_{t^*}^t \sum_{j=1}^d \|\partial_{x_j} K_\varepsilon(.,t-s)*D^\alpha f_j(u^\varepsilon_m(.,s))\|_{L^\infty(\mathbb{R}^d)}\,ds \\ &\quad + \int_{t^*}^t \sum_{j=1}^d \|\partial_{x_j} K_\varepsilon(.,t-s)*D^\alpha q^\varepsilon[u^\varepsilon_m(.,s)])\|_{L^\infty(\mathbb{R}^d)}\,ds \\ &\le \frac{C_\varepsilon}{\sqrt{t-t^*}}\|D^{\alpha-e_i}u^\varepsilon_{m+1}(.,t^*)\|_{L^\infty(\mathbb{R}^d)} \\ &\quad + 2C_\varepsilon\sqrt{t-t^*}\left(\sum_{j=1}^d |f_j'(u^\varepsilon_m)| + 2\gamma d\right)\|D^\alpha u^\varepsilon_m\|_{L^\infty(\Omega_{t^*,\bar t})} \\ &\quad + 2C_\varepsilon\sqrt{t-t^*}\sum_{j=1}^d \|D^\alpha(f_j(u^\varepsilon_m)) - f_j'(u^\varepsilon_m)D^\alpha u^\varepsilon_m\|_{L^\infty(\Omega_{t^*,\bar t})}. \end{aligned}$$

Here we used $\Omega_{t^*,\bar{t}} := \mathbb{R}^d \times (t^*, \bar{t}]$. The estimate (3.18) implies

$$
\begin{aligned}
&\|D^\alpha u^\varepsilon_{m+1}(.,t)\|_{L^\infty(\mathbb{R}^d)} \\
&\leq \frac{C_\varepsilon}{\sqrt{t-t^*}} \|D^{\alpha-e_i} u^\varepsilon_{m+1}(.,t^*)\|_{L^\infty(\mathbb{R}^d)} + 2C_\varepsilon \sqrt{t-t^*}(L_{\mathbf{f}}(\bar{r}) + 2\gamma d)\|D^\alpha u^\varepsilon_m\|_{L^\infty(\Omega_{t^*,\bar{t}})} \\
&+ 2C_\varepsilon \sqrt{t-t^*} \mathcal{A}^\alpha\big(\|u^\varepsilon_m\|_{L^\infty(t^*,\bar{t};W^{\infty,|\alpha|-1}(\mathbb{R}^d))}\big).
\end{aligned} \tag{3.19}
$$

The first and third term on the right-hand side of (3.19) contain only derivatives of lower order than $|\alpha|$. Therefore we can apply the induction hypothesis (3.17). The first term is evaluated in t^* directly. The induction hypothesis holds for arbitrary $t^* \in (0, \bar{t}]$ and if we apply it for, e.g., $t^*/2$ at the position of t^* in (3.17) we see that there is a continuous function $\mathcal{B}^\alpha_1 : [0,\infty) \to [0,\infty)$ (depending on $\mathcal{C}^{\alpha-e_i}$ in (3.17)) such that

$$
\|D^{\alpha-e_i} u^\varepsilon_{m+1}(.,t^*)\|_{L^\infty(\mathbb{R}^d)} \leq \mathcal{B}^\alpha_1\Big(\|u_0\|_{L^\infty(\mathbb{R}^d)} + \|u_0\|_{L^2(\mathbb{R}^d)} + \frac{1}{\sqrt{t^*}}\Big)
$$

holds. For the third term we obtain directly from the hypothesis that there is a continuous function $\mathcal{B}^\alpha_2 : [0,\infty) \to [0,\infty)$ (depending on $\mathcal{A}^\alpha$ in (3.19) and $\mathcal{C}^{\alpha-e_i}$ in (3.17)) such that

$$
\mathcal{A}^\alpha\big(\|u^\varepsilon_m\|_{L^\infty(t^*,\bar{t};W^{\infty,|\alpha|-1}(\mathbb{R}^d))}\big) \leq \mathcal{B}^\alpha_2\Big(\|u_0\|_{L^\infty(\mathbb{R}^d)} + \|u_0\|_{L^2(\mathbb{R}^d)} + \frac{1}{\sqrt{t^*}}\Big)
$$

holds. Altogether with $2C_\varepsilon\sqrt{t-t^*}(L_{\mathbf{f}}(\bar{r}) + 2\gamma d) < 1$ we see that there is a continuous function $\mathcal{C}^\alpha : [0,\infty) \to [0,\infty)$ (depending on $\mathcal{B}^\alpha_1$, $\mathcal{B}^\alpha_2$) such that

$$
\|D^\alpha u^\varepsilon_m(.,t)\|_{L^\infty(\mathbb{R}^d)} \leq \frac{1}{\sqrt{t-t^*}} \mathcal{C}^\alpha\Big(\|u_0\|_{L^\infty(\mathbb{R}^d)} + \frac{1}{\sqrt{t^*}}\Big).
$$

Mixed derivatives are treated in the same manner. Also the case $p = 2$ can be treated analogously. ∎

3.4. Proof of Theorem 3.2. In this section we extend the local existence results from Lemmas 3.3 and 3.4 to time intervals $[0,T]$ for the arbitrary but fixed number $T > 0$. The key tool is the following a priori estimate.

LEMMA 3.5 (a priori estimate). *Let the assumptions of Theorem* 3.2 *be valid. For* $\varepsilon > 0$ *and* $T > 0$ *assume that* $u^\varepsilon \in C^2_1(\Omega_T)$ *is a classical solution of* (3.1), (3.2) *that satisfies for* $t \in (0,T]$ *the decay estimates*

$$
\lim_{|\mathbf{x}|\to\infty} |D^\alpha u^\varepsilon(\mathbf{x},t)| = 0 \quad (\alpha \in \mathbb{N}^d_0, |\alpha| \leq 1). \tag{3.20}
$$

Then u^ε *satisfies for* $t \in [0,T]$

$$
\frac{1}{2}\|u^\varepsilon(.,t)\|^2_{L^2(\mathbb{R}^d)} + \varepsilon \sum_{j=1}^d \|\partial_{x_j} u^\varepsilon\|^2_{L^2(\Omega_t)} = \frac{1}{2}\|u_0\|^2_{L^2(\mathbb{R}^d)}. \tag{3.21}
$$

Proof. Let $w \in C^1(\mathbb{R}^d)$ be a function that tends to 0 for $|\mathbf{x}| \to \infty$ and $j \in \{1, \ldots, d\}$. Since ϕ_ε is even we observe

$$\begin{aligned}\int_{\mathbb{R}^d} w(\mathbf{x})\big[[\phi_\varepsilon * w](\mathbf{x})\big]_{x_j}\,d\mathbf{x} &= -\int_{\mathbb{R}^d} w(\mathbf{y})\int_{\mathbb{R}^d}\phi_\varepsilon(\mathbf{y}-\mathbf{x})w_{x_j}(\mathbf{x})\,d\mathbf{x}d\mathbf{y}\\ &= -\int_{\mathbb{R}^d} w(\mathbf{y})[\phi_\varepsilon * w_{x_j}](\mathbf{y})\,d\mathbf{y}\\ &= -\int_{\mathbb{R}^d} w(\mathbf{y})\big[[\phi_\varepsilon * w](\mathbf{y})\big]_{y_j}\,d\mathbf{y}.\end{aligned} \tag{3.22}$$

Interchanging the rôles of $\mathbf{x}$ and $\mathbf{y}$ in the last term we obtain

$$\int_{\mathbb{R}^d} w(\mathbf{x})\big[[\phi_\varepsilon * w](\mathbf{x})\big]_{x_j}\,d\mathbf{x} = 0 \quad (j = 1, \ldots, d). \tag{3.23}$$

To prove (3.21) we multiply (3.1) by u^ε, integrate with respect to space and time, and obtain by (3.2), (3.20) the equality

$$\begin{aligned}&\frac{1}{2}\|u^\varepsilon(.,t)\|^2_{L^2(\mathbb{R}^d)} + \varepsilon\sum_{j=1}^d \|\partial_{x_j}u^\varepsilon\|^2_{L^2(\Omega_T)}\\ =\;& \frac{1}{2}\|u_0\|^2_{L^2(\mathbb{R}^d)} + \gamma\sum_{j=1}^d\int_0^t\int_{\mathbb{R}^d} u^\varepsilon(\mathbf{x},s)\Big[\big[\phi_\varepsilon * u^\varepsilon(.,s)\big](\mathbf{x}) - u^\varepsilon(\mathbf{x},s)\Big]_{x_j}\,d\mathbf{x}ds.\end{aligned}$$

Using (3.23) leads to the statement (3.21). ◻

For the proof of Theorem 3.2 we rely on the L^2-a priori estimate in Lemma 3.5 and Sobolev's inequalities.

Proof of Theorem 3.2. Let u^ε be the unique fixed point of L_ε which has been determined in Lemma 3.3. From Lemma 3.4 we have $u^\varepsilon(.,s) \in W^{k,\infty}(\mathbb{R}^d)\cap W^{k,2}(\mathbb{R}^d)$ for $s \in (0, \bar{t}\,]$. Since $k > d/2$ we deduce by Sobolev's inequality that there is a constant $C > 0$ such that

$$\begin{aligned}\|u^\varepsilon(.,s)\|_{L^\infty(\mathbb{R}^d)} &\le C\|u^\varepsilon(.,s)\|_{W^{k,2}(\mathbb{R}^d)}\\ &\le C\|u^\varepsilon(.,s)\|^{1-\theta}_{L^2(\mathbb{R}^d)}\left(\sum_{|\alpha|=k}\|D^\alpha u^\varepsilon(.,s)\|_{L^2(\mathbb{R}^d)}\right)^\theta.\end{aligned}$$

The last estimate follows from the interpolation theory for Sobolev spaces with $\theta \in (0,1)$ (cf. [6]). Moreover the regularity of u^ε implies that (3.20) holds in $(0, \bar{t}\,]$. Thus the estimate (3.21) applies and we obtain with Lemma 3.4 for $s \in (\bar{t}/4, \bar{t})$ and $t^* = \bar{t}/4$ the estimate

$$\|u^\varepsilon(.,s)\|_{L^\infty(\mathbb{R}^d)} \le C\|u_0\|^{1-\theta}_{L^2(\mathbb{R}^d)}\mathcal{C}\Big(\|u_0\|_{L^\infty(\mathbb{R}^d)} + \|u_0\|_{L^2(\mathbb{R}^d)} + \frac{2}{\sqrt{\bar{t}}}\Big)^\theta (s - \bar{t}/4)^{-\frac{\theta}{2}}.$$

The function $\mathcal{C}$ is a continuous nonnegative function depending on the functions $\mathcal{C}^\alpha$, $|\alpha| = k$, in (3.16). Now we fix s, say $s = \bar{t}/2$, and choose r_2 so small such that for all u_0 with $\|u_0\|_{L^2(\mathbb{R}^d)} \le r_2$ the right-hand side in the last inequality is bounded from above by $\|u_0\|_{L^\infty(\mathbb{R}^d)}$. For the (new) initial function $u_0 = u^\varepsilon(.,s)$ with $s = \bar{t}/2$ the assumptions of Lemma 3.3 are satisfied. We can extend our local solution to the time interval $[0, 3\bar{t}/2)$. Repeating all arguments we finally obtain a global unique solution u^ε for $t \in [0,T]$. The function u^ε satisfies $u^\varepsilon \in C_1^2(\Omega_T)$. Since we have (after extension of u_m^ε to Ω_T) $u_m^\varepsilon \in C([0,T]; L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d))$ by construction and $u_m^\varepsilon \to u^\varepsilon$ in $L^\infty(0,T; L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d))$ from the proof of Lemma 3.3 we have also $u^\varepsilon \in C([0,T]; L^2(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d))$. Thus u^ε provides a classical solution of (3.1), (3.2). It is unique in this class since each classical solution is a fixed point of (3.10), which is unique. □

4. Sharp-interface limits for the scalar model problem in one dimension. We let $d = 1$. For $x = x_1$ and $f(u) = f_1(u)$ (1.2) with (1.6) takes the form

$$u_t^\varepsilon + f(u^\varepsilon)_x = \varepsilon u_{xx}^\varepsilon + \gamma \big[\phi_\varepsilon * u^\varepsilon - u^\varepsilon\big]_x \tag{4.1}$$

in $\mathbb{R} \times (0,T)$, $T > 0$. As before we consider the Cauchy problem for (4.1) with initial condition

$$u^\varepsilon(x,0) = u_0(x) \quad (x \in \mathbb{R}). \tag{4.2}$$

The assumptions on u_0 will be specified below. We restrict ourselves to the choice

$$f(u) = u^3.$$

In this section we assume that a unique classical solution of the Cauchy problem for (4.1) with certain properties exists. Actually this can be proven using Theorem 3.2. For a short discussion of this issue and restrictions concerning f, we refer to Note 4.6. We study the sharp-interface limit $\varepsilon \to 0$. Section 4.1 is devoted to the derivation of certain a priori estimates. Using the compensated compactness framework in section 4.2 we treat the limit problem in one space dimension with $\gamma = \mathcal{O}(1)$.

4.1. A priori estimates. For $\varepsilon, \gamma > 0$ we introduce the mapping $E^\varepsilon : L^2(\mathbb{R}) \to L^1(\mathbb{R})$ by

$$E^\varepsilon[w](x) = \frac{\gamma}{4} \int_{\mathbb{R}} \phi_\varepsilon(x-y)\big(w(y) - w(x)\big)^2 \, dy \quad (w \in L^2(\mathbb{R}),\ x \in \mathbb{R}). \tag{4.3}$$

We assume that Assumption 3.1 for $d = 1$ is valid throughout the section. We need the following technical lemma.

LEMMA 4.1. *Let $w \in C_1^2(\mathbb{R} \times (0,T))$ such that we have for all $t \in (0,T)$*

$$w(.,t),\ w_t(.,t),\ w_x(.,t),\ w_{xx}(.,t) \in L^2(\mathbb{R}).$$

Then we have for $t \in (0,T)$

$$\frac{d}{dt} \int_{\mathbb{R}} E^\varepsilon[w(.,t)](x) \, dx = -\gamma \int_{\mathbb{R}} \Big[\big[\phi_\varepsilon * w(.,t)\big](x) - w(x,t)\Big] w_t(x,t) \, dx \tag{4.4}$$

and

$$2 \int_{\mathbb{R}} E^\varepsilon[w_x(.,t)](x) \, dx = \gamma \int_{\mathbb{R}} \Big[\big[\phi_\varepsilon * w(.,t)\big](x) - w(x,t)\Big] w_{xx}(x,t) \, dx. \tag{4.5}$$

Proof. We compute for $t \in (0,T)$

$$\begin{aligned}
&\frac{\gamma}{4}\frac{d}{dt}\int_{\mathbb{R}}\int_{\mathbb{R}}\phi_\varepsilon(x-y)(w(y,t)-w(x,t))^2\,dydx\\
&= \frac{\gamma}{2}\int_{\mathbb{R}}\int_{\mathbb{R}}\phi_\varepsilon(x-y)(w(y,t)-w(x,t))w_t(y,t)\,dydx\\
&\quad +\frac{\gamma}{2}\int_{\mathbb{R}}\int_{\mathbb{R}}\phi_\varepsilon(y-x)(w(x,t)-w(y,t))w_t(x,t)\,dydx\\
&= \gamma\int_{\mathbb{R}}\int_{\mathbb{R}}\phi_\varepsilon(x-y)(w(y,t)-w(x,t))w_t(y,t)\,dydx\\
&= -\gamma\int_{\mathbb{R}}\big([\phi_\varepsilon * w(.,t)](x)-w(x,t)\big)w_t(x,t)\,dx.
\end{aligned}$$

This is (4.4). Note that we used the symmetry of ϕ. To derive (4.5) consider for $t \in (0,T)$

$$\begin{aligned}
\int_{\mathbb{R}}\int_{\mathbb{R}} E^\varepsilon[w_x(.,t)](x)\,dx &= -\frac{\gamma}{2}\int_{\mathbb{R}}\int_{\mathbb{R}}\phi_\varepsilon(x-y)(w_x(x,t)w_x(y,t)-w_x^2(x,t))\,dydx\\
&= -\frac{\gamma}{2}\int_{\mathbb{R}} w_x(x,t)\big([\phi_\varepsilon * w_x(.,t)](x)-w_x(x,t)\big)\,dx\\
&= \frac{\gamma}{2}\int_{\mathbb{R}} w_{xx}(x,t)\big([\phi_\varepsilon * w(.,t)]-w(x,t)\big)\,dx. \qquad \square
\end{aligned}$$

With this lemma we now prove the crucial a priori estimates on u^ε.

LEMMA 4.2. *Assume that we have $u_0 \in L^2(\mathbb{R})\cap L^\infty(\mathbb{R})$. Let $u^\varepsilon \in C([0,T];L^2(\mathbb{R})\cap L^\infty(\mathbb{R}))\cap C_1^2(\mathbb{R}\times(0,T))$ be a classical solution of (4.1), (4.2) that satisfies for $t \in (0,T]$ the relations*

$$u_t^\varepsilon(.,t),\ u_x^\varepsilon(.,t),\ u_{xx}^\varepsilon(.,t) \in L^2(\mathbb{R}). \tag{4.6}$$

Then we have for all $t \in [0,T]$

$$\frac{1}{2}\|u^\varepsilon(.,t)\|^2_{L^2(\mathbb{R})} + \varepsilon\|u_x^\varepsilon\|^2_{L^2(\mathbb{R}\times(0,t))} = \frac{1}{2}\|u_0\|^2_{L^2(\mathbb{R})} \tag{4.7}$$

and

$$\begin{aligned}
&\frac{1}{4}\|u^\varepsilon(.,t)\|^4_{L^4(\mathbb{R})}\\
&\quad+\int_{\mathbb{R}} E^\varepsilon[u^\varepsilon(.,t)](x)\,dx + 3\varepsilon\|u^\varepsilon u_x^\varepsilon\|^2_{L^2(\mathbb{R}\times(0,t))}\\
&\quad+2\varepsilon\gamma\|E^\varepsilon[u_x^\varepsilon]\|_{L^1(\mathbb{R}\times(0,t))} = \frac{1}{4}\|u_0\|^4_{L^4(\mathbb{R})} + \int_{\mathbb{R}} E^\varepsilon[u_0](x)\,dx.
\end{aligned} \tag{4.8}$$

Proof. The proof of (4.7) has already been given in Lemma 3.5. We turn to prove the second estimate (4.8). By (4.6) and Sobolev embedding we get for $t \in (0,T]$ the limit

$$\lim_{|x|\to\infty}\big(|u^\varepsilon(x,t)| + |u_x^\varepsilon(x,t)|\big) = 0. \tag{4.9}$$

Using (4.9) we obtain by multiplication of (4.1) with $f(u) = u^3$

$$\begin{aligned}
&\frac{1}{4}\|u^\varepsilon(.,t)\|^4_{L^4(\mathbb{R})} + 3\varepsilon\|u^\varepsilon u^\varepsilon_x\|^2_{L^2(\mathbb{R}\times(0,t))} \\
&\quad = \frac{1}{4}\|u_0\|^4_{L^4(\mathbb{R})} - \gamma\int_0^t\int_{\mathbb{R}} [(u^\varepsilon(x,s))^3]_x\big[[\phi_\varepsilon * u^\varepsilon(.,s)](x) - u^\varepsilon(x,s)\big]\,dxds.
\end{aligned} \tag{4.10}$$

Moreover, multiplication of (4.1) with the term $-\gamma[\phi_\varepsilon * u^\varepsilon(.,t) - u^\varepsilon]$ leads to

$$\begin{aligned}
&-\gamma\int_0^t\int_{\mathbb{R}} u^\varepsilon_t(x,s)\big[[\phi_\varepsilon * u^\varepsilon(.,s)](x) - u^\varepsilon(x,s)\big]\,dxds \\
&-\gamma\int_0^t\int_{\mathbb{R}} [(u^\varepsilon(x,s))^3]_x\big[[\phi_\varepsilon * u^\varepsilon(.,s)](x) - u^\varepsilon(x,s)\big]\,dxds \\
&\quad = -\varepsilon\gamma\int_0^t\int_{\mathbb{R}} u^\varepsilon_{xx}(.,s)\big[[\phi_\varepsilon * u^\varepsilon(.,s)](x) - u^\varepsilon(x,s)\big]\,dxds.
\end{aligned}$$

We now apply both identities from Lemma 4.1 with $w = u^\varepsilon$. Then the last equation simplifies to

$$\begin{aligned}
&\int_0^t \frac{d}{ds}\left(\int_{\mathbb{R}} E^\varepsilon[u^\varepsilon(.,s)](x)dx\right)ds - \gamma\int_0^t\int_{\mathbb{R}} (u^\varepsilon(x,s))^3_x[[\phi_\varepsilon * u^\varepsilon(.,s)](x) - u^\varepsilon(x,s)]dxds \\
&\quad = -2\varepsilon\gamma\int_0^t\int_{\mathbb{R}} E^\varepsilon[u^\varepsilon_x(.,s)](x)dxds.
\end{aligned} \tag{4.11}$$

Adding (4.10) and (4.11) implies the estimate (4.8). ∎

It is interesting to compare estimate (4.8) with analogous estimates that can be obtained for the local counterpart of (4.1) where the convolution-type term $[\phi_\varepsilon * u^\varepsilon(.,s)](x) - u^\varepsilon(x,s)$ is substituted by the dispersive term $\varepsilon^2 u^\varepsilon_{xxx}$ (cf. (1.3)). In the latter case one gets ([19])

$$\begin{aligned}
&\frac{1}{4}\|u^\varepsilon(.,t)\|^4_{L^4(\mathbb{R})} + \frac{\gamma\varepsilon^2}{2}\int_{\mathbb{R}}(u^\varepsilon_x(x,t))^2\,dx + 3\varepsilon\|u^\varepsilon u^\varepsilon_x\|^2_{L^2(\mathbb{R}\times(0,t))} + \varepsilon^3\gamma\|(u^\varepsilon_{xx})^2\|_{L^1(\mathbb{R}\times(0,t))} \\
&\quad = \frac{1}{4}\|u_0\|^4_{L^4(\mathbb{R})} + \frac{\gamma\varepsilon^2}{2}\int_{\mathbb{R}}(u_{0,x}(x))^2\,dx.
\end{aligned}$$

The term $\frac{\gamma\varepsilon^2}{2}(u^\varepsilon_x)^2$ takes the rôle of E^ε. In the realistic models for phase transitions like, e.g., the Navier–Stokes–Korteweg system these two expressions are exactly the additional terms in the free energy functional that model the contributions of surface tension ([31]).

Note moreover that the energy estimate for the nonlocal dispersion does not involve the spatial derivative $u_{0,x}$. It is a major advantage of nonlocal regularizations that they come along with less restrictive assumptions on the regularity of the problem.

In the limit process one main difficulty is to deal with mixed terms which include local and nonlocal terms. To cope with this problem we shall use the following lemma.

Lemma 4.3. *Let the assumptions of Lemma* 4.2 *be valid. Then there exists a constant $C > 0$ such that*

$$\|\phi_\varepsilon * u^\varepsilon - u^\varepsilon\|_{L^2(\mathbb{R}\times(0,T))} \le C\sqrt{\varepsilon}.$$

The constant C depends on T, u_0 and ϕ but not on ε.

Proof. Let $(x,t) \in \mathbb{R} \times (0,T)$ be arbitrary but fixed. Denote $B_\varepsilon(x) = \{y \in \mathbb{R} \,|\, |x-y| \leq \varepsilon\}$. We consider $I : \mathbb{R} \times (0,T) \to \mathbb{R}$ with

$$I(x,t) = [\phi_\varepsilon * u^\varepsilon(.,t)](x) - u^\varepsilon(x,t) = \int_{\mathbb{R}} \phi_\varepsilon(x-y)(u^\varepsilon(x,t) - u^\varepsilon(y,t))\, dy.$$

Assumption 3.1(ii) and the Morrey-type inequality (see [13, section 5.6.2])

$$|w(x) - w(y)| \leq C_1 \sqrt{\varepsilon} \Big(\int_{x-2\varepsilon}^{x+2\varepsilon} |w_x(z)|^2 \, dz \Big)^{1/2} \quad (x \in \mathbb{R},\, y \in B_\varepsilon(x),\, w \in C^1(\mathbb{R}))$$

show that the following estimate holds:

$$\begin{aligned}
|I(x,t)| &\leq \int_{B_\varepsilon(x)} \phi_\varepsilon(x-y) |(u^\varepsilon(x,t) - u^\varepsilon(y,t))| \, dy \\
&\leq C_1 \sqrt{\varepsilon} \int_{B_\varepsilon(x)} \phi_\varepsilon(x-y) \left(\int_{x-2\varepsilon}^{x+2\varepsilon} |u_x^\varepsilon(z,t)|^2 \, dz \right)^{1/2} dy \\
&= C_1 \sqrt{\varepsilon} \left(\int_{x-2\varepsilon}^{x+2\varepsilon} |u_x^\varepsilon(z,t)|^2 \, dz \right)^{1/2}.
\end{aligned}$$

Now we integrate $|I(x,t)|^2$ with respect to space and obtain with the substitution $z = x + \frac{4\varepsilon}{\pi} \arctan(\tilde{z})$,

$$\begin{aligned}
\int_{\mathbb{R}} |I(x,t)|^2 \, dx &= C_1 \varepsilon \int_{\mathbb{R}} \left(\int_{x-2\varepsilon}^{x+2\varepsilon} |u_x^\varepsilon(z,t)|^2 \, dz \right) dx, \\
&= C_1 \varepsilon \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \Big| u_x^\varepsilon \Big(x + \frac{4\varepsilon}{\pi} \arctan(\tilde{z}), t \Big) \Big|^2 \, dx \right) \frac{4\varepsilon}{\pi} \frac{1}{1+\tilde{z}^2} \, d\tilde{z}, \\
&= C_1 \frac{4\varepsilon^2}{\pi} \int_{\mathbb{R}} \left(\int_{\mathbb{R}} |u_x^\varepsilon(x,t)|^2 \, dx \right) \frac{1}{1+\tilde{z}^2} \, d\tilde{z}, \\
&\leq C_2 \varepsilon^2 \|u_x^\varepsilon(.,t)\|^2_{L^2(\mathbb{R})}.
\end{aligned}$$

Integration with respect to time and estimate (4.7) of Lemma 4.2 yield

$$\|I\|_{L^2(\mathbb{R} \times (0,T))} \leq C_3 \varepsilon \|u_x^\varepsilon\|_{L^2(\mathbb{R} \times (0,T))} \leq C\sqrt{\varepsilon}.$$

This is the statement of the lemma. ∎

4.2. The limit process of vanishing diffusion and dispersion. We now prove that the classical solutions of the regularized problem (4.1), (4.2) converge in a strong norm for vanishing ε to a limit function u which is a weak solution of

$$u_t(x,t) + f(u(x,t))_x = 0, \quad f(u) = u^3. \tag{4.12}$$

We need a more general notation of an entropy pair as in section 2. The pair $(\eta, q) \in C^2(\mathbb{R})$ is called an entropy pair for (4.12) if and only if $\eta \in C^2(\mathbb{R})$ and the consistency relation $\eta'(w) f'(w) = q'(w)$ holds for all $w \in \mathbb{R}$. Convexity of η is not claimed in this section. To perform the limit $\varepsilon \to 0$ for the problem (4.1), (4.2) we make use of the following theorem which is an adaption of the original work of Schonbek and Murat [29, 32].

Theorem 4.4 (Schonbek/Murat). *We suppose that the family* $\{u^\varepsilon\}_{\varepsilon>0}$ *of classical solutions of* (4.1), (4.2) *is uniformly bounded in* $L^p(\mathbb{R}\times(0,T))$ *for some* $p>1$ *and satisfies*

$$\frac{\partial}{\partial t}\eta(u^\varepsilon)+\frac{\partial}{\partial x}q(u^\varepsilon)\subset \text{ compact set in } W^{-1,2}(Q) + \text{ bounded set in } \mathcal{M}(Q) \tag{4.13}$$

for all entropy pairs (η,q) *for* (4.12) *such that there is a constant* $C>0$ *with*

$$|\eta'(w)|+|\eta''(w)|\le C \quad (w\in\mathbb{R}) \tag{4.14}$$

and all open bounded sets $Q\subset\mathbb{R}\times(0,T)$.
Then the following statements are valid.

(i) *There is a subsequence* $\{u^{\varepsilon_k}\}_{k\in\mathbb{N}}$ *of* $\{u^\varepsilon\}_{\varepsilon>0}$ *and a function* $u\in L^p(\mathbb{R}\times(0,T))$ *such that* $\varepsilon_k\to 0$ *for* $k\to\infty$ *and the subsequence converges for* $k\to\infty$ *to* u *in* $L^r_{loc}(\mathbb{R}\times(0,T))$, $r\in[1,p)$.

(ii) *If moreover* $p>3$ *and Assumption* 3.1(ii) *with* $d=1$ *holds, we have that* u *is a weak solution of* (4.12), *i.e.,*

$$\int_0^T\int_{\mathbb{R}} u(x,t)\psi_t(x,t)+f(u(x,t))\psi_x(x,t)\,dxdt=0 \tag{4.15}$$

for all $\psi\in C_0^\infty(\mathbb{R}\times(0,T))$.

Proof. The first statement (i) is the content of Theorem 3.2 and Corollary 3.2 in [32]. We note that the origin of the sequence $\{u^\varepsilon\}_{\varepsilon>0}$ as a solution of an initial-value problem with nonlocal regularization is not important in order to apply these results but only the compactness property (4.13) related to the flux f via the entropy pairs. In this form the compactness property has been introduced in [29]. To prove (ii) we state that the converging subsequence $\{u^{\varepsilon_k}\}_{k\in\mathbb{N}}$ is a family of classical solutions of (4.1), (4.2) and satisfies in particular for all $\psi\in C_0^\infty(\mathbb{R}\times(0,T))$,

$$\begin{aligned}
&\int_0^T\int_{\mathbb{R}} u^{\varepsilon_k}(x,t)\psi_t(x,t)+f(u^{\varepsilon_k}(x,t))\psi_x(x,t)\,dxdt\\
&\quad= -\int_0^T\int_{\mathbb{R}}\varepsilon_k u^{\varepsilon_k}(x,t)\psi_{xx}(x,t)\,dxdt\\
&\qquad +\gamma\int_0^T\int_{\mathbb{R}}\big([\phi_{\varepsilon_k}*u^{\varepsilon_k}(.,t)](x)-u^{\varepsilon_k}(x,t)\big)\psi_x(x,t)\,dxdt.
\end{aligned} \tag{4.16}$$

From (i), $p>3$, and f being the cubic function, we deduce that the left-hand side in (4.16) converges to the expression on the left-hand side in (4.15) if k tends to ∞. It remains to be seen whether the terms on the right-hand side in (4.16) converge to 0. This is clear for the viscosity term. For the capillarity term we have

$$\begin{aligned}
&\left|\gamma\int_0^T\int_{\mathbb{R}}\big([\phi_{\varepsilon_k}*u^{\varepsilon_k}(.,t)](x)-u^{\varepsilon_k}(x,t)\big)\psi_x(x,t)\,dxdt\right|\\
&\quad= \left|\gamma\int_0^T\int_{\mathbb{R}}u^{\varepsilon_k}(x,t)\big([\phi_{\varepsilon_k}*\psi_x(.,t)](x)-\psi_x(x,t)\big)\,dxdt\right|\\
&\quad\le \gamma C\|u^{\varepsilon_k}\|_{L^p(\Omega_T)}\|\phi_{\varepsilon_k}*\psi_x-\psi_x\|_{L^2(\Omega_T)}\\
&\quad\to 0 \quad (k\to\infty).
\end{aligned}$$

The last line follows from the pointwise convergence of the convolution of a smooth function towards the function if the convolution parameter vanishes. Note moreover that the L^p-norm of $\{u^{\varepsilon_k}\}_{k\in\mathbb{N}}$ is uniformly bounded and that $p > 2$ holds; C is an interpolation constant that depends on the support of ψ_x and its convolution. Statement (ii) is now proven. $\square$

We now have all the tools to prove the main theorem of this section.

THEOREM 4.5. *Suppose that we have $u_0 \in L^2(\mathbb{R}) \cap L^\infty(\mathbb{R})$ and that the function $\phi \in C^\infty(\mathbb{R}^d)$ is even, bounded, nonnegative, and satisfies*

$$\int_{\mathbb{R}} \phi(x)\,dx = 1, \quad \operatorname{supp}(\phi) \subset [-1,1].$$

Let $u^\varepsilon \in C([0,T]; L^\infty(\mathbb{R})\cap L^2(\mathbb{R})) \cap C_1^2(\mathbb{R}\times(0,T))$ be a classical solution of (4.1), (4.2) *that satisfies the decay estimate* (4.6) *for $t \in (0,T]$. Then there exists a subsequence $\{u^{\varepsilon_k}\}_{k\in\mathbb{N}}$ of $\{u^\varepsilon\}_{\varepsilon>0}$ and a function $u \in L^p(\mathbb{R}\times(0,T))$, $p \in [2,4]$, such that*

(i) *the subsequence converges to u in $L^r_{loc}(\mathbb{R}\times(0,T))$, $r \in [1,4)$,*

(ii) *u is a weak solution of* (4.12).

Proof. From Lemma 4.2 we know that $\{u^\varepsilon\}_{\varepsilon>0}$ is bounded uniformly in $L^p(\mathbb{R}\times(0,T))$ for each $p \in [2,4]$. Let (η,q) be an arbitrary but fixed entropy pair with the property (4.14). We obtain

$$\begin{aligned} \eta(u^\varepsilon)_t + q(u^\varepsilon)_x &= \varepsilon\eta(u^\varepsilon)_{xx} - \varepsilon\eta''(u^\varepsilon)(u^\varepsilon_x)^2 \\ &\quad + \gamma[\eta'(u^\varepsilon)(\phi_\varepsilon * u^\varepsilon - u^\varepsilon)]_x - \gamma\eta''(u^\varepsilon)u^\varepsilon_x(\phi_\varepsilon * u^\varepsilon - u^\varepsilon) \\ &=: I_1^\varepsilon + I_2^\varepsilon + I_3^\varepsilon + I_4^\varepsilon. \end{aligned}$$

We will now prove that we have for each bounded open subset Q of $\mathbb{R}\times(0,T)$ the relations

$$|\langle I_1^\varepsilon,\theta\rangle| \overset{\varepsilon\to 0}{\longrightarrow} 0 \quad \forall\,\theta \in W_0^{1,2}(Q), \tag{4.17}$$

$$|\langle I_2^\varepsilon,\psi\rangle| \;\le\; C_2\|\psi\|_{L^\infty(Q)} \quad \forall\,\psi \in C_0^\infty(Q), \tag{4.18}$$

$$|\langle I_3^\varepsilon,\theta\rangle| \overset{\varepsilon\to 0}{\longrightarrow} 0 \quad \forall\,\theta \in W_0^{1,2}(Q), \tag{4.19}$$

$$|\langle I_4^\varepsilon,\psi\rangle| \;\le\; C_4\|\psi\|_{L^\infty(Q)} \quad \forall\,\psi \in C_0^\infty(Q). \tag{4.20}$$

The lemma of Murat ([29]) implies then that assumption (4.13) of Theorem 4.4 can be satisfied for $p \in (1,4]$. Thus the statement of Theorem 4.5 is proven. We start with I_1^ε and observe with (4.14), Hölder's inequality, and Lemma 4.2 that

$$\begin{aligned} |\langle I_1^\varepsilon,\theta\rangle| &= \left|\int_Q \varepsilon\eta(u^\varepsilon(x,t))_{xx}\theta(x,t)\,dxdt\right| \\ &\le \varepsilon\|\eta'(u^\varepsilon)u^\varepsilon_x\|_{L^2(Q)}\|\theta_x\|_{L^2(Q)} \\ &\le \varepsilon C_1\|u^\varepsilon_x\|_{L^2(Q)}\|\theta\|_{W^{1,2}(Q)} \\ &\le \sqrt{\varepsilon}C_1\|\theta\|_{W^{1,2}(Q)} \overset{\varepsilon\to 0}{\longrightarrow} 0. \end{aligned}$$

This is (4.17). To check (4.18) we use again Lemma 4.2 and consider the estimate

$$
\begin{aligned}
|\langle I_2^\varepsilon,\psi\rangle| &= \left|\int_Q \varepsilon\eta''(u^\varepsilon(x,t))(u_x^\varepsilon(x,t))^2\psi(x,t)\,dxdt\right| \\
&\le \varepsilon C_2\|u_x^\varepsilon\|^2_{L^2(Q)}\|\psi\|_{L^\infty(Q)} \\
&\le C_2\|\psi\|_{L^\infty(Q)}.
\end{aligned}
$$

Statement (4.19) is a consequence of

$$
\begin{aligned}
|\langle I_3^\varepsilon,\theta\rangle| &= \left|\gamma\int_Q \Big[\eta'(u^\varepsilon(x,t))[\phi_\varepsilon * u^\varepsilon(.,t)](x) - u^\varepsilon(x,t)\Big]_x\theta(x,t)\,dxdt\right| \\
&\le C_3\gamma\|\phi_\varepsilon * u^\varepsilon - u^\varepsilon\|_{L^2(Q)}\|\theta_x\|_{L^2(Q)} \\
&\le C_3\gamma\sqrt{\varepsilon}\|\theta\|_{W^{1,2}(Q)} \overset{\varepsilon\to 0}{\longrightarrow} 0.
\end{aligned}
$$

For the last inequality we used Lemma 4.3. Finally we have for I_ε^4

$$
\begin{aligned}
|\langle I_4^\varepsilon,\psi\rangle| &= \left|\gamma\int_Q \eta''(u^\varepsilon(x,t))u_x^\varepsilon(x,t)([\phi_\varepsilon * u^\varepsilon(.,t)](x) - u^\varepsilon(x,t))\psi(x,t)\,dxdt\right| \\
&\le C_4\gamma\|u_x^\varepsilon\|_{L^2(Q)}\|\phi_\varepsilon * u^\varepsilon - u^\varepsilon\|_{L^2(Q)}\|\psi\|_{L^\infty(Q)} \\
&\le C_4\gamma\|\psi\|_{L^\infty(Q)}.
\end{aligned}
$$

For the last inequality we used Lemmas 4.2 and 4.3. □

NOTE 4.6.

(i) *In Theorem* 4.5 *we supposed that a classical solution* $u^\varepsilon \in C([0,T];L^\infty(\mathbb{R})\cap L^2(\mathbb{R}))\cap C_1^2(\mathbb{R}\times(0,T))$ *of* (4.1), (4.2) *exists and satisfies the decay estimate* (4.6) *for* $t\in(0,T]$. *In view of Lemma* 3.4, *in particular the estimate* (3.16), *we can guarantee all these properties if Theorem* 3.2 *applies, i.e., if* $\|u_0\|_{L^2(\mathbb{R})}$ *is small enough.*

(ii) *Theorem* 4.5 *is tailored to the choice* $f(u) = u^3$. *To apply the* L^p*-theory of Schonbek a growth condition on the flux* f *is necessary. However, the restriction to cubic functions can be avoided. One can apply similar techniques as in* [23] *to treat the case of arbitrary fluxes which are only assumed to have a globally bounded derivative.*

5. Sharp-interface limit in the case of multiple space dimensions. In this section we return to multiple space dimensions and consider the Cauchy problem for (3.1), (3.2). We do not consider a special choice for the flux functions $f_1,\dots,f_d$ as in section 4.2 but we require them to satisfy Assumption 3.1 and for some $p>1$ to be specified below the following growth condition:

(5.1) $\quad \exists\, C>0: \quad |f_j(u)| \le C(1+u^r)$ for all $u\in\mathbb{R}$, $r\in[0,p)$, and $j=1,\dots,d$.

Throughout the section we assume that (3.1), (3.2) has a unique classical solution u^ε in $\Omega_T := \mathbb{R}^d\times(0,T)$, $T>0$. This is true, e.g., under the conditions of Theorem 3.2. We study the sharp-interface limit $\varepsilon\to 0$. A similar study for the Cauchy problem with the local diffusion and dispersion term has been performed in [23].

5.1. Review of the theory of -measure-valued solutions. In the limit $\varepsilon \to 0$ we will deal with Kruzkov solutions for the first order conservation law

$$u_t + \operatorname{div}(\mathbf{f}(u)) = 0 \quad \text{in } \Omega_T. \tag{5.2}$$

A **Kruzkov solution of** (5.2)

$$u(.,0) = u_0 \text{ in } \mathbb{R}^d.$$

Equation [24] is a function $u \in L^p_{loc}(\Omega_T)$ such that for all $k \in \mathbb{R}$ and all $\psi \in C_0^\infty(\Omega_T)$, $\psi \geq 0$,

$$\int_{\Omega_T} |u(\mathbf{x},t) - k|\psi_t(\mathbf{x},t) + \operatorname{sgn}(u(\mathbf{x},t) - k)\sum_{j=1}^{d}(f_j(u(\mathbf{x},t)) - f_j(k))\psi_{x_j}(\mathbf{x},t)\, d\mathbf{x}dt \geq 0$$

holds and moreover the initial condition is satisfied in the sense that for all compact subsets K of $\mathbb{R}^d$,

$$\lim_{T\to 0^+} \frac{1}{T}\int_0^T \int_K |u(\mathbf{x},t) - u_0(\mathbf{x})|\, d\mathbf{x}dt = 0 \tag{5.3}$$

holds.

By an **entropy solution of** (5.2)

$$u(.,0) = u_0 \text{ in } \mathbb{R}^d,$$

we mean in this section function $u \in L^p_{loc}(\Omega_T)$ such that for all smooth functions $\eta, q_1, \dots, q_d : \mathbb{R} \to \mathbb{R}$ with η' bounded and $\eta' f_j' = q_j'$, $j = 1, \dots, d$ and all test functions $\psi \in C_0^\infty(\Omega_T)$, $\psi \geq 0$, we have

$$\int_{\Omega_T} \eta(\mathbf{x},t)\psi_t(\mathbf{x},t) + \sum_{j=1}^{d} q_j(u(\mathbf{x},t)\psi_{x_j}(\mathbf{x},t)\, d\mathbf{x}dt \geq 0$$

and also (5.3). Let us mention that Kruzkov solutions are entropy solutions and vice versa [16] but we need both notions in what follows. Our analysis relies on results by Schonbek and Szepessy [32, 35] which are extensions of the concept of measure-valued solutions initiated by DiPerna [10] for L^∞-functions to the case of L^p-functions.

We need some preparatory notations. Let $p \in (1,\infty)$ be arbitrary but fixed. Denote by Prob($\mathbb{R}$) the set of probability measures on $\mathbb{R}$. A mapping $\nu = \nu_{(\mathbf{x},t)}$ from Ω_T to Prob($\mathbb{R}$) is called a **p-Young measure** if the function

$$(\mathbf{x},t) \mapsto \langle \nu_{(\mathbf{x},t)}, g(\lambda)\rangle := \int_{\mathbb{R}} g(\lambda)\, d\nu_{(\mathbf{x},t)}(\lambda)$$

is in $L^p(\Omega_T)$ for all $g \in C^0(\mathbb{R})$ with

$$g(\lambda) = \mathcal{O}(1 + |\lambda|^r) \quad (r \in [0,p)). \tag{5.4}$$

A p-Young measure is called a **measure-valued Kruzkov solution of** (5.2)

$$u(.,0) = u_0 \text{ in } \mathbb{R}^d,$$

if it satisfies for all $k \in \mathbb{R}$ and $\psi \in C_0^\infty(\Omega_T)$, $\psi \geq 0$, the inequality

(5.5)
$$\int_{\Omega_T} \langle \nu_{(\mathbf{x},t)}, |\lambda - k| \rangle \psi_t(\mathbf{x},t) + \sum_{j=1}^{d} \langle \nu_{(\mathbf{x},t)}, \operatorname{sgn}(\lambda - k)(f_j(\lambda) - f_j(k)) \rangle \psi_{x_j}(\mathbf{x},t)\, d\mathbf{x}dt \geq 0$$

and moreover

$$\lim_{T \to 0^+} \frac{1}{T} \int_0^T \int_K \langle \nu_{(\mathbf{x},t)}, |\lambda - u_0(\mathbf{x})| \rangle \, d\mathbf{x}dt = 0 \tag{5.6}$$

for all compact subsets K of $\mathbb{R}^d$. We recover the definition of the Kruzkov solution if we set $\nu_{(\mathbf{x},t)} = \delta_{u(\mathbf{x},t)}$, where $\delta_{u(\mathbf{x},t)}$ is the Dirac-measure concentrated in $u(\mathbf{x},t)$ for almost all $(\mathbf{x},t) \in \Omega_T$.

Now we collect in a theorem results that have been proven in [32, 35].

THEOREM 5.1 (Schonbek, Szepessy). *For $p > 1$ let $\{u^\varepsilon\}_{\varepsilon > 0} \subset L^p(\Omega_T)$ be a sequence of functions that is uniformly bounded in $L^p(\Omega_T)$. Then there exists a p-Young measure $\nu = \nu_{(\mathbf{x},t)}$ such that*

$$g(u^\varepsilon) \rightharpoonup \langle \nu_{(\mathbf{x},t)}, g(\lambda) \rangle \ \textit{in}\ L^p(\Omega) \tag{5.7}$$

holds for all continuous functions g satisfying (5.4)*.*

Let $u_0 \in L^p(\mathbb{R}^d)$. If $\{u^\varepsilon\}_{\varepsilon>0}$ is uniformly bounded in $L^\infty(0,T;L^p(\mathbb{R}^d))$ and ν provides a measure-valued Kruzkov solution of (1.1) *the following statements are valid:*

(i) *There is a function $u \in L^p(\Omega_T)$ such that $\lim_{\varepsilon\to 0} \|u - u^\varepsilon\|_{L^p_{loc}(\Omega_T)} = 0$.*

(ii) *The function u is a Kruzkov solution of* (5.2)*,* (5.3)*.*

5.2. The main result. We conclude with the main result of the section: the convergence of the sequence $\{u^\varepsilon\}_{\varepsilon>0}$ of classical solutions of (3.1), (3.2) to a Kruzkov solution of the corresponding first order problem (5.2), (5.3). We relax the scaling of the nonlocal term with respect to ε and suppose that the parameter $\gamma > 0$ satisfies

$$\gamma = \gamma(\varepsilon) = o(\sqrt{\varepsilon}). \tag{5.8}$$

THEOREM 5.2. *Let Assumption* 3.1 *and condition* (5.1) *with $p = 2$ for $f_1, \ldots, f_d$ be satisfied. For $u_0 \in L^2(\Omega_T)$ assume that $\{u^\varepsilon\}_{\varepsilon>0}$ is a sequence of classical solutions for* (3.1)*,* (3.2)*. If we choose γ according to* (5.8) *then there is a Kruzkov solution u of the Cauchy problem* (5.2)*,* (3.2) *such that*

$$\lim_{\varepsilon \to 0} \|u - u^\varepsilon\|_{L^2_{loc}(\Omega_T)} = 0.$$

Since u in Theorem 5.2 is a Kruzkov solution it is not only an entropy solution in the sense of this section but also in the sense of the definition in section 2 for $d = 1$. Then we conjecture that the existence of undercompressive waves in the limit is excluded.

However, the result of Theorem 5.2 shows that the new nonlocal regularization leads also in multiple space dimensions to a well-defined limit behavior at least in the regime governed by (5.8).

Proof of Theorem 5.2. From Lemma 3.5 we know that the sequence $\{u^\varepsilon\}_{\varepsilon>0}$ is in particular uniformly bounded in $L^\infty(0,T;L^2(\mathbb{R}^d)) \subset L^2(\Omega_T)$. Thus choose $p = 2$

to apply Theorem 5.1. The first part of that theorem defines the 2-Young measure $\nu = \nu_{(\mathbf{x},t)}$ associated with $\{u^\varepsilon\}_{\varepsilon>0}$. We show first that ν satisfies (5.5). It suffices to show that

$$\int_{\Omega_T} \langle \nu_{(\mathbf{x},t)}\eta(\lambda)\rangle \psi_t(\mathbf{x},t) + \sum_{j=1}^{d} \langle \nu_{(\mathbf{x},t)}, q_j(\lambda)\rangle \psi_{x_j}(\mathbf{x},t)\, d\mathbf{x}dt \geq 0 \tag{5.9}$$

holds for all nonnegative test functions $\psi \in C_0^\infty(\Omega_T)$ and all smooth functions $\eta, q_1, \ldots, q_d$ such that η' is a bounded function and $\eta' f_j' = q_j'$, $j = 1, \ldots, d$ is satisfied. An approximation argument for the less regular Kruzkov entropies $|\cdot - k|$ yields that ν satisfies (5.9). We obtain for arbitrary nonnegative test function $\psi \in C_0^\infty(\Omega_T)$ after multiplication of (3.1) with $\eta'(u^\varepsilon)$ and integration with respect to $(\mathbf{x}, t) \in \Omega_T$,

(5.10)

$$\begin{aligned}
&\int_{\Omega_T} \eta(u^\varepsilon(\mathbf{x},t))\psi_t(\mathbf{x},t) + \sum_{j=1}^{d} q_j(u^\varepsilon(\mathbf{x},t))\psi_{x_j}(\mathbf{x},t)\, d\mathbf{x}dt \\
&= \quad -\varepsilon \int_{\Omega_T} \eta(u^\varepsilon(\mathbf{x},t))\Delta\psi(\mathbf{x},t)\, dxdt + \varepsilon \int_{\Omega_T} \eta''(u^\varepsilon(\mathbf{x},t))|\nabla u^\varepsilon(\mathbf{x},t)|^2 \psi(\mathbf{x},t)\, d\mathbf{x}dt \\
&\quad\quad + \gamma(\varepsilon) \int_{\Omega_T} \eta'(u^\varepsilon(\mathbf{x},t))([\phi_\varepsilon * u^\varepsilon(.,t)](\mathbf{x}) - u^\varepsilon(\mathbf{x},t)) \sum_{j=1}^{d} \psi_{x_j}(\mathbf{x},t)\, d\mathbf{x}dt \\
&\quad\quad + \gamma(\varepsilon) \int_{\Omega_T} \eta''(u^\varepsilon(\mathbf{x},t)) \left(\sum_{j=1}^{d} u^\varepsilon_{x_j}(\mathbf{x},t) \right) ([\phi_\varepsilon * u^\varepsilon(.,t)](\mathbf{x}) - u^\varepsilon(\mathbf{x},t))\psi(\mathbf{x},t)\, d\mathbf{x}dt \\
&= \quad I_1^\varepsilon + I_2^\varepsilon + I_3^\varepsilon + I_4^\varepsilon.
\end{aligned}$$

With Lemma 3.5 and $\|\phi_\varepsilon * u^\varepsilon\|_{L^2(\Omega_T)} \leq \|u^\varepsilon\|_{L^2(\Omega_T)}$ we can estimate for appropriate ε-independent constants $C_1, C_3, C_4 > 0$

$$|I_1^\varepsilon| \leq C_1 \varepsilon \|\eta'(u^\varepsilon)\|_{L^\infty(\Omega_T)} \|\nabla u^\varepsilon\|_{L^2(\Omega_T)} \|\nabla\psi\|_{L^2(\Omega_T)} \leq C_1\sqrt{\varepsilon} \tag{5.11}$$

and

$$|I_3^\varepsilon| \leq C_3\gamma(\varepsilon) \|\eta'(u^\varepsilon)\|_{L^\infty(\Omega_T)} \|[\phi_\varepsilon * u^\varepsilon] - u^\varepsilon\|_{L^2(\Omega_T)} \|\nabla\psi\|_{L^2(\Omega_T)} \leq C_3\gamma(\varepsilon) \tag{5.12}$$

and

(5.13)

$$|I_4^\varepsilon| \leq \gamma(\varepsilon) \|\eta''(u^\varepsilon)\|_{L^\infty(\Omega_T)} \|\nabla u^\varepsilon\|_{L^2(\Omega_T)} \|[\phi_\varepsilon * u^\varepsilon] - u^\varepsilon\|_{L^2(\Omega_T)} \|\psi\|_{L^\infty(\Omega_T)} \leq C_4 o(\varepsilon).$$

If we use (5.11), (5.12), (5.13), and $I_2^\varepsilon \geq 0$ for the limit $\varepsilon \to 0$ in (5.10), the weak convergence statement (5.7) in Theorem 5.1 shows that the Young measure $\nu_{(\mathbf{x},t)}$ satisfies (5.9). Note that $\eta, q_1, \ldots, q_d$ satisfies (5.4) by construction.

It remains to be seen whether the initial condition (5.6) is valid. We follow the argumentation in [35, Theorem 3.2] where it turns out that it is sufficient to prove that there is a constant $C_5 > 0$ such that

$$J = -\frac{1}{T} \int_0^T \int_{\mathbb{R}^d} \int_0^t u_t^\varepsilon(\mathbf{x},s)\rho(\mathbf{x})\, dsd\mathbf{x}dt \leq C_5 T \tag{5.14}$$

holds for all $\rho \in C_0^\infty(\mathbb{R}^d)$. The constant $C_5 > 0$ might depend on ρ but not on ε. Since u^ε is a classical solution of (3.1) we can use the differential equation and estimate

$$\begin{aligned} J &\leq -\frac{1}{T}\int_0^T \int_0^t \int_{\mathbb{R}^d} \sum_{j=1}^d f_j(u^\varepsilon(\mathbf{x},s))\rho_{x_j}(\mathbf{x}) + \varepsilon u^\varepsilon(\mathbf{x},s)\Delta\rho(\mathbf{x})\, d\mathbf{x}dsdt \\ &\quad + \frac{1}{T}\int_0^T \int_0^t \int_{\mathbb{R}^d} \sum_{j=1}^d \gamma(\varepsilon)\big([\phi_\varepsilon * u^\varepsilon(.,s)](\mathbf{x}) - u^\varepsilon(\mathbf{x},s)\big)\rho_{x_j}(\mathbf{x})\, d\mathbf{x}dsdt \\ &\leq C(\mathrm{supp}(\rho),\mathbf{f})\Big((1 + \|u^\varepsilon\|_{L^\infty(0,T;L^2(\mathbb{R}^d))})\|\nabla\rho\|_{L^\infty(\mathbb{R}^d)} \\ &\quad + \varepsilon\|u^\varepsilon\|_{L^\infty(0,T;L^2(\mathbb{R}^d))}\|\Delta\rho\|_{L^\infty(\mathbb{R}^d)} + \gamma(\varepsilon)\|u^\varepsilon\|_{L^\infty(0,T;L^2(\mathbb{R}^d))}\|\nabla\rho\|_{L^\infty(\mathbb{R}^d)}\Big)T. \end{aligned}$$

Note that we used (5.4) to estimate the fluxes in terms of the L^2-norm of $u^\varepsilon(.,t)$ for $t \in [0,T]$ and (3.3). Since $\|u^\varepsilon\|_{L^\infty(0,T;L^2(\mathbb{R}^d))}$ is bounded uniformly with respect to ε due to Lemma 3.5 we have proven (5.14). The second part of Theorem 5.1 concludes the proof. $\square$

REFERENCES

[1] G. Alberti, G. Bellettini, M. Cassandro, and E. Presutti, *Surface tension in Ising systems with Kac potentials*, J. Statist. Phys., 82 (1996), pp. 743–796.

[2] D. Anderson, G. McFadden, and A. Wheeler, *Diffuse-interface methods in fluid mechanics*, Annu. Rev. Fluid Mech., 30 (1998), pp. 139–165.

[3] P. Bates, P. Fife, X. Ren, and X. Wang, *Traveling waves in a convolution model for phase transitions*, Arch. Rational Mech. Anal., 138 (1997), pp. 105–136.

[4] N. Bedjaoui and P. LeFloch, *Diffusive-dispersive traveling waves and kinetic relations. I: Nonconvex hyperbolic conservation laws*, J. Differential Equations, 178 (2002), pp. 574–607.

[5] S. Benzoni-Gavage, *Stability of subsonic planar phase boundaries in a van der Waals fluid*, Arch. Rational Mech. Anal., 150 (1999), pp. 23–55.

[6] J. Bergh and J. Löfström, *Interpolation Spaces: An Introduction*, Springer-Verlag, Berlin, 1976.

[7] A. Bertozzi and M. Shearer, *Existence of undercompressive traveling waves in thin film equations*, SIAM J. Math. Anal., 32 (2000), pp. 194–213.

[8] C. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Fundamental Principles of Mathematical Sciences, Springer-Verlag, 2000.

[9] A. De Masi, T. Gobron, and E. Presutti, *Travelling fronts in non-local evolution equations*, Arch. Rational Mech. Anal., 132 (1995), pp. 143–205.

[10] R. DiPerna, *Measure-valued solutions to conservation laws*, Arch. Rational Mech. Anal., 88 (1985), pp. 223–270.

[11] J. Dunn and J. Serrin, *On the thermomechanics of interstitial working*, Arch. Rational Mech. Anal., 88 (1985), pp. 95–133.

[12] G. Ermentrout and J. McLeod, *Existence and uniqueness of traveling waves for a neural network*, Proc. Roy. Soc. Edinburgh, 123 Sect A (1993), pp. 461–478.

[13] L. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[14] H. Fan and M. Slemrod, *Dynamic flows with liquid/vapor phase transitions*, in Handbook of Mathematical Fluid Dynamics, S. Friedlander and D. Serre, eds., North-Holland, Amsterdam, 2002, pp. 373–420.

[15] R. Fosdick and D. Mason, *On a model of nonlocal continuum mechanics Part I: Existence and regularity*, SIAM J. Appl. Math., 58 (1998), pp. 1278–1306.

[16] E. Godlewski and P.-A. Raviart, *Hyperbolic systems of conservation laws*, Ellipses, Paris, 1991.

[17] R. Hagan and M. Slemrod, *The viscosity-capillarity criterion for shocks and phase transitions*, Arch. Rational Mech. Anal., 83 (1983), pp. 333–361.

[18] H. Hattori and D. Li, *Global solutions of a high-dimensional system for Korteweg materials*, J. Math. Anal. Appl., 198 (1996), pp. 84–97.

[19] B. Hayes and P. LeFloch, *Non-classical shocks and kinetic relations: Scalar conservation laws*, Arch. Rational Mech. Anal., 139 (1997), pp. 1–56.

[20] B. Hayes and M. Shearer, *Undercompressive shocks and Riemann problems for scalar conservation laws with non-convex fluxes*, Proc. Roy. Soc. Edinburgh, Sect. A, Math., 129 (199), pp. 733–754.

[21] D. Jacobs, W. McKinney, and M. Shearer, *Travelling wave solutions of the modified Korteweg-deVries-Burgers equation*, J. Differential Equations, 116 (1995), pp. 448–467.

[22] A. Kluwick, E. Cox, and S. Scheichl, *Non-classical kinematic shocks in suspensions of particles in fluids*, Acta Mech., 144 (2000), pp. 197–210.

[23] C. Kondo and P. LeFloch, *Zero diffusion-dispersion limits for scalar conservation laws*, SIAM J. Math. Anal., 33 (2002), pp. 1320–1329.

[24] S. Kruzkov, *First order quasilinear equations with several independent variables*, Math. USSR-Sb., 123 (1970), pp. 228–255.

[25] J. Lebowitz and O. Penrose, *Rigorous treatment of the van der Waals-Maxwell theory of the liquid-vapor transition*, J. Math. Phys., 7 (1966), pp. 98–113.

[26] P. LeFloch, *An introduction to nonclassical shocks of systems of conservation laws*, in An Introduction to Recent Developments in Theory and Numerics for Conservation Laws. D. Kröner, M. Ohlberger, and C. Rohde eds., Lect. Notes Comput. Sci. Engrg. 5, Springer-Verlag, 1999, pp. 28–72.

[27] P. LeFloch, *Hyperbolic systems of conservation laws. The theory of classical and nonclassical shock waves*, Lect. Math., Birkhauser Verlag, Basel, Switzerland, 2002.

[28] P. LeFloch and C. Rohde, *Existence theory for the Riemann problem with vanishing diffusion and dispersion*, Indiana Univ. Math. J., 50 (2001), pp. 1707–1743.

[29] F. Murat, *L'injection du cône positif de H^{-1} dans $W^{-1,q}$ est compacte pour tout $q < 2$*, J. Math. Pures Appl., 60 (1981), pp. 309–322.

[30] X. Ren and L. Truskinovsky, *Finite scale microstructures in nonlocal elasticity*, J. Elasticity, 59 (2000), pp. 319–355.

[31] C. Rohde, *On local and non-local Navier-Stokes-Korteweg systems for liquid-vapour phase transitions*, Technical report, Math. Institut, Universität Freiburg No. 18, Freiburg, Germany, 2004.

[32] M. Schonbek, *Convergence of solutions to nonlinear dispersive equations*, Comm. Partial Differential Equations, 7 (1982), pp. 959–1000.

[33] D. Serre, *Systems of Conservations Laws, I*, Cambridge University Press, Cambridge, 1999.

[34] M. Slemrod, *Admissibility criteria for propagating phase boundaries in a van der Waals fluid*, Arch. Rational Mech. Anal., 81 (1983), pp. 301–315.

[35] A. Szepessy, *An existence result for scalar conservation laws using measure valued solutions*, Comm. Partial Differential Equations, 14 (1989), pp. 1329–1350.

[36] J. van der Waals, *Thermodynamische Theorie der Kapillarität unter Voraussetzung stetiger Dichteänderung*, Z. Phys. Chem., 13 (1894), pp. 657–725.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 130–155

ANALYSIS OF A THERMOMECHANICAL MODEL FOR SHAPE MEMORY ALLOYS*

ULISSE STEFANELLI†

Abstract. This note addresses the mathematical study of a nonlinear model arising in the description of the macroscopic thermomechanical behavior of shape memory materials and previously introduced in [R. Peyroux, A. Chrysochoos, Ch. Licht, and M. Löbel, *J. Phys. C4 Suppl.,* 6 (1996), pp. 347–356]. In particular, we discuss the model derivation and investigate a system of PDEs coupled with a vectorial variational inequality. We develop the analysis in both the dissipative and the nondissipative cases, providing indeed a quantitative asymptotic connection between the two regimes. Moreover, we prove the global in time well-posedness for suitable initial and boundary value problems. As a by-product of the well-posedness analysis, we address a variable time-step discretization procedure, proving indeed its convergence and providing some a priori error bounds. Finally, we deal with the asymptotic behavior of the system for large times and establish the convergence of the trajectories to the solution of a suitable stationary problem.

Key words. shape memory alloys, well-posedness, discretization, long-time behavior

AMS subject classifications. 74C05, 35K55, 65M12, 35B40

DOI. 10.1137/S0036141004444251

1. Introduction.

The present analysis is concerned with the evolution of four unknown fields θ, χ_1, χ_2, and u governed by the following system of equations and inclusion:

$$\partial_t(c_s\theta - \ell\chi_1) - k\Delta\theta = r, \tag{1.1}$$

$$\operatorname{div}\sigma + b = 0, \tag{1.2}$$

$$\mathbb{A}(\varepsilon(u) + \beta\chi_2) = \sigma, \tag{1.3}$$

$$\mu\partial_t\begin{pmatrix}\chi_1\\ \chi_2\end{pmatrix} + \gamma\begin{pmatrix}\chi_1\\ \chi_2\end{pmatrix} + \begin{pmatrix}\frac{\ell}{\theta_*}(\theta-\theta_*)\\ \sigma:\beta\end{pmatrix} + \partial I_K\begin{pmatrix}\chi_1\\ \chi_2\end{pmatrix} \ni \begin{pmatrix}0\\ 0\end{pmatrix}. \tag{1.4}$$

These relations are asked to be fulfilled in the space-time domain $Q := \Omega\times(0,T)$ for some open and bounded subset $\Omega\subset\mathbb{R}^3$ with smooth boundary Γ and some reference time $T>0$. In addition c_s, ℓ, k, γ, and θ_* are positive parameters, $\mathbb{A}$ and β are, respectively, a 4-tensor and a 2-tensor, and μ is a nonnegative constant (see below). Here, $u=(u_1,u_2,u_3)\in\mathbb{R}^3$ and $\varepsilon(u)$ denotes the 2-tensor

$$\varepsilon_{ij}(u) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \quad \text{for } i,j = 1,2,3,$$

while ∂I_K stands for the subdifferential of the indicator function of the nonempty, bounded, convex, and closed subset K of $\mathbb{R}^2$, i.e., $I_K(x) = 0$ if $x\in K$ and

*Received by the editors June 4, 2004; accepted for publication (in revised form) January 12, 2005; published electronically August 31, 2005. The main part of this research was performed during a visit to France under the partial sponsorship of the MIUR-COFIN 2002 research program on *Free boundary problems in applied sciences.*

http://www.siam.org/journals/sima/37-1/44425.html

†Istituto di Matematica Applicata e Tecnologie Informatiche – CNR, via Ferrata 1, I-27100 Pavia, Italy (ulisse@imati.cnr.it)

$I_K(x) = +\infty$ elsewhere. Namely, $\partial I_K : \mathbb{R}^2 \to 2^{\mathbb{R}^2}$ is defined by

$$y \in \partial I_K(x) \ \text{ iff } \ x \in K \ \text{ and } \ y \cdot (w - x) \leq 0 \quad \forall w \in K.$$

Moreover, $\sigma : \beta := \sigma_{ij}\beta_{ij}$ (summation convention) denotes the standard contraction product of 2-tensors, $(\operatorname{div} \sigma)_i := \partial(\sigma_{ij})/\partial x_j$ for $i = 1, 2, 3$, and $r : Q \to \mathbb{R}$ and $b : Q \to \mathbb{R}^3$ are given functions.

The nonlinear system (1.1)–(1.4) arises in connection with the study of the thermomechanical behavior of shape memory materials. These are metallic alloys with an intrinsic ability of undergoing a thermoelastic solid-solid transformation between crystallographic configurations with different physical and mechanical properties: *austenite*, which is stable at higher temperatures, and variants of *martensite*, stable at lower temperatures [2, 21]. At the macroscopic level such a reversible phase transformation results in the so-called *shape memory effect*. Namely, shape memory alloys can be permanently deformed (up to 8% under traction) and then be forced to recover their original shape just by thermal means. This unusual macroscopic mechanical effect is nowadays exploited in several innovative devices. Indeed shape memory materials are actually used in order to realize a variety of actuators (also of microscopic size) and structures. The field of application of shape memory technologies ranges from bioengineering to structures-engineering and aerospace sciences [13, 16, 30]. Here we are concerned with a macroscopic modelization previously introduced in [31] and able to describe the shape memory effect in a small deformation realm. We refer the reader to section 2 for a derivation of the model as well as for some discussion of its thermodynamical consistency. As for the justification of this modeling perspective as well as some experimental validation one should refer to the original paper [31] as well as to [14, 20, 32]. For the purposes of this introductory discussion, let us remark that the scalar field θ in (1.1)–(1.4) represents the absolute temperature of the medium while the vector field u is its displacement. Hence, the 2-tensors σ and $\varepsilon(u)$ stand for the tension and the linearized strain, respectively. Finally, K is the admissible convex and closed range for the internal variables $[\chi_1, \chi_2]$ (see below).

The current literature on the mathematical modeling of shape memory alloys is quite rich, and it is not our purpose to provide here an exhaustive review. Indeed, let us just mention that the problem of describing the thermomechanical behavior of shape memory alloys has been tackled both from the microscopic [2, 4, 5] and the macroscopic viewpoint [1, 3, 6, 18, 19, 33]. Among the latter, we shall particularly mention the so-called Frémond model for shape memory alloys. This model was originally presented in [19] and analyzed in [9]. Indeed both the present model and Frémond's model [19] are formulated in the framework of generalized standard materials by means of analogous free-energy and dissipation considerations. In particular, the phase relation of Frémond's model corresponds, in the present setting, to the choices $\mu > 0$ and $\gamma = 0$. On the contrary, the modeling considerations and the experimental evidence of [31] suggests considering the degenerate case $\mu = 0$ (taking indeed $\gamma > 0$). We will refer to the situation $\mu > 0$ as the *dissipative case* and $\mu = 0$ as the *nondissipative* one.

From the mathematical point of view, the present situation appears to be more delicate with respect to that of [9] because of the time-degeneracy of the phase relation (1.4). Moreover, let us stress that the momentum balance equation of [9] includes of a fourth order regularizing term that is actually not present in our situation.

The system (1.1)–(1.4) has to be supplied with suitable initial and boundary conditions. To this aim we ask for

$$\theta(\cdot,0) = \theta^0 \quad \text{on} \quad \Omega, \tag{1.5}$$

$$\mu\big[\chi_1(\cdot,0),\chi_2(\cdot,0)\big] = \mu\big[\chi_1^0,\chi_2^0\big] \quad \text{on} \quad \Omega, \tag{1.6}$$

$$k\partial_\nu\theta + h(\theta-\theta_e) = f \quad \text{on} \quad \Gamma, \tag{1.7}$$

$$\sigma\nu = g \quad \text{on} \quad \Gamma_t, \tag{1.8}$$

$$u = 0 \quad \text{on} \quad \Gamma_0, \tag{1.9}$$

where θ^0, χ_1^0, χ_2^0 are initial values, $h, \theta_e > 0$, $f : \Gamma \to \mathbb{R}$, and $g : \Gamma_t \to \mathbb{R}^3$ are prescribed. Moreover, ν is the unit outward normal vector to the boundary, and $\{\Gamma_0, \Gamma_t\}$ is a partition of Γ into two disjoint subsets of positive surface measure.

This paper addresses the mathematical study of the system (1.1)–(1.9) in both the dissipative ($\mu > 0$) and nondissipative ($\mu = 0$) regimes. First of all, we shall comment on the model derivation and its thermodynamical consistency. Then, we investigate the dissipative situation and provide a well-posedness result for a global variational solution (Theorem 4.1). As a by-product of this analysis we provide a variable time-step discretization scheme which turns out to be stable and convergent. Moreover, we are in the position of providing an a priori bound of optimal order on the discretization error. Then, we prove an asymptotic result that connects the dissipative and the nondissipative regimes. In particular, we prove that, as μ goes to zero, the solution of the dissipative model converges to the solution of the nondissipative one (Theorem 4.2). Indeed, we also achieve some estimate in terms of μ on the distance between the latter two solutions. As a corollary of this asymptotic result, one obtains the global variational well-posedness for the nondissipative case as well (Theorem 4.1). Then, we focus on the long-time behavior of solutions for the nondissipative model. In particular, we prove that the model actually converges to a unique equilibrium which is characterized as the solution of a suitable elliptic problem (Theorem 4.3). Finally, we turn to the proof of a suitable maximum principle which entails an essential lower bound for the temperature in terms of data (Theorem 4.4). The latter in particular ensures that, starting from a positive datum, the temperature θ remains positive for all times.

The paper is outlined as follows. We shall discuss the derivation of the model in section 2. Then, we introduce the variational formulation of the problem in section 3. Our main results are stated in section 4, while section 5 is devoted to the study of the dissipative case. In particular, it contains the details of the discretization method. The nondissipative model is investigated in section 6, and section 7 focuses on the long-time behavior of solutions. The crucial proof of the positivity of the temperature is then given in section 8.

2. Model. We devote this section to a derivation of the thermomechanical model in study [31]. Our aim is to possibly clarify the meaning of relations (1.1)–(1.4) and check for the thermodynamical consistency of the model. In particular, it is beyond our purposes to provide the reader with a full justification of this modeling perspective. Indeed, for a comprehensive discussion on the model as well as some experimental validation, the reader should refer to [31], where the model was introduced.

We will describe the thermomechanical evolution of a shape memory material with respect to its smooth reference configuration $\Omega \subset \mathbb{R}^3$ by means of the absolute temperature θ, the (small) deformation $\varepsilon(u)$ (u is the displacement), and a pair of internal variables $[\chi_1, \chi_2]$ introduced below. In particular, for the purposes of this section, θ is *assumed* to be strictly positive (this will turn out to be Theorem 4.4 later on). Let us suppose from the very beginning that only two martensitic variants are present besides one austenite and indicate the respective local proportions as η_1, η_2, and η_A, respectively. This assumption is of course extremely reductive since, in some particular alloy, up to 24 martensitic variants have been detected. Nevertheless our somewhat crude simplification is still suitable for describing the basic features of the physical phenomenon [9, 31]. We, moreover, assume that the phases possibly coexist at each point of the body, that no overlapping between different phases can occur, and that no void appears in the mixture. Hence, the phase proportions η_1, η_2, and η_A are constrained to fulfill the obvious relations

$$0 \leq \eta_1, \eta_2, \eta_A \leq 1, \quad \eta_1 + \eta_2 + \eta_A = 1.$$

We exploit these relations in order to eliminate η_A by introducing the internal variables

$$\chi_1 := \eta_1 + \eta_2, \quad \chi_2 := \eta_1 - \eta_2.$$

Of course the set $\{\chi_1 = 1\}$ corresponds to the situation where no austenite is present, the set $\{\chi_1 = \chi_2\}$ corresponds to the set where just the first variant of martensite is present, etc. Owing to the above discussion it is clear that $[\chi_1, \chi_2]$ are constrained in the triangle

$$K := \{[x_1, x_2] \in \mathbb{R}^2 : 0 \leq |x_2| \leq x_1 \leq 1\}. \tag{2.1}$$

In order to deduce the differential relations governing the evolutions of the state quantities above we will follow the approach via *microscopic motions* originally proposed by Frémond. The basic novelty of this theory is to take into account the thermomechanical effect of the microscopic rearrangements of the phases at the macroscales. In particular, one admits that the microscopic movements of the substance might give rise to some thermal macroscopic effect which influences the overall energy balance of the body. We will not review here the full theory of thermomechanics of continua with microscopic motions and just refer the reader to the recent monograph [20] for both a comprehensive discussion and a specific application to the description of shape memory materials.

To the aim of dealing with microscopic motions, let us postulate from the very beginning that the proper quantities describing such micromovements are $[\dot\chi_1, \dot\chi_2]$, where of course the dot denotes time differentiation (as it is customary, at this stage we assume that all the quantities occurring in the analysis are as smooth as needed in order to go through the differentiations).

Hence, it seems convenient to regard the vector $(u, [\dot\chi, \dot\chi_2])$ as an *actual rigid velocity vector*. Moreover, let us assume that there exists a suitable linear space of virtual rigid body velocities R (see [20] for a full discussion). Finally, we suppose that, for all times $t \in [0, T]$, the virtual power of the internal forces of the body with respect to the generic smooth subdomain $D \subset \Omega$ and virtual rigid body velocities $(v, c) \in R$ is

$$P_{int}(D, v, c) := -\int_D \sigma : \varepsilon(v) - \int_D B \cdot c.$$

The first term above is classical while the second one describes the power of microscopic internal forces. In the latter the quantity $B(\cdot) \in \mathbb{R}^2$ comes into play and an obvious dimensional argument entails that it shall be regarded as an *energy density*.

In particular, B represents and vector energy density per units of $[\dot\chi_1, \dot\chi_2]$ (see [20, sect. 13.3, p. 360]). We now introduce the virtual power of the external and acceleration forces as

$$P_{ext}(D, v, c) := \int_D b \cdot v + \int_{\partial D} g \cdot v \, d\mathcal{H}^{n-1}, \quad P_{acc}(D, v, c) := \int_D \rho \, \zeta \cdot v.$$

Here b represents an action density at distance (body force) while g is an action density at contact (traction) and we use a standard notation for the Hausdorff measure. Moreover, $\zeta = \ddot u$ is the macroscopic acceleration and ρ is the material density (no microscopic accelerations are considered). By recalling the *virtual power principle* [23], choosing arbitrarily the regular and connected domain D and the virtual rigid body velocities $(v, c) \in R$ we deduce from the relation

$$P_{acc}(D, v, c) = P_{int}(D, v, c) + P_{ext}(D, v, c)$$

two systems of momentum balance equations, namely, [20, sect. 2.4, p. 5]

$$\rho \ddot u = \operatorname{div} \sigma + b \quad \text{in } \Omega \times (0, T), \tag{2.2}$$

$$\sigma \nu = g \quad \text{on } \Gamma \times (0, T), \tag{2.3}$$

which stands for the *macroscopic momentum balance*, and

$$B = 0 \quad \text{in } \Omega \times (0, T), \tag{2.4}$$

which corresponds to the *microscopic momentum balance*. Of course, ν stands for the unit normal vector field pointing outward Γ.

Now letting e denote the internal energy density of the system and Q the entropy flux, we can follow [20, sect. 13.4, p. 361] and deduce that, in our situation, the energy balance is expressed by

$$\dot e + \operatorname{div}(\theta Q) - r = \sigma : \varepsilon(\dot u) + B \cdot [\dot\chi_1, \dot\chi_2] \quad \text{in } \Omega \times (0, T), \tag{2.5}$$

$$-\theta Q \cdot \nu = \pi \quad \text{on } \Gamma \times (0, T), \tag{2.6}$$

where r and π denote some volume and surface heat source densities, respectively. In particular, we note that the right-hand side of (2.5) takes into account the contribution to the energy balance provided by both macroscopic and microscopic movements.

The next step is to define the quantities e, Q, σ, and B in terms of the state variables in such a way that the second principle of thermodynamics, in the form of the Clausius–Duhem inequality, is fulfilled. In particular, the latter in our case reduces to

$$\dot s + \operatorname{div} Q - \frac{r}{\theta} \geq 0 \quad \text{in } \Omega \times (0, T), \tag{2.7}$$

where s is the entropy of the system and r/θ represents an external entropy source density. In order to accomplish the above requirement we will exploit the Ginzburg–Landau theory by introducing the *free energy density* $\psi = \psi(\theta, [\chi_1, \chi_2], \varepsilon(u))$ and the *pseudopotential of dissipation* $\phi = \phi(\nabla\theta, [\dot\chi_1, \dot\chi_2])$ and defining

$$s := -\frac{\partial \psi}{\partial \theta}, \quad e := \psi + \theta s, \quad Q := -\frac{\partial \phi}{\partial(\nabla\theta)}, \tag{2.8}$$

$$\sigma := \frac{\partial \psi}{\partial(\varepsilon(u))}, \tag{2.9}$$

$$B := [B_1, B_2], \quad B_j := \frac{\partial \psi}{\partial \chi_j} + \frac{\partial \phi}{\partial (\dot{\chi}_j)}, \quad j = 1, 2. \tag{2.10}$$

The above choice splits B into a *nondissipative* and a *dissipative* part, respectively, and is inspired by thermodynamic considerations (see below and [20]). Moreover, the latter notions of derivative are intended to be properly generalized in case ψ, ϕ are nonsmooth functions. At the present stage, our only requirement on the potentials ψ and ϕ is that [20, 27] ϕ is convex, nonnegative, and vanishes at 0.

We shall now check for the thermodynamic consistency of this class of models by recalling (2.5) and the above definitions in order to compute that

$$\begin{aligned} \dot{s} + \operatorname{div} Q - \frac{r}{\theta} &= \frac{1}{\theta}\left(\theta \dot{s} + \operatorname{div}(\theta Q) - r\right) - \frac{1}{\theta} Q \cdot \nabla \theta \\ &= \frac{1}{\theta}\left(\left[\frac{\partial \phi}{\partial \dot{\chi}_1}, \frac{\partial \phi}{\partial \dot{\chi}_2} \right] \cdot [\dot{\chi}_1, \dot{\chi}_2] - Q \cdot \nabla \theta \right) \geq 0, \end{aligned}$$

where we used the properties of the pseudopotential ϕ. Hence, the Clausius–Duhem inequality (2.7) easily follows from the positivity of θ. As a consequence, the general positivity proof implied by Theorem 4.4 entails the thermodynamic consistency of the whole class of models.

We now come to our actual choice of ψ [31]. In particular, we let

$$\begin{aligned} \psi(\theta, [\chi_1, \chi_2], \varepsilon(u)) &= -c_s \theta \ln \theta + \frac{1}{2}(\varepsilon(u) + \beta \chi_2) : \mathbb{A}(\varepsilon(u) + \beta \chi_2) \\ &\quad + \ell \frac{\theta_* - \theta_{**}}{2\theta_*}(\chi_1^2 + \chi_2^2) + \frac{\ell}{\theta_*}(\theta - \theta_*)\chi_1 + I_K(\chi_1, \chi_2). \end{aligned} \tag{2.11}$$

In the latter expression, the first term is purely caloric and c_s represents a specific heat density. The second term corresponds to the mechanical energy. In particular $\mathbb{A}$ is the elasticity tensor and the extra term $\beta \chi_2$ represents the mechanical effect of the presence of the two different martensitic variants. Indeed, one assumes that the mechanical potential of the material is

$$\frac{1}{2}(\varepsilon(u) + \beta_1 \eta_1 + \beta_2 \eta_2) : \mathbb{A}(\varepsilon(u) + \beta_1 \eta_1 + \beta_2 \eta_2),$$

where β_j are *transformation strain tensors* encoding the mechanical effect of the martensite-austenite phase change and are assumed to verify $\beta_1 = -\beta_2 =: \beta$ in order to take into account the so-called *self-accommodating* properties of the two martensitic variants [20, 31]. Note that the thermal expansion of the system is neglected.

The indicator function I_K forces $[\chi_1, \chi_2]$ to take solely admissible values in K and the term $\frac{\ell}{\theta_*}(\theta - \theta_*)\chi_1$ classically represents the phase–temperature interaction. In particular, ℓ is a latent heat density related to the martensite-austenite transformation and $\theta_* > 0$ is the critical martensite-to-austenite transition temperature.

The modeling novelty of this framework with respect to the original Frémond model for shape memory alloys [9, 19] consists in including into ψ the term

$$\ell \frac{\theta_* - \theta_{**}}{2\theta_*}(\chi_1^2 + \chi_2^2), \tag{2.12}$$

where a second critical transition temperature $0 < \theta_{**} < \theta_*$ is introduced for the austenite-to-martensite transformation. By referring to the zero-stress situation, one observes that in Frémond's model [9] no austenite is present for temperatures θ

below θ_* nor martensites for $\theta > \theta_*$. This simplification is, however, fairly crude and experiments suggest that one should consider a suitable temperature range where the three phases may coexist in the zero-stress situation [14, 32]. The present model [31] extends Frémond's approach in the direction of including some description of this effect. In particular, looking back to (2.4) in the zero-stress equilibrium ($[\dot\chi_1, \dot\chi_2] = 0$, $\sigma = 0$) the internal variables are asked to fulfill

$$\ell \frac{\theta_* - \theta_{**}}{\theta_*} \begin{pmatrix} \chi_1 \\ \chi_2 \end{pmatrix} + \begin{pmatrix} \frac{\ell}{\theta_*}(\theta - \theta_*) \\ 0 \end{pmatrix} + \partial I_K \begin{pmatrix} \chi_1 \\ \chi_2 \end{pmatrix} \ni \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

which entails that

$$\chi_1 = \min \left\{ \max \left\{ \frac{\theta_* - \theta}{\theta_* - \theta_{**}}, 0 \right\}, 1 \right\}, \quad \chi_2 = 0.$$

Hence, no martensites are present for $\theta > \theta_*$, no austenite is allowed for $\theta < \theta_{**}$, and possibly all the phases are admissible for intermediate temperatures [31] (see also [20, Rem. 13.4, p. 364]).

As for the pseudopotential of dissipation we will ask for

$$\phi(\nabla\theta, [\dot\chi_1, \dot\chi_2]) := \frac{k}{2\theta} |\nabla\theta|^2 + \frac{\mu}{2} \dot\chi_1^2 + \frac{\mu}{2} \dot\chi_2^2.$$

Here $k > 0$ stands for a constant thermal conductivity coefficient and $\mu \geq 0$ measures some dissipation effect on the phase variables. In particular, the heat flux $q := \theta Q = -k\nabla\theta$ is of Fourier type.

Finally, the balance relations (2.5), (2.2), and (2.4) read as follows:

$$c_s \dot\theta - \ell \dot\chi_1 - k\Delta\theta - r = -\frac{\ell}{\theta_*}(\theta - \theta_*)\dot\chi_1 + \mu \dot\chi_1^2 + \mu \dot\chi_2^2, \tag{2.13}$$

$$\rho \ddot u = \operatorname{div} \sigma + b, \tag{2.14}$$

$$\mu \begin{pmatrix} \dot\chi_1 \\ \dot\chi_2 \end{pmatrix} + \ell \frac{\theta_* - \theta_{**}}{\theta_*} \begin{pmatrix} \chi_1 \\ \chi_2 \end{pmatrix} + \begin{pmatrix} \frac{\ell}{\theta_*}(\theta - \theta_*) \\ \sigma : \beta \end{pmatrix} + \partial I_K \begin{pmatrix} \chi_1 \\ \chi_2 \end{pmatrix} \ni \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{2.15}$$

We shall focus our attention from the very beginning on the *quasi-static* situation where the inertial term $\rho \ddot u$ is negligible in (2.14). Indeed, let us stress that the latter approximation of the momentum balance equation is rather standard in connection with the Frémond model [9, 10, 12, 24] and translates into the belief that the mechanical evolution takes place on some faster time scale when compared with the thermal evolution. On the other hand, the reader is referred to [8], where the full momentum problem is considered for Frémond's model (see also [15, 22, 25] and the recent monograph [26] for the analysis of mechanical evolution under different nonconvex settings).

In order to deduce the system (1.1)–(1.4) from (2.9) and (2.13)–(2.15) we now apply some further modification to the balance relations by means of suitable small perturbation assumptions. At first, one supposes to be interested in a temperature range close to the critical temperature θ_* and neglects the first term in the right-hand side of (2.13). By setting, for the sake of notational simplicity, $\gamma := \ell(\theta_* - \theta_{**})/\theta_*$, we readily check that the system (1.1)–(1.4) in the nondissipative regime $\mu = 0$ follows directly from (2.9) and (2.13), (2.15), and (2.14) in its quasi-static form.

As for the nondissipative regime $\mu > 0$ we shall additionally assume to be interested in a situation where the phase evolution is suitably slow, i.e., $\mu \dot\chi_j^2 = 0$ in the energy balance equation (2.13). On the other hand we retain the dissipation term $\mu[\dot{\chi}_1, \dot{\chi}_2]$ in (2.15). This is of course again an assumption of small perturbation type.

Finally, as for boundary conditions (1.7)–(1.9), one assumes to know g in (2.3) just on Γ_t and imposes the body to be clamped on Γ_0. Moreover, we choose $\pi := f + h(\theta_e - \theta)$, where f is a prescribed surface heat source density, $h > 0$ is a thermal exchange coefficient, and $\theta_e > 0$ is a given external temperature.

3. Variational formulation. We start by fixing some notation. Let

$$H := L^2(\Omega), \quad \mathcal{H} := H \times H \times H, \quad V := H^1(\Omega),$$
$$\mathbb{H} := \{\sigma : \Omega \to \mathbb{R}^{3\times 3}_{symm} \text{ measurable such that } \sigma : \sigma \in L^1(\Omega)\},$$

where $\mathbb{R}^{3\times 3}_{symm}$ denotes of course the space of 3×3 symmetric tensors. All the above spaces are endowed with their respective natural scalar products. In particular, we will use the symbols $(\cdot, \cdot)$ and $\|\cdot\|$ for all products and norms in the above L^2-type spaces. Moreover, the notation $(\cdot, \cdot)_\Gamma$ will stand for the scalar product in both $L^2(\Gamma)$ and $(L^2(\Gamma))^3$, $|\cdot|$ denotes any Euclidean norm, $\|\cdot\|_E$ will stand for the norm in the generic normed space E, and $[\cdot, \cdot]$ denotes the generic pair. We introduce the Hilbert space

$$\mathcal{V} := \{v \in V^3 \text{ such that } v = 0 \text{ on } \Gamma_0\},$$

endowed with the standard norm, and set, for any $u, v \in \mathcal{V}$,

$$a(u, v) := (\mathbb{A}\varepsilon(u), \varepsilon(v)),$$

where $\varepsilon : \mathcal{V} \to \mathbb{H}$ stands for the linearized strain tensor. Following the classic linear elasticity theory, we ask $\mathbb{A} = (a_{ijkh})$ to be symmetric and positive definite on $\mathbb{R}^{3\times 3}_{symm}$, namely,

$$a_{ijkh} = a_{ijhk} = a_{khij} \quad \forall i, j, h, k = 1, 2, 3 \quad \text{and} \quad \mathbb{A}\sigma : \sigma > 0 \quad \forall \sigma \in \mathbb{R}^{3\times 3}_{symm}/\{0\}.$$

Namely, for all $\sigma, \tau \in \mathbb{R}^{3\times 3}_{symm}$ one has that $\mathbb{A}\sigma : \tau = \sigma : \mathbb{A}\tau = \mathbb{A}^{\frac{1}{2}}\sigma : \mathbb{A}^{\frac{1}{2}}\tau$, where $\mathbb{A}^{\frac{1}{2}}$ stands for the well-defined square root of $\mathbb{A}$. In particular, since $\beta \neq 0$, we readily compute that $\mathbb{A}^{\frac{1}{2}}\beta \neq 0$ as well. Moreover, recalling the Γ_0 has a positive surface measure, and thanks to Korn's inequality (see, e.g., [17, Thm. 3.3, p. 115]), there exists a positive constant $c_\mathcal{V}$ depending on $\mathbb{A}$ such that

$$a(v, v) \ = \ \|\mathbb{A}^{\frac{1}{2}}\varepsilon(v)\|^2 \ \geq \ c_\mathcal{V}\|v\|^2_\mathcal{V} \quad \forall v \in \mathcal{V}.$$

Finally, let the notation $\langle \cdot, \cdot \rangle$ stand for the duality pairing between V' and V or $\mathcal{V}'$ and $\mathcal{V}$, where the prime denotes the topological duals. Since the special triangular form of K specified above is not needed for our analysis, let K be an arbitrary nonempty, bounded, convex, and closed subset of $\mathbb{R}^2$, and define the (convex and closed) set $\mathcal{K} := \big\{[x_1, x_2] \in (L^2(\Omega))^2 \text{ such that } [x_1, x_2] \in K \text{ a.e. in } \Omega\big\}$. For almost every $t \in (0, T)$ let us define the functionals $F(t) : V \to V'$ and $G(t) : \mathcal{V} \to \mathcal{V}'$ as

$$\langle F(t), \varphi \rangle := (r(t), \varphi) + (f(t), \varphi)_\Gamma \quad \forall \varphi \in V,$$
$$\langle G(t), v \rangle := (b(t), v) + (g(t), v)_\Gamma \quad \forall v \in \mathcal{V}.$$

We shall make precise our variational formulation of (1.1)–(1.9) by posing the following problem.

Problem P_μ. Find $\theta \in H^1(0,T;H) \cap L^\infty(0,T;V)$, $[\chi_1, \chi_2] \in (H^1(0,T;H) \cap L^\infty(Q))^2$, and $u \in H^1(0,T;\mathcal{V})$ such that $\mu(\chi_1,\chi_2) \in (W^{1,\infty}(0,T;H))^2$ and

$$\big((c_s\theta - \ell\chi_1)_t, \varphi\big) + k(\nabla\theta, \nabla\varphi) + h(\theta - \theta_e, \varphi)_\Gamma = \langle F, \varphi\rangle \quad \forall \varphi \in V \quad \text{a.e. in } (0,T), \tag{3.1}$$

$$a(u,v) + (\mathbb{A}\beta\chi_2, \varepsilon(v)) = \langle G, v\rangle \quad \forall v \in \mathcal{V} \text{ a.e. in } (0,T), \tag{3.2}$$

$$\mathbb{A}(\varepsilon(u) + \beta\chi_2) = \sigma \quad \text{a.e. in } Q, \tag{3.3}$$

$$\mu\partial_t\begin{pmatrix}\chi_1\\ \chi_2\end{pmatrix} + \gamma\begin{pmatrix}\chi_1\\ \chi_2\end{pmatrix} + \partial I_K\begin{pmatrix}\chi_1\\ \chi_2\end{pmatrix} \ni \begin{pmatrix}\frac{\ell}{\theta_*}(\theta_* - \theta)\\ -\sigma : \beta\end{pmatrix} \quad \text{a.e. in } Q, \tag{3.4}$$

$$c_s\theta(\cdot,0) - \ell\chi_1(\cdot,0) = c_s\theta^0 - \ell\chi_1^0 \quad \text{a.e. in } \Omega, \tag{3.5}$$

$$\mu\big[\chi_1(\cdot,0), \chi_2(\cdot,0)\big] = \mu\big[\chi_1^0, \chi_2^0\big] \quad \text{a.e. in } \Omega. \tag{3.6}$$

Remark 3.1. Let us stress that the above regularity requirements and (3.3) entail, in particular, that $\sigma \in H^1(0,T;\mathbb{H})$. Namely, relation (3.4) makes sense.

4. Main results. We shall assume the following:

(A1) $F \in L^2(0,T;H) + W^{1,1}(0,T;V')$, $G \in H^1(0,T;\mathcal{V}')$,
(A2) $\theta^0 \in V$,
(A3) $[\chi_1^0, \chi_2^0] \in \mathcal{K}$.

In particular, the first formula in (A1) entails that there exist $F_1 \in L^2(0,T;H)$ and $F_2 \in W^{1,1}(0,T;V')$ such that $F = F_1 + F_2$. We observe that whenever $r \in L^2(0,T;H)$, $f \in W^{1,1}(0,T;L^2(\Gamma))$, $b \in H^1(0,T;\mathcal{H})$, and $g \in H^1(0,T;(L^2(\Gamma))^3)$ the regularities in (A1) follow. As a consequence of (A1)–(A3), we introduce $u^0 \in \mathcal{V}$ as the unique solution to (3.2) at time $t=0$, $\sigma^0 \in \mathbb{H}$ via u^0 and (3.3), and finally

$$\mu[\chi_{1,\mu,t}(0), \chi_{2,\mu,t}(0)] := \left[\frac{\ell}{\theta_*}(\theta_* - \theta^0) - \gamma\chi_1^0, -\sigma^0 : \beta - \gamma\chi_2^0\right].$$

In particular, we observe that the left-hand side above is bounded in $H \times H$, uniformly with respect to μ, and that relation (3.4) is fulfilled also for $t=0$.

We are now in a position to state our results.

THEOREM 4.1 (well-posedness). *Let $\mu \geq 0$. Under the assumptions* (A1)–(A3), *there exists a unique solution to Problem P_μ. Moreover, given two sets of data $(F_i, G_i, \theta_i^0, \chi_{1,i}^0, \chi_{2,i}^0)$ fulfilling* (A1)–(A3) *and two external temperatures $\theta_{e,i}$, for $i=1,2$, the respective solutions $(\theta_i, \chi_{1,i}, \chi_{2,i}, u_i)$ to the corresponding problems P_μ fulfill*

$$\begin{aligned}
&\|\theta_1 - \theta_2\|^2_{L^2(0,T;H)} + \sup_{t\in[0,T]}\left\|\int_0^t \nabla(\theta_1 - \theta_2)\right\|^2 + \sup_{t\in[0,T]}\left\|\int_0^t (\theta_1 - \theta_2)\right\|^2_{L^2(\Gamma)} \\
&\quad + \sum_{j=1}^2 \big(\mu\|\chi_{j,1} - \chi_{j,2}\|^2_{C([0,T];H)} + \|\chi_{j,1} - \chi_{j,2}\|^2_{L^2(0,T;H)}\big) + \|u_1 - u_2\|^2_{L^2(0,T;\mathcal{V})} \\
&\leq c_0\left(\int_0^T \left\|\int_0^t (F_{1,1} - F_{1,2})\right\|^2 dt + \|F_{2,1} - F_{2,2}\|^2_{L^1(0,T;V')} + \|G_1 - G_2\|^2_{L^2(0,T;\mathcal{V}')}\right. \\
&\qquad \left. + \|c_s(\theta_1^0 - \theta_2^0) - \ell(\chi_{1,1}^0 - \chi_{1,2}^0)\|^2 + \mu\|\chi_{2,1}^0 - \chi_{2,2}^0\|^2 + |\theta_{e,1} - \theta_{e,2}|^2\right),
\end{aligned} \tag{4.1}$$

where c_0 depends on c_s, k, h, ℓ, γ, Γ, $\mathbb{A}^{\frac12}\beta$, θ_*, and $c_{\mathcal{V}}$ but is independent of μ.

THEOREM 4.2 (dissipation asymptotics). *Under the assumptions* (A1)–(A3), *the solution* $(\theta_\mu, \chi_{1,\mu}, \chi_{2,\mu}, u_\mu)$ *to Problem* P_μ *converges as* $\mu \to 0$ *to the solution* $(\theta, \chi_1, \chi_2, u)$ *of Problem* P_0 *at least weakly in the respective natural spaces.*

Moreover, we will address the study of the long-time behavior of the solution to Problem P_0. Indeed, the reader should notice that the above-stated well-posedness result is actually independent of the choice of the reference time T. Hence, in particular, the solution $(\theta, \chi_1, \chi_2, u)$ to Problem P_0 may be uniquely extended for all times. Now let the ω-limit set be defined as

$$\begin{aligned}\omega\big(\theta, \chi_1, \chi_2, u\big) := \{\big(\theta_\infty, \chi_{1,\infty}, \chi_{2,\infty}, u_\infty\big) \in H \times H \times H \times \mathcal{V} \quad &\text{such that there exists}\\ \text{a sequence of positive real numbers} \quad \{t_n\} \quad \text{with} \quad t_n \longrightarrow +\infty \quad &\text{and}\\ \big(\theta(t_n), \chi_1(t_n), \chi_2(t_n), u(t_n)\big) \longrightarrow \big(\theta_\infty, \chi_{1,\infty}, \chi_{2,\infty}, u_\infty\big) \quad \text{in} \quad &H \times H \times H \times \mathcal{V}\}.\end{aligned}$$

To establish a long-time behavior result, we need some further assumptions on the data. In particular, we ask for

(A4) $F \in L^2(0, +\infty; H)$, $G \in H^1(0, +\infty, \mathcal{V}')$.

Hence, the following holds true.

THEOREM 4.3 (long-time behavior). *Under assumptions* (A2)–(A4)*, the* ω*-limit set* $\omega\big(\theta, \chi_1, \chi_2, u\big)$ *reduces to the unique solution to the problem*

$$\theta_\infty = \theta_e \quad a.e. \ in \ \ \Omega, \tag{4.2}$$

$$a(u_\infty, v) + (\mathbb{A}\beta\chi_{2,\infty}, \varepsilon(v)) = 0 \quad \forall v \in \mathcal{V}, \tag{4.3}$$

$$\mathbb{A}(\varepsilon(u_\infty) + \beta\chi_{2,\infty}) = \sigma_\infty \quad a.e. \ in \ \ \Omega, \tag{4.4}$$

$$\gamma\begin{pmatrix}\chi_{1,\infty}\\ \chi_{2,\infty}\end{pmatrix} + \partial I_K\begin{pmatrix}\chi_{1,\infty}\\ \chi_{2,\infty}\end{pmatrix} \ni \begin{pmatrix}\frac{\ell}{\theta_*}(\theta_* - \theta_e)\\ -\sigma_\infty : \beta\end{pmatrix} \quad a.e. \ in \ \ \Omega. \tag{4.5}$$

Namely, the whole trajectory converges to $(\theta_\infty, \chi_{1,\infty}, \chi_{2,\infty}, u_\infty)$ *as* $t \to +\infty$*.*

We close this section by stating precisely a maximum principle for the temperature θ which entails its positivity in the frame of our concrete modeling situation. To this aim, of course some sign assumption on the external heat sources is needed and we will ask for

(A5) $\langle F(t), v\rangle \geq 0$ for a.e. $t \in (0, T)$ and all $v \in V$ with $v \geq 0$ a.e. in Ω.

The latter follows for instance when f, $r \geq 0$ almost everywhere in their respective domains and could clearly be weakened. One has the following.

THEOREM 4.4 (lower bound). *Let* $\mu \geq 0$*. Under assumptions* (2.1), (A1)–(A3), *and* (A5)*, let* $\theta_d \in \mathbb{R}$ *be such that*

$$\gamma + \frac{\ell}{\theta_*}(\theta_d - \theta_*) \leq 0. \tag{4.6}$$

Then, the unique solution $(\theta, [\chi_1, \chi_2], u)$ *to Problem* P_μ *fulfills*

$$\inf\{\inf \theta_0, \, \theta_e, \, \theta_*, \theta_d\} \leq \theta(x, t) \quad for\ a.e. \ \ (x, t) \in Q, \tag{4.7}$$

where $\inf \theta_0$ *stands for the essential infimum of* θ_0 *on* Ω*.*

Clearly, in order to deduce from the above-stated lower bound (4.7) a positivity result for θ (and consequently a proof of the thermodynamical consistency of the model; see section 2) one has to start from a positive initial datum θ_0 and ask for the existence of a positive constant θ_d fulfilling (4.6). Let us stress that this second requirement is compatible with our modeling situation since, owing to the discussion of section 2,

$$\gamma+\frac{\ell}{\theta_*}(\theta_d-\theta_*)=\frac{\ell}{\theta_*}(\theta_*-\theta_{**})+\frac{\ell}{\theta_*}(\theta_d-\theta_*)=\frac{\ell}{\theta_*}(\theta_d-\theta_{**}),$$

and it suffices to choose $0<\theta_d\leq\theta_{**}$ in order to achieve (4.6).

5. Dissipative problem. Throughout this section, the dissipation parameter μ is fixed and strictly positive.

5.1. Continuous dependence. Let us denote by $(\theta_i,\chi_{1,i},\chi_{2,i},u_i)$ for $i=1,2$ two solutions to Problem P_μ associated to the given two sets of data $(F_i,G_i,\theta_i^0,\chi_{1,i}^0,$ $\chi_{2,i}^0)$, and $\theta_{e,i}$ for $i=1,2$. We set $\overline{\theta}:=\theta_1-\theta_2$, $\overline{u}:=u_1-u_2$, and so on. Let us take the integral on $(0,t)$, for $t\in(0,T]$, of relation (3.1) written for $(\theta_1,\chi_{1,1},\chi_{2,1},u_1)$ and subtract the same relation for $(\theta_2,\chi_{1,2},\chi_{2,2},u_2)$, choose $\varphi:=\overline{\theta}$, and integrate on $(0,t)$ for $t\in(0,T]$. We readily obtain that

$$\begin{aligned}(5.1)\qquad & c_s\int_0^t\|\overline{\theta}\|^2+\frac{k}{2}\left\|\int_0^t\nabla\overline{\theta}\right\|^2+\frac{h}{2}\left\|\int_0^t\overline{\theta}\right\|^2_{L^2(\Gamma)}\\ &\leq\int_0^t\left(\|c_s\overline{\theta}^0-\ell\overline{\chi}_1^0\|+\left\|\int_0^s\overline{F}_1\right\|\right)\|\overline{\theta}\|\,ds\\ &+\int_0^t\left\langle\int_0^s\overline{F}_2,\overline{\theta}\right\rangle ds+h\int_0^t(s\overline{\theta}_e,\overline{\theta})_\Gamma ds+\ell\int_0^t(\overline{\theta},\overline{\chi}_1).\end{aligned}$$

Let us now take the difference between relation (3.4) written for $(\theta_1,\chi_{1,1},\chi_{2,1},u_1)$ and the same relation for $(\theta_2,\chi_{1,2},\chi_{2,2},u_2)$, multiply the corresponding relation by $[\overline{\chi}_1,\overline{\chi}_2]$, exploit the monotonicity of the subdifferential, and integrate on $\Omega\times(0,t)$ for $t\in(0,T]$. We get

$$(5.2)\quad \sum_{j=1}^2\left(\frac{\mu}{2}\|\overline{\chi}_j(t)\|^2+\gamma\int_0^t\|\overline{\chi}_j\|^2\right)\leq\sum_{j=1}^2\frac{\mu}{2}\|\overline{\chi}_j^0\|^2-\frac{\ell}{\theta_*}\int_0^t(\overline{\theta},\overline{\chi}_1)-\int_0^t(\overline{\sigma}:\beta,\overline{\chi}_2).$$

As for the last term in the right-hand side of (5.2) we take advantage of (3.2)–(3.3) and readily compute that

$$(5.3)\qquad -\overline{\sigma}:\beta=-\mathbb{A}^{\frac12}\varepsilon(\overline{u}):\mathbb{A}^{\frac12}\beta-|\mathbb{A}^{\frac12}\beta|^2\overline{\chi}_2,$$

$$(5.4)\qquad \int_0^t\|\mathbb{A}^{\frac12}\varepsilon(\overline{u})\|^2=-\int_0^t(\mathbb{A}^{\frac12}\varepsilon(\overline{u}):\mathbb{A}^{\frac12}\beta,\overline{\chi}_2)+\int_0^t\langle\overline{G},\overline{u}\rangle.$$

Hence, by choosing $1<\rho<(\gamma+|\mathbb{A}^{\frac12}\beta|^2)/|\mathbb{A}^{\frac12}\beta|^2$, we take the sum between (5.2) and (5.4) in order to obtain

$$\begin{aligned}&\sum_{j=1}^2\left(\frac{\mu}{2}\left(\|\overline{\chi}_j(t)\|^2-\|\overline{\chi}_j^0\|^2\right)+\gamma\int_0^t\|\overline{\chi}_j\|^2\right)+\int_0^t\|\mathbb{A}^{\frac12}\varepsilon(\overline{u})\|^2\\ &\quad\leq-\frac{\ell}{\theta_*}\int_0^t(\overline{\theta},\overline{\chi}_1)-2\int_0^t(\mathbb{A}^{\frac12}\varepsilon(\overline{u}):\mathbb{A}^{\frac12}\beta,\overline{\chi}_2)-|\mathbb{A}^{\frac12}\beta|^2\int_0^t\|\overline{\chi}_2\|^2+\int_0^t\langle\overline{G},\overline{u}\rangle\\ &\quad\leq-\frac{\ell}{\theta_*}\int_0^t(\overline{\theta},\overline{\chi}_1)+(\rho-1)|\mathbb{A}^{\frac12}\beta|^2\int_0^t\|\overline{\chi}_2\|^2+\frac{1}{\rho}\int_0^t\|\mathbb{A}^{\frac12}\varepsilon(\overline{u})\|^2+\int_0^t\langle\overline{G},\overline{u}\rangle.\end{aligned}$$

Then, multiplying (5.2) by $1/\theta_*$ and taking the sum with the above relation, one has

$$
\begin{aligned}
(5.5)\quad & \frac{c_s}{\theta_*}\int_0^t \|\overline{\theta}\|^2 + \frac{k}{2\theta_*}\left\|\int_0^t \nabla\overline{\theta}\right\|^2 + \frac{h}{2\theta_*}\left\|\int_0^t \overline{\theta}\right\|^2_{L^2(\Gamma)} + \frac{\mu}{2}\sum_{j=1}^{2}\|\overline{\chi}_j(t)\|^2 \\
&+\gamma\int_0^t \|\overline{\chi}_1\|^2 + \big(\gamma+(1-\rho)|\mathbb{A}^{\frac12}\beta|^2\big)\int_0^t \|\overline{\chi}_2\|^2 + \frac{\rho-1}{\rho}\int_0^t \|\mathbb{A}^{\frac12}\varepsilon(\overline{u})\|^2 \\
&\le \frac{1}{\theta_*}\int_0^t \left(\|c_s\overline{\theta}^0 - \ell\overline{\chi}_1^0\| + \left\|\int_0^s \overline{F}_1\,ds\right\|\right)\|\overline{\theta}\| + \frac{1}{\theta_*}\int_0^t \left\langle \int_0^s \overline{F}_2\,ds, \overline{\theta}\right\rangle \\
&+\frac{h}{\theta_*}\int_0^t (s\overline{\theta}_e,\overline{\theta})_\Gamma ds + \sum_{j=1}^{2}\frac{\mu}{2}\|\overline{\chi}_j^0\|^2 + \int_0^t \langle \overline{G},\overline{u}\rangle.
\end{aligned}
$$

Finally, the assertion follows from an integration by parts.

5.2. Discretization. Let us introduce our variable time-step discretization of P_μ. To this aim we define the partition $\mathcal{P} := \big\{0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T\big\}$ with variable time-step $\tau_i := t_i - t_{i-1}$ and let $\tau := \max_{1\le i\le N}\tau_i$ denote the diameter of the partition $\mathcal{P}$. In the forthcoming analysis the following notation will be extensively used: $\{w_i\}_{i=0}^N$ being a vector, we denote by $w_{\mathcal{P}}$ and $\overline{w}_{\mathcal{P}}$ two functions of the time interval $[0,T]$ which interpolate the values of the vector $\{w_i\}$ piecewise linearly and backward constantly on the partition $\mathcal{P}$, respectively. Namely,

$$
\begin{aligned}
w_{\mathcal{P}}(0) &:= w_0, \quad w_{\mathcal{P}}(t) := g_i(t)w_i + \big(1-g_i(t)\big)w_{i-1}, \\
\overline{w}_{\mathcal{P}}(0) &:= w_0, \quad \overline{w}_{\mathcal{P}}(t) := w_i \quad \text{for } t\in(t_{i-1},t_i], \quad i=1,\dots,N,
\end{aligned}
$$

where $g_i(t) := (t-t_{i-1})/\tau_i$ for $t\in(t_{i-1},t_i]$, $i=1,\dots,N$. Moreover, given a vector $\{w_i\}_{i=0}^N$, we define another vector $\{\delta w_i\}_{i=1}^N$ as $\delta w_i := (w_i - w_{i-1})/\tau_i$.

Finally, we introduce some approximation of the data. Hence, let $F = F_1 + F_2$, where $F_1 \in L^2(0,T;H)$ and $F_2 \in W^{1,1}(0,T;V')$, and set

$$(5.6)\qquad F_{1,i} := \frac{1}{\tau_i}\int_{t_{i-1}}^{t_i} F_1(s)\,ds \ \in H \quad \text{for } i=1,\dots,N,$$

$$(5.7)\qquad F_{2,1} := F_2(t_i) \ \in V' \quad \text{for } i=0,1,\dots,N,$$

$$(5.8)\qquad G_i := G(t_i) \ \in \mathcal{V}' \quad \text{for } i=0,1,\dots,N.$$

Of course, owing to (A1), the latter positions are justified. In particular, let us remark that one has

$$(5.9)\qquad \overline{F}_{1,\mathcal{P}} \to F_1 \ \text{ strongly in } L^2(0,T;H),$$

$$(5.10)\qquad F_{2,\mathcal{P}} \to F_2 \ \text{ strongly in } W^{1,1}(0,T;V'),$$

$$(5.11)\qquad G_{\mathcal{P}} \to G \ \text{ strongly in } H^1(0,T;\mathcal{V}')$$

whenever the diameter τ of partition $\mathcal{P}$ goes to 0.

Hence, we are interested in the following discrete problem.

Problem D_μ. Find $\{\theta_i\}_{i=0}^N \in V^{N+1}$, $\{[\chi_{1,i},\chi_{2,i}]\}_{i=0}^N \in \mathcal{K}^{N+1}$, and $\{u_i\}_{i=1}^N \in \mathcal{V}^N$ such that, for all $i=1,\dots,N$,

$$\big((c_s\delta\theta_i - \ell\delta\chi_{1,i}),\varphi\big) + k(\nabla\theta_i,\nabla\varphi) + h(\theta_i-\theta_e,\varphi)_\Gamma = \langle F_i,\varphi\rangle \quad \forall\varphi\in V, \tag{5.12}$$

$$a(u_i,v) + (\mathbb{A}\beta\chi_{2,i},\varepsilon(v)) = \langle G_i,v\rangle \quad \forall v\in\mathcal{V}, \tag{5.13}$$

$$\mathbb{A}(\varepsilon(u_i)+\beta\chi_{2,i}) = \sigma_i \quad \text{a.e. in } \Omega, \tag{5.14}$$

$$\mu\begin{pmatrix}\delta\chi_{1,i}\\ \delta\chi_{2,i}\end{pmatrix} + \gamma\begin{pmatrix}\chi_{1,i}\\ \chi_{2,i}\end{pmatrix} + \begin{pmatrix}\frac{\ell}{\theta_*}(\theta_i-\theta_*)\\ \sigma_i:\beta\end{pmatrix} + \partial I_K\begin{pmatrix}\chi_{1,i}\\ \chi_{2,i}\end{pmatrix} \ni \begin{pmatrix}0\\0\end{pmatrix} \quad \text{a.e. in } \Omega, \tag{5.15}$$

$$\theta_0 = \theta^0 \quad \text{a.e. in } \Omega, \tag{5.16}$$

$$\mu[\chi_{1,0},\chi_{2,0}] = \mu[\chi_1^0,\chi_2^0] \quad \text{a.e. in } \Omega. \tag{5.17}$$

5.3. Discrete well-posedness. We prove the following lemma.

LEMMA 5.1. *Under the assumptions* (A1)–(A3), (5.6)–(5.8)*, and for all* τ *sufficiently small, problem* D_μ *has a unique solution.*

Proof. We proceed by induction. Namely, we assume to know the solution of the problem up to level $i-1$ and solve for level i. In particular, we are concerned with the problem of finding $\theta_i\in V$, $[\chi_{1,i},\chi_{2,i}]\in\mathcal{K}$, and $u_i\in\mathcal{V}$ such that

$$\big((c_s\theta_i - \ell\chi_{1,i}),\varphi\big) + \tau_i k(\nabla\theta_i,\nabla\varphi) + \tau_i h(\theta_i-\theta_e,\varphi)_\Gamma = \langle F_i^*,\varphi\rangle \quad \forall\varphi\in V, \tag{5.18}$$

$$a(u_i,v) + (\mathbb{A}\beta\chi_{2,i},\varepsilon(v)) = \langle G_i,v\rangle \quad \forall v\in\mathcal{V}, \tag{5.19}$$

$$\mathbb{A}(\varepsilon(u_i)+\beta\chi_{2,i}) = \sigma_i \quad \text{a.e. in } \Omega, \tag{5.20}$$

$$\mu\begin{pmatrix}\chi_{1,i}\\ \chi_{2,i}\end{pmatrix} + \tau_i\begin{pmatrix}\gamma\chi_{1,i}+\frac{\ell}{\theta_*}(\theta_i-\theta_*)\\ \gamma\chi_{2,i}+\sigma_i:\beta\end{pmatrix} + \partial I_K\begin{pmatrix}\chi_{1,i}\\ \chi_{2,i}\end{pmatrix} \ni \mu\begin{pmatrix}\chi_{1,i-1}\\ \chi_{2,i-1}\end{pmatrix} \quad \text{a.e. in } \Omega, \tag{5.21}$$

where we collected in the right-hand sides of (5.18)–(5.19) and (5.21) the quantities known at level i and let $F_i^* := \tau_i F_i + c_s\theta_{i-1} - \ell\chi_{1,i-1}$. We shall stress that the latter scheme is of course fully implicit.

Let us now fix $[\tilde\chi_1,\tilde\chi_2]\in\mathcal{K}$. It is then straightforward to find the unique solutions $\theta\in V$ and $u\in\mathcal{V}$ to (5.18) with $\tilde\chi_1$ instead of $\chi_{1,i}$ and (5.19) with $\tilde\chi_2$ instead of $\chi_{2,i}$, respectively. Hence, we have implicitly defined a mapping $T_1:\mathcal{K}\to V\times\mathcal{V}$ as $T_1[\tilde\chi_1,\tilde\chi_2] := [\theta,u]$. On the other hand, for all $(\tilde\theta,\tilde u,\tilde\chi_2)\in V\times\mathcal{V}\times H$ there exists a unique pair $[\chi_1,\chi_2]\in\mathcal{K}$ solving relation (5.21) with $(\tilde\theta,\tilde u)$ instead of (θ_i,u_i) and σ_i is defined by (5.20) with $\tilde\chi_2$ instead of $\chi_{2,i}$. Thus, one may define a mapping $T_2: V\times\mathcal{V}\times H\to\mathcal{K}$ as $T_2(\tilde\theta,\tilde u,\tilde\chi_2) = [\chi_1,\chi_2]$.

Our next aim is to prove that, for sufficiently small τ, the mapping $T_3:\mathcal{K}\to\mathcal{K}$ defined as $T_3[\tilde\chi_1,\tilde\chi_2] := T_2(T_1[\tilde\chi_1,\tilde\chi_2],\tilde\chi_2)$ is a contraction in $H\times H$. To this end let $[\tilde\chi_{1,j},\tilde\chi_{2,j}]\in\mathcal{K}$, $[\theta_j,u_j] = T_1[\tilde\chi_{1,j},\tilde\chi_{2,j}]$, $[\chi_{1,j},\chi_{2,j}] := T_3[\tilde\chi_{1,j},\tilde\chi_{2,j}]$ for $j=1,2$, and define $\overline\theta := \theta_1-\theta_2$, $\overline u := u_1-u_2$, etc. Hence, we readily check that

$$(c_s\overline\theta,\varphi) + \tau_i(\nabla\overline\theta,\nabla\varphi) + \tau_i h(\overline\theta,\varphi)_\Gamma = (\ell\overline{\tilde\chi}_1,\varphi) \quad \forall\varphi\in V,$$
$$a(\overline u,v) + (\mathbb{A}\beta\overline{\tilde\chi}_2,\varepsilon(v)) = 0 \quad \forall v\in\mathcal{V},$$
$$\sigma:\beta = \mathbb{A}\varepsilon(\overline u):\beta + \mathbb{A}\beta\overline{\tilde\chi}_2:\beta.$$

Namely, by choosing $[\varphi,v] = [\overline\theta,\overline u]$ above one gets that

$$\|\overline\theta\| \le \frac{\ell}{c_s}\|\overline{\tilde\chi}_1\|, \quad \|\mathbb{A}^{\frac12}\varepsilon(\overline u)\| \le c_1\|\overline{\tilde\chi}_2\|, \quad \|\overline\sigma:\beta\| \le c_1\|\overline{\tilde\chi}_2\|, \tag{5.22}$$

where c_1 depends only on $\mathbb{A}^{\frac{1}{2}}\beta$. On the other hand, we exploit (5.21) and get that

$$\begin{pmatrix}\chi_{1,j}\\ \chi_{2,j}\end{pmatrix} = (1+\partial I_K)^{-1}\left(\frac{\tau_i}{\mu+\tau_i\gamma}\right)\begin{pmatrix}\frac{\ell}{\theta_*}(\theta_*-\theta_j)+\mu\chi_{1,i-1}/\tau_i\\ -\sigma_j:\beta+\mu\chi_{2,i-1}/\tau_i\end{pmatrix} \quad \text{for} \quad j=1,2.$$

In particular, also using (5.22), one computes that

$$\begin{aligned}\|\overline{\chi}_1\|^2+\|\overline{\chi}_2\|^2 &\le \left(\frac{\tau_i}{\mu+\tau_i\gamma}\right)^2\left(\left\|\frac{\ell}{\theta_*}\overline{\theta}\right\|^2+\|\overline{\sigma}:\beta\|^2\right)\\ &\le \left(\frac{\tau_i}{\mu+\tau_i\gamma}\right)^2\left(\frac{\ell^4}{c_s^2\theta_*^2}\|\overline{\widetilde{\chi}}_1\|^2+c_1^2\|\overline{\widetilde{\chi}}_2\|^2\right).\end{aligned}$$

Finally it suffices to fix $\tau\le\mu/\max\{\ell^2/(c_s\theta_*),c_1\}$ in order to get that T_3 is actually a contraction in $H\times H$. The assertion follows from the fact that $T_3(H)\subset\mathcal{K}$ which is closed in $H\times H$. □

For the sake of later convenience, we rewrite the scheme (5.12)–(5.15) in more compact form as

$$\begin{aligned}&(5.23)\quad \big(\partial_t(c_s\theta_{\mathcal{P}}-\ell\chi_{1,\mathcal{P}}),\varphi\big)+k(\nabla\overline{\theta}_{\mathcal{P}},\nabla\varphi)+h(\overline{\theta}_{\mathcal{P}}-\theta_e,\varphi)_\Gamma\\ &\qquad\qquad =\langle\overline{F}_{\mathcal{P}},\varphi\rangle\quad\forall\varphi\in V,\ \text{ a.e. in } (0,T),\\ &(5.24)\quad a(\overline{u}_{\mathcal{P}},v)+(\mathbb{A}\beta\overline{\chi}_{2,\mathcal{P}},\varepsilon(v))=\langle\overline{G}_{\mathcal{P}},v\rangle\quad\forall v\in\mathcal{V},\ \text{ a.e. in } (0,T),\\ &(5.25)\quad \mathbb{A}(\varepsilon(\overline{u}_{\mathcal{P}})+\beta\overline{\chi}_{2,\mathcal{P}})=\overline{\sigma}_{\mathcal{P}}\quad\text{a.e. in } Q,\\ &(5.26)\quad \mu\partial_t\begin{pmatrix}\chi_{1,\mathcal{P}}\\ \chi_{2,\mathcal{P}}\end{pmatrix}+\gamma\begin{pmatrix}\overline{\chi}_{1,\mathcal{P}}\\ \overline{\chi}_{2,\mathcal{P}}\end{pmatrix}+\partial I_K\begin{pmatrix}\overline{\chi}_{1,\mathcal{P}}\\ \overline{\chi}_{2,\mathcal{P}}\end{pmatrix}\ni\begin{pmatrix}\frac{\ell}{\theta_*}(\theta_*-\overline{\theta}_{\mathcal{P}})\\ -\overline{\sigma}_{\mathcal{P}}:\beta\end{pmatrix}\quad\text{a.e. in } Q.\end{aligned}$$

Remark 5.2. In order to completely justify the above notation one could consider, for instance, $u_0:=u^0$ and $\sigma_0:=\sigma^0$, where u^0 and σ^0 are defined above.

5.4. Stability. Our approximation scheme fulfills some suitable *conditional stability* property. In particular, this subsection brings us to the proof of the following lemma.

LEMMA 5.3. *Under the assumptions* (A1)–(A3), (5.6)–(5.8), *and for all τ sufficiently small, let $\{\theta_i\}_{i=0}^N\in V^{N+1}$, $\{[\chi_{1,i},\chi_{2,i}]\}_{i=0}^N\in\mathcal{K}^{N+1}$, and $\{u_i\}_{i=1}^N\in\mathcal{V}^N$ be the unique solution to Problem D_μ. Then there exists a positive constant c_2 depending only on c_s, k, h, θ_e, Γ, θ_*, γ, $\mathbb{A}^{\frac{1}{2}}\beta$, $c_{\mathcal{V}}$, θ^0, $[\chi_1^0,\chi_2^0]$, $\|F_1\|_{L^2(0,T;H)}$, $\|F_2\|_{W^{1,1}(0,T;V')}$, and $\|G\|_{H^1(0,T;\mathcal{V}')}$ such that*

$$\begin{aligned}(5.27)\qquad &\|\theta_{\mathcal{P}}\|_{H^1(0,T;H)\cap C^0([0,T];V)}+\|[\chi_{1,\mathcal{P}},\chi_{2,\mathcal{P}}]\|_{(H^1(0,T;H))^2}\\ &+\sqrt{\mu}\|[\chi_{1,\mathcal{P}},\chi_{2,\mathcal{P}}]\|_{(W^{1,\infty}(0,T;H))^2}+\|u\|_{H^1(0,T;\mathcal{V})}\le c_2.\end{aligned}$$

In particular, c_2 is independent of μ and τ.

Proof. Henceforth we will denote by c any positive constant, possibly depending on data but on neither μ nor $\mathcal{P}$. In particular, c may vary from line to line.

Let us take the difference between relation (5.15) written at level i and the same relation at level $i-1$. By defining $[\delta\chi_{1,0},\delta\chi_{2,0}]:=[\chi_{1,\mu,t}(0),\chi_{2,\mu,t}(0)]$ we are entitled to do so for $i=1,\dots,N$. Next, we multiply the resulting relation by $[\delta\chi_{1,i},\delta\chi_{2,i}]$, integrate in space, and sum for $i=1,\dots,m$ for some $m=1,\dots,N$. By exploiting the monotonicity of the subdifferential, we obtain that

$$\sum_{j=1}^{2}\left(\frac{\mu}{2}\|\delta\chi_{j,m}\|^2-\frac{\mu}{2}\|\delta\chi_j(0)\|^2+\gamma\sum_{i=1}^{m}\tau_i\|\delta\chi_{j,i}\|^2\right)$$
$$\leq -\frac{\ell}{\theta_*}\sum_{i=1}^{m}\tau_i(\delta\theta_i,\delta\chi_{1,i})-\sum_{i=1}^{m}\tau_i(\delta\sigma_i:\beta,\delta\chi_{2,i}). \tag{5.28}$$

We now take the difference between relation (5.13) written at level i and the same relation at level $i-1$, choose $v:=\delta u_i$, and sum for $i=1,\dots,m$. One readily gets that

$$\sum_{i=1}^{m}\tau_i\|\mathbb{A}^{\frac12}\varepsilon(\delta u_i)\|^2=\sum_{i=1}^{m}\tau_i\langle\delta G_i,\delta u_i\rangle-\sum_{i=1}^{m}\tau_i(\mathbb{A}^{\frac12}\varepsilon(\delta u_i):\mathbb{A}^{\frac12}\beta,\delta\chi_{2,i}). \tag{5.29}$$

On the other hand, owing to (5.14), it may be easily computed that

$$\delta\sigma_i:\beta=\mathbb{A}^{\frac12}\varepsilon(\delta u_i):\mathbb{A}^{\frac12}\beta+|\mathbb{A}^{\frac12}\beta|^2\delta\chi_{2,i}\quad\forall i=1,\dots,N. \tag{5.30}$$

Hence, taking into account (5.30) and adding (5.28) to (5.29), we may again choose a suitable ρ such that $1<\rho<(\gamma+|\mathbb{A}^{\frac12}\beta|^2)/|\mathbb{A}^{\frac12}\beta|^2$ and deduce that

$$\sum_{j=1}^{2}\left(\frac{\mu}{2}\|\delta\chi_{j,m}\|^2-\frac{\mu}{2}\|\delta\chi_j(0)\|^2\right)$$
$$+\sum_{i=1}^{m}\tau_i\Big(\gamma\|\delta\chi_{1,i}\|^2+(\gamma+|\mathbb{A}^{\frac12}\beta|^2-\rho|\mathbb{A}^{\frac12}\beta|^2)\|\delta\chi_{2,i}\|^2\Big)$$
$$+\frac{\rho-1}{\rho}\sum_{i=1}^{m}\tau_i\|\mathbb{A}^{\frac12}\varepsilon(\delta u_i)\|^2$$
$$\leq -\frac{\ell}{\theta_*}\sum_{i=1}^{m}\tau_i(\delta\theta_i,\delta\chi_{1,i})+\sum_{i=1}^{m}\tau_i\langle\delta G_i,\delta u_i\rangle. \tag{5.31}$$

Next, we test relation (5.12) by $\varphi=\tau_i\delta\theta_i$ and take the sum for $i=1,\dots,m$. Because of (A2) we obtain that

$$c_s\sum_{i=1}^{m}\tau_i\|\delta\theta_i\|^2+\frac{k}{2}\|\nabla\theta_m\|^2+\frac{h}{2}\|\theta_m\|^2_{L^2(\Gamma)}$$
$$\leq c+h(\theta_e,\theta_m)_\Gamma+\ell\sum_{i=1}^{m}\tau_i(\delta\theta_i,\delta\chi_{1,i})$$
$$+\sum_{i=1}^{m}\tau_i(F_{1,i},\delta\theta_i)+\langle F_{2,m},\theta_m\rangle-\langle F_{2,1},\theta^0\rangle-\sum_{i=2}^{m}\langle F_{2,i}-F_{2,i-1},\theta_{i-1}\rangle. \tag{5.32}$$

Finally, it suffices to take the sum between (5.31) and (5.32) multiplied by $1/\theta_*$, consider (A3), and perform some standard computations in order to obtain that

$$\begin{aligned}(5.33)\qquad & \mu\sum_{j=1}^{2}\|\delta\chi_{j,m}\|^2+\|\theta_m\|_V^2+\sum_{i=1}^{m}\tau_i\left(\sum_{j=1}^{2}\|\delta\chi_{j,i}\|^2+\|\delta\theta_i\|^2+\|\delta u_i\|_{\mathcal{V}}^2\right)\\ & \le c\sum_{i=2}^{m-1}\tau_i\langle\delta F_{2,i},\theta_{i-1}\rangle\\ & +c\left(1+\|F_{2,m}\|_{V'}^2+\|F_{2,1}\|_{V'}^2+\sum_{i=1}^{m}\tau_i\left(\|\delta_i G\|_{\mathcal{V}'}^2+\|F_{1,i}\|^2\right)\right).\end{aligned}$$

Finally, from (5.6) and (5.8) and an application of the discrete Gronwall lemma we readily obtain the bounds of (5.28). □

5.5. Convergence. Let us now consider the limit as the diameter τ of partition $\mathcal{P}$ goes to zero. We are actually in a position to prove the following.

LEMMA 5.4. *Under the assumptions* (A1)–(A3), (5.6)–(5.8)*, and for all τ sufficiently small, let $\{\theta_i\}_{i=0}^N\in V^{N+1}$, $\{[\chi_{1,i},\chi_{2,i}]\}_{i=0}^N\in\mathcal{K}^{N+1}$, and $\{u_i\}_{i=1}^N\in\mathcal{V}^N$ be the unique solution to problem D_μ. Then the following convergences hold:*

$$\begin{aligned}(5.34)\qquad & \theta_{\mathcal{P}}\longrightarrow\theta_\mu \quad \textit{weakly star in}\ \ H^1(0,T;H)\cap L^\infty(0,T;V)\\ & \qquad\qquad \textit{and strongly in}\ \ C([0,T];H),\\ (5.35)\qquad & \overline{\theta}_{\mathcal{P}}\longrightarrow\theta_\mu \quad \textit{weakly star in}\ \ L^\infty(0,T;V)\\ & \qquad\qquad \textit{and strongly in}\ \ L^\infty(0,T;H),\\ (5.36)\qquad & \chi_{j,\mathcal{P}}\longrightarrow\chi_{j,\mu} \quad \textit{weakly star in}\ \ W^{1,\infty}(0,T;H)\\ & \qquad\qquad \textit{and strongly in}\ \ C([0,T];H),\quad j=1,2,\\ (5.37)\qquad & \overline{\chi}_{j,\mathcal{P}}\longrightarrow\chi_{j,\mu} \quad \textit{strongly in}\ \ L^\infty(0,T;H),\quad j=1,2,\\ (5.38)\qquad & u_{\mathcal{P}}\longrightarrow u_\mu \quad \textit{weakly in}\ \ H^1(0,T;\mathcal{V})\\ & \qquad\qquad \textit{and strongly in}\ \ C([0,T];\mathcal{V}),\\ (5.39)\qquad & \overline{u}_{\mathcal{P}}\longrightarrow u_\mu \quad \textit{strongly in}\ \ L^\infty(0,T;\mathcal{V}),\end{aligned}$$

where $(\theta_\mu,\chi_{1,\mu},\chi_{2,\mu},u_\mu)$ is the unique solution to Problem P_μ.

In particular, let us stress that the latter lemma entails the proof of the existence statement of Theorem 4.1.

Proof. Taking into account the estimate (5.28) and well-known compactness results, we readily find a quadruple $(\theta_\mu,\chi_{1,\mu},\chi_{2,\mu},u_\mu)$ such that, possibly taking nonrelabeled subsequences, the weak and weak-star convergences of Lemma 5.4 hold true together with the following:

$$\begin{aligned}(5.40)\qquad & \theta_{\mathcal{P}}\longrightarrow\theta_\mu \quad \text{strongly in}\ \ C([0,T];H),\\ (5.41)\qquad & \overline{\theta}_{\mathcal{P}}\longrightarrow\theta_\mu \quad \text{strongly in}\ \ L^\infty(0,T;H).\end{aligned}$$

We now turn to the proof of some strong convergence for $\chi_{1,\mathcal{P}},\chi_{2,\mathcal{P}}$, and $u_{\mathcal{P}}$ by a direct Cauchy argument. To this aim, let $\mathcal{P}_m$ denote the extracted sequence of partitions. We denote $\theta_m:=\theta_{\mathcal{P}_m}$, $u_m:=u_{\mathcal{P}_m}$, etc. By taking the difference between (5.26) written for $\mathcal{P}_n$ and the same relation for $\mathcal{P}_m$, multiplying it by $[\overline{\chi}_{1,n}-\overline{\chi}_{1,m},\overline{\chi}_{2,n}-\overline{\chi}_{2,m}]$, integrating on Ω, and exploiting the monotonicity of the subdifferential we readily obtain that

$$
\begin{aligned}
(5.42)\quad & \frac{\mu}{2}\frac{d}{dt}\sum_{j=1}^{2}\|(\chi_{j,n}-\chi_{j,m})(t)\|^2+\gamma\sum_{j=1}^{2}\|(\overline{\chi}_{j,n}-\overline{\chi}_{j,m})(t)\|^2\\
&\leq \frac{\ell}{\theta_*}\|(\overline{\theta}_n-\overline{\theta}_m)(t)\|\,\|(\overline{\chi}_{1,n}-\overline{\chi}_{1,m})(t)\|-\big((\overline{\sigma}_n-\overline{\sigma}_m)(t):\beta,(\overline{\chi}_{2,n}-\overline{\chi}_{2,m})(t)\big)
\end{aligned}
$$

for almost every $t\in(0,T)$. In particular, we made a crucial use of the fact that, given any vector $\{w_i\}_{i=0}^N\in H^{N+1}$, one has that $(w'_{\mathcal{P}},\overline{w}_{\mathcal{P}})\geq(w'_{\mathcal{P}},w_{\mathcal{P}})$ since of course the residual term $(w'_{\mathcal{P}},\overline{w}_{\mathcal{P}}-w_{\mathcal{P}})$ is nonnegative. On the other hand, by taking the difference of the corresponding relations (5.24) with $v=\overline{u}_n-\overline{u}_m$ and of (5.25) we readily check that

$$
\begin{aligned}
(5.43)\qquad \|\mathbb{A}^{\frac12}\varepsilon(\overline{u}_n-\overline{u}_m)\|^2 &= -\big(\mathbb{A}^{\frac12}\varepsilon(\overline{u}_n-\overline{u}_m):\mathbb{A}^{\frac12}\beta,(\overline{\chi}_{2,n}-\overline{\chi}_{2,m})\big)\\
&\quad+\langle\overline{G}_n-\overline{G}_m,\overline{u}_n-\overline{u}_m\rangle,
\end{aligned}
$$

$$
\begin{aligned}
(5.44)\qquad & -\big((\overline{\sigma}_n-\overline{\sigma}_m):\beta,\overline{\chi}_{2,n}-\overline{\chi}_{2,m}\big)\\
&= -\,(\mathbb{A}^{\frac12}\varepsilon(\overline{u}_n-\overline{u}_m):\mathbb{A}^{\frac12}\beta,\overline{\chi}_{2,n}-\overline{\chi}_{2,m})-|\mathbb{A}^{\frac12}\beta|^2\|\overline{\chi}_{2,n}-\overline{\chi}_{2,m}\|^2.
\end{aligned}
$$

Hence, owing to the latter relation, taking the sum of (5.42) and (5.43) and integrating on $(0,t)$ for some $t\in(0,T]$ we easily infer that

$$
\begin{aligned}
&\sum_{j=1}^{2}\left(\frac{\mu}{2}\|(\chi_{j,n}-\chi_{j,m})(t)\|^2+\gamma\int_0^t\|\overline{\chi}_{j,n}-\overline{\chi}_{j\,m}\|^2\right)+\int_0^t\|\mathbb{A}^{\frac12}\varepsilon(\overline{u}_n-\overline{u}_m)\|^2\\
&\leq\frac{\ell}{\theta_*}\int_0^t\|\overline{\theta}_n-\overline{\theta}_m\|\,\|\overline{\chi}_{1,n}-\overline{\chi}_{1,m}\|-2\int_0^t(\mathbb{A}^{\frac12}\varepsilon(\overline{u}_n-\overline{u}_m):\mathbb{A}^{\frac12}\beta,\overline{\chi}_{2,n}-\overline{\chi}_{2,m})\\
&-|\mathbb{A}^{\frac12}\beta|^2\int_0^t\|\overline{\chi}_{2,n}-\overline{\chi}_{2,m}\|^2+\int_0^t\langle\overline{G}_n-\overline{G}_m,\overline{u}_n-\overline{u}_m\rangle\\
&\leq\frac{\gamma}{2}\int_0^t\|\overline{\chi}_{1,n}-\overline{\chi}_{1,m}\|^2+(\rho-1)|\mathbb{A}^{\frac12}\beta|^2\int_0^t\|\overline{\chi}_{2,n}-\overline{\chi}_{2,m}\|^2\\
&+\frac{1}{\rho}\int_0^t\|\mathbb{A}^{\frac12}\varepsilon(\overline{u}_n-\overline{u}_m)\|^2+\frac{\ell^2}{2\theta_*^2\gamma}\int_0^t\|\overline{\theta}_n-\overline{\theta}_m\|^2+\int_0^t\langle\overline{G}_n-\overline{G}_m,\overline{u}_n-\overline{u}_m\rangle
\end{aligned}
$$

for some $1<\rho<(\gamma+|\mathbb{A}^{\frac12}\beta|^2)/|\mathbb{A}^{\frac12}\beta|^2$. Hence, in particular,

$$
\begin{aligned}
(5.45)\quad & \sum_{j=1}^{2}\left(\mu\|\chi_{j,n}-\chi_{j,m}\|^2_{C([0,t];H)}+\int_0^t\|\overline{\chi}_{j,n}-\overline{\chi}_{j\,m}\|^2\right)+\int_0^t\|\overline{u}_n-\overline{u}_m\|^2_{\mathcal{V}}\\
&\leq c\left(\int_0^t\|\overline{\theta}_n-\overline{\theta}_m\|^2+\int_0^t\|\overline{G}_n-\overline{G}_m\|^2_{\mathcal{V}'}\right),
\end{aligned}
$$

where c depends on ℓ, θ_*, γ, and $c_{\mathcal{V}}$. Finally it suffices to exploit (5.11) and (5.41) in order to obtain that $[\chi_{1,\mathcal{P}},\chi_{2,\mathcal{P}}]$ is a Cauchy sequence in $(C([0,T];H))^2$. Namely, we checked the strong convergences in (5.36)–(5.37). The strong convergences of (5.38)–(5.39) are now an easy consequence of (5.11), (5.36)–(5.37), and (5.43).

We prove that indeed the quadruple $(\theta_\mu, \chi_{1,\mu}, \chi_{2,\mu}, u_\mu)$ solves Problem P_μ. Let us introduce the new auxiliary variables $[\overline{\xi}_{1,\mathcal{P}}, \overline{\xi}_{2,\mathcal{P}}] \in \partial I_K(\overline{\chi}_{1,\mathcal{P}}, \overline{\chi}_{2,\mathcal{P}})$ almost everywhere in Q such that (5.26) reduces to the following equality:

$$\mu\partial_t\begin{pmatrix}\chi_{1,\mathcal{P}}\\ \chi_{2,\mathcal{P}}\end{pmatrix} + \gamma\begin{pmatrix}\overline{\chi}_{1,\mathcal{P}}\\ \overline{\chi}_{2,\mathcal{P}}\end{pmatrix} + \begin{pmatrix}\frac{\ell}{\theta_*}(\overline{\theta}_{\mathcal{P}} - \theta_*)\\ \overline{\sigma}_{\mathcal{P}} : \beta\end{pmatrix} + \begin{pmatrix}\overline{\xi}_{1,\mathcal{P}}\\ \overline{\xi}_{2,\mathcal{P}}\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix} \quad \text{a.e. in } Q.$$

Hence, moving from (5.28), it is straightforward to possibly extract a further subsequence such that the convergences of Lemma 5.4 hold and there exists $[\xi_{1,\mu}, \xi_{2,\mu}]$ such that

$$\xi_{j,\mathcal{P}} \to \xi_{j,\mu} \quad \text{weakly in } L^2(0,T;H) \quad \text{for } j = 1,2.$$

Owing to the convergences proved above and to (5.9)–(5.11) it is now possible to pass to the limit as the diameter τ of partition $\mathcal{P}$ goes to zero and check that $(\theta_\mu, \chi_{1,\mu}, \chi_{2,\mu}, u_\mu, \xi_{1,\mu}, \xi_{2,\mu})$ fulfills relations (3.1)–(3.3), (3.5)–(3.6), and

$$\mu\partial_t\begin{pmatrix}\chi_{1,\mu}\\ \chi_{2,\mu}\end{pmatrix} + \gamma\begin{pmatrix}\chi_{1,\mu}\\ \chi_{2,\mu}\end{pmatrix} + \begin{pmatrix}\xi_{1,\mu}\\ \xi_{2,\mu}\end{pmatrix} \ni \begin{pmatrix}\frac{\ell}{\theta_*}(\theta_* - \theta_\mu)\\ -\sigma_\mu : \beta\end{pmatrix} \quad \text{a.e. in } Q. \tag{5.46}$$

Moreover, it is straightforward to check that

$$\int_0^T (\overline{\xi}_{j,\mathcal{P}}, \overline{\chi}_{j,\mathcal{P}}) \to \int_0^T (\xi_{j,\mu}, \chi_{j,\mu}) \qquad \text{for } j = 1,2.$$

In particular, classical results on maximal monotone operators [7, Prop. 2.5, p. 27] entail that $[\xi_{1,\mu}, \xi_{2,\mu}] \in \partial I_K(\chi_{1,\mu}, \chi_{2,\mu})$ almost everywhere in Q and the assertion of the lemma follows. Before closing this proof, we observe that the convergences stated in the lemma hold for all the sequence of partitions and not just for a subsequence since the solution to Problem P_μ is unique. □

5.6. Error estimates. For the sake of completeness, we state here a priori bound on the discretization error. To this aim, recalling from (A1) that $F := F_1 + F_2$, we sharpen our regularity requirements by asking for

$$F_1 \in BV([0,T];H), \tag{5.47}$$

where the above notation refers to the space of real functions valued in H with bounded variation. We have the following estimate.

LEMMA 5.5. *Under the assumptions* (A1)–(A3), (5.6)–(5.8), *and* (5.47), *let* $(\theta, [\chi_1, \chi_2], u)$ *and* $\{\theta_i\}_{i=0}^N \in V^{N+1}$, $\{[\chi_{1,i}, \chi_{2,i}]\}_{i=0}^N \in \mathcal{K}^{N+1}$, *and* $\{u_i\}_{i=1}^N \in \mathcal{V}^N$ *be the unique solutions to Problems* P_μ *and* D_μ, *respectively. Hence, there exists a positive constant* c_3 *depending on* c_2 *such that, possibly taking* τ *small enough, one has that*

$$\begin{aligned} &\|\theta - \theta_{\mathcal{P}}\|_{L^2(0,t;H)} + \sup_{s\in[0,t]}\left\|\int_0^s \nabla(\theta - \overline{\theta}_{\mathcal{P}})\right\| + \sup_{s\in[0,t]}\left\|\int_0^s (\theta - \overline{\theta}_{\mathcal{P}})\right\|_{L^2(\Gamma)} \\ &\quad + \mu\sum_{j=1}^2 \|\chi_j - \chi_{j,\mathcal{P}}\|_{C([0,t];H)} + \sum_{j=1}^2 \|\chi_j - \chi_{j,\mathcal{P}}\|_{L^2(0,t,H)} \\ &\quad + \|u - u_{\mathcal{P}}\|_{C([0,t];\mathcal{V})} \le c_3\tau \quad \forall t \in [0,T]. \end{aligned} \tag{5.48}$$

We will not give here the detailed proof of the latter result. Indeed, the argument follows closely the lines of the proof of the continuous dependence estimate (4.1). The additional intricacy related to the fact that the continuous and the discrete solutions do not solve the same equations may be overcome by the same techniques of the two papers [34, 35], where indeed the abstract analysis of [28, 29] is applied in a similar context. On the other hand, some comment is in order. We point out that the latter estimate is *optimal* with respect to the order of convergence, since the backward Euler scheme is used in order to approximate time derivatives in Problem P_μ. Moreover, no a priori constraints between consecutive time-steps are imposed. Hence (5.48) ensures the possibility of implementing an adaptive procedure as in [29]. Finally, let us remark that c_3 depends exponentially on T since the Gronwall lemma is used in the proof of (5.48).

6. Nondissipative problem. We now proceed to the proof of Theorem 4.1 for $\mu = 0$ and Theorem 4.2. To this aim $(\theta_\mu, \chi_{1,\mu}, \chi_{2,\mu}, u_\mu)$ will denote a sequence of solutions to Problem P_μ as μ converges toward 0.

6.1. Continuous dependence. First of all, we address the continuous dependence claim in the nondissipative situation. To this end, it suffices to follow exactly the lines of the corresponding proof of subsection 5.1 with the choice of the parameter $\mu = 0$.

6.2. A priori estimates. We shall prove the following.

LEMMA 6.1. *Under the assumptions* (A1)–(A3)*, let* $(\theta_\mu, \chi_{1,\mu}, \chi_{2,\mu}, u_\mu)$ *be solutions to Problem* P_μ*. Then, there exists a positive constant* c_4 *depending only on* c_s, k, h, θ_e, Γ, γ, $\mathbb{A}^{\frac{1}{2}}\beta$, $c_{\mathcal{V}}$, θ_*, θ^0, $[\chi_1^0, \chi_2^0]$, $\|F_1\|_{L^2(0,T;H)}$, $\|F_2\|_{W^{1,1}(0,T;V')}$*, and* $\|G\|_{H^1(0,T;\mathcal{V}')}$ *such that*

$$\|\theta_\mu\|_{H^1(0,T;H)\cap L^\infty(0,T;V)} + \|[\chi_{1,\mu}, \chi_{2,\mu}]\|_{(H^1(0,T;H))^2} + \|u_\mu\|_{H^1(0,T;\mathcal{V})} \le c_4. \tag{6.1}$$

In particular, c_4 *is independent of* μ.

Proof. This argument is just sketched here since it is very close to that of Lemma 5.3. Indeed, the above-stated result represents the continuous version of the stability estimates for the discrete scheme. Namely, the key a priori estimate consists in taking the sum between (3.1) with $\varphi := \theta_t/\theta_*$, the time derivative of (3.4) multiplied by $[\chi_{1,\mu,t}, \chi_{2,\mu,t}]$, and the time derivative of (3.2) with $v = u_{\mu,t}$. We shall remark that the latter choices of test functions are not admissible at the present stage. However, the above calculation is to be intended at an appropriate approximation level (for instance, that of the discrete scheme). We prefer to skip the details of this discussion for the sake of clarity. Next, taking the integral on $(0,t)$ for some $t \in (0,T]$ of the upcoming relation, the assertion of Lemma 6.1 follows along the same lines as the proof of Lemma 5.3. ◻

6.3. Passage to the limit. Let us now finally turn to the proof of Theorem 4.2. Precisely, we prove the following.

LEMMA 6.2. *Under the assumptions* (A1)–(A3)*, let* $(\theta_\mu, \chi_{1,\mu}, \chi_{2,\mu}, u_\mu)$ *be the solution to Problem* P_μ*. Then the following convergences hold:*

$$\theta_\mu \longrightarrow \theta \quad \textit{weakly star in } H^1(0,T;H) \cap L^\infty(0,T;V) \text{ and strongly in } C([0,T];H), \tag{6.2}$$

$$\chi_{j,\mu} \longrightarrow \chi_j \quad \textit{weakly in } H^1(0,T;H) \text{ and strongly in } C([0,T];H), \quad j = 1,2, \tag{6.3}$$

$$u_\mu \longrightarrow u \quad \textit{weakly in} \ \ H^1(0,T;\mathcal{V}) \ \textit{and strongly in} \ \ C([0,T];\mathcal{V}), \tag{6.4}$$

where (θ,χ_1,χ_2,u) *is the unique solution to* P_0.

Proof. This argument is very close to that of Lemma 5.4. Thanks to Lemma 6.1 we are in the position of finding a quadruple (θ,χ_1,χ_2,u) such that, possibly taking non-relabeled subsequences and owing to well-known compactness results, the following convergences hold:

$$\theta_\mu \longrightarrow \theta \quad \text{weakly star in} \ \ H^1(0,T;H)\cap L^\infty(0,T;V) \ \text{and strongly in} \ \ C([0,T];H), \tag{6.5}$$

$$\chi_{j,\mu} \longrightarrow \chi_j \quad \text{weakly in} \ \ H^1(0,T;H), \quad j=1,2, \tag{6.6}$$

$$u_\mu \longrightarrow u \quad \text{weakly in} \ \ H^1(0,T;\mathcal{V}). \tag{6.7}$$

We now turn to the proof of a direct Cauchy argument. To this aim, we fix two parameters μ_n,μ_m of the extracted subsequence and denote by $(\theta_r,\chi_{1,r},\chi_{2,r},u_r)$ the solution to Problem P_{μ_r} for $r=n,m$ (σ_n, σ_m are defined accordingly). Next, we take the difference between (3.4) written for μ_1 and the same relation for μ_2, multiply it by $(\chi_{1,n}-\chi_{1,m},\chi_{2,n}-\chi_{2,m})$, and integrate in space. We readily obtain that, for almost all $t\in(0,T)$,

$$\begin{aligned}\gamma\sum_{j=1}^{2}\|(\chi_{j,n}-\chi_{j,m})(t)\|^2 &\le \frac{\ell}{\theta_*}\|(\theta_n-\theta_m)(t)\|\,\|(\chi_{1,n}-\chi_{1,m})(t)\|\\ &- \big(\beta:(\sigma_n-\sigma_m)(t),(\chi_{2,n}-\chi_{2,m})(t)\big)\\ &+\sum_{j=1}^{2}\|(\mu_n\chi_{j,n,t}-\mu_m\chi_{j,m,t})(t)\|\,\|(\chi_{j,n}-\chi_{j,m})(t)\|.\end{aligned} \tag{6.8}$$

Once again we readily compute that

$$\|\mathbb{A}^{\frac12}\varepsilon(u_n-u_m)\|^2 = -(\mathbb{A}^{\frac12}\varepsilon(u_n-u_m):\mathbb{A}^{\frac12}\beta,\chi_{2,n}-\chi_{2,m}) \quad \text{a.e. in} \ \ (0,T), \tag{6.9}$$

$$\begin{aligned}&-(\sigma_n-\sigma_m):\beta(\chi_{2,n}-\chi_{2,m})\\ &\quad= -\mathbb{A}^{\frac12}\varepsilon(u_n-u_m):\mathbb{A}^{\frac12}\beta(\chi_{2,n}-\chi_{2,m}) - |\mathbb{A}^{\frac12}\beta|^2(\chi_{2,n}-\chi_{2,m})^2 \quad \text{a.e. in} \ \ Q.\end{aligned} \tag{6.10}$$

Finally, taking the sum between (6.8) and (6.9) and using (6.11), we readily obtain that

$$\begin{aligned}&\gamma\|(\chi_{1,n}-\chi_{1,m})(t)\|^2 + \big(\gamma+|\mathbb{A}^{\frac12}\beta|^2-\rho|\mathbb{A}^{\frac12}\beta|^2\big)\|(\chi_{2,n}-\chi_{2,m})(t)\|^2\\ &\quad+\frac{\rho-1}{\rho}\|\mathbb{A}^{\frac12}\varepsilon(u_n-u_m)(t)\|^2 \le \frac{\ell}{\theta_*}\|(\theta_n-\theta_m)(t)\|\,\|(\chi_{1,n}-\chi_{1,m})(t)\|\\ &\quad+\sum_{j=1}^{2}\|(\mu_n\chi_{j,n,t}-\mu_m\chi_{j,m,t})(t)\|\,\|(\chi_{j,n}-\chi_{j,m})(t)\|,\end{aligned}$$

where once again ρ is such that $1<\rho<(\gamma+|\mathbb{A}^{\frac12}\beta|^2)/|\mathbb{A}^{\frac12}\beta|^2$. Now, it suffices to recall (5.28) and (6.5) in order to infer the strong convergences

$$\begin{aligned}\chi_{j,\mu} &\longrightarrow \chi_j \quad \text{strongly in} \ \ C([0,T];H) \ \ \text{for} \ \ j=1,2,\\ u_\mu &\longrightarrow u \quad \text{strongly in} \ \ C([0,T];\mathcal{V}).\end{aligned}$$

Moving from the above positions, the proof of this lemma may be concluded exactly as that of Lemma 5.4. ∎

6.4. Error control. The Cauchy argument devised in the latter subsection may be used in order to achieve some quantitative control on the distance between the dissipative and the nondissipative regimes. In particular, we have the following.

LEMMA 6.3. *Under the assumptions* (A1)–(A3)*, let* $(\theta_\mu, \chi_{1,\mu}, \chi_{2,\mu}, u_\mu)$ *and* $(\theta, \chi_1, \chi_2, u)$ *denote the unique solutions to Problems* P_μ *and* P_0*, respectively. Then, there exists a constant* c_5 *with the same dependences of* c_2 *(in particular independent of* μ*) such that*

$$\|\theta - \theta_\mu\|_{L^2(0,T;H)} + \sup_{t\in[0,T]} \left\| \int_0^t \nabla(\theta - \theta_\mu) \right\| + \sup_{t\in[0,T]} \left\| \int_0^t (\theta - \theta_\mu) \right\|_{L^2(\Gamma)}$$
$$+ \sum_{j=1}^2 \|\chi_j - \chi_{j,\mu}\|_{L^2(0,T;H)} + \|u - u_\mu\|_{L^2(0,T;\mathcal{V})} \le c_5\sqrt{\mu}.$$

The proof of Lemma 6.3 follows along the same lines as subsections 5.1 and 6.3 and is therefore omitted.

6.5. Discretization. Let us comment here the possible discretization of Problem P_0. First of all we observe that the limit procedure of section 5 is completely independent of μ. On the other hand, the positivity of μ is exploited in order to implement the contraction argument. Namely, one could choose $\mu = \mu(\tau) > 0$ such that $\lim_{\tau\to 0^+} \mu(\tau) = 0$ and prove that the resulting discrete solution exists and converges indeed to a solution to Problem P_0.

It is, however, remarkable that we would also be in the position of providing a variable time-step discretization scheme for Problem P_0 as well. Namely, we could directly work at the nondissipative level $\mu = 0$ and prove Lemmas 5.1, 5.3, and 5.4 (and hence Theorem 4.1) directly for Problem D_0. On the other hand, we prefer to analyze here the discretization of the dissipative problem because the well-posedness proof for Problem D_0 relies on some nonconstructive technique (Schauder fixed point) and hence shows a merely theoretical interest (while the scheme for problem D_μ is effectively computable). Finally, we are interested in establishing the asymptotic connection within the dissipative and the nondissipative models at all levels, namely, the continuous and the discrete ones.

As for the discretization error estimate (5.48) one could actually prove that an analogous bound holds true in the case $\mu = 0$. In particular, in the latter nondissipative case we will be forced to replace τ with $\sqrt{\tau}$, i.e., we reduce ourselves to a suboptimal convergence rate.

7. Long-time behavior. Let us now turn to the proof of Theorem 4.3. In particular, let us recall that the dissipation parameter μ is set to be zero throughout this section. Namely, we will carry out the long-time behavior analysis in the nondissipative regime. Owing to Theorem 4.1 it is a standard matter to check for the existence and uniqueness of a quadruple $(\theta, \chi_1, \chi_2, u)$ such that, for each $T \in (0, +\infty)$, one has that $\theta \in H^1(0,T;H) \cap L^\infty(0,T;V)$, $[\chi_1, \chi_2] \in (H^1(0,T;H))^2$, $u \in H^1(0,T;\mathcal{V})$, fulfilling conditions (3.1)–(3.5) with $\mu = 0$. We proceed by establishing some lemmas.

LEMMA 7.1. *Under the assumptions* (A2)–(A4)*, there exists a positive constant* c_6 *depending on* c_s, k, h, θ_e, Γ, γ, $\mathbb{A}^{\frac{1}{2}}\beta$, $c_{\mathcal{V}}$, θ_*, θ^0, $[\chi_1^0, \chi_2^0]$, $\|F\|_{L^2(0,+\infty;H)}$, *and* $\|G\|_{H^1(0,+\infty;\mathcal{V}')}$ *such that*

$$\int_0^t \left(\|\theta_t\|^2 + \sum_{j=1}^2 \|\chi_{j,t}\|^2 + \|u_t\|_{\mathcal{V}}^2 \right) + \|\theta(t) - \theta_e\|_V^2 \le c_6 \qquad \forall t > 0. \tag{7.1}$$

We do not provide here a detailed proof of the latter estimate. Indeed, the argument of Lemma 6.1 (together with the long-time assumption (A4)) may be easily adapted to ensure the validity of (7.1).

A first consequence of Lemma 7.1 is that the set

$$\big\{(\theta(t),\chi_1(t),\chi_2(t),u(t)),\ t>0\big\}\ \text{ is bounded in }\ V\times H\times H\times\mathcal{V}.$$

Therefore, there exists a sequence $t_n\to+\infty$ and a quadruple $(\theta_\infty,\chi_{1,\infty},\chi_{2,\infty},u_\infty)$ such that

$$\begin{aligned}\theta(t_n)&\longrightarrow\theta_\infty &&\text{strongly in }\ H,\\ \chi_j(t_n)&\longrightarrow\chi_{j,\infty} &&\text{weakly in }\ H,\ j=1,2,\\ u(t_n)&\longrightarrow u_\infty &&\text{weakly in }\ \mathcal{V}.\end{aligned}$$

Indeed, by observing that relations (3.2)–(3.4) are actually fulfilled everywhere in time and almost everywhere in Ω and arguing as in the proof of Lemma 6.2, we easily check that the direct Cauchy argument devised above entails that the latter convergences are strong. In particular, the set $\omega(\theta,\chi_1,\chi_2,u)$ is nonempty.

Consider now any $(\theta_\infty,\chi_{1,\infty},\chi_{2,\infty},u_\infty)\in\omega(\theta,\chi_1,\chi_2,u)$. Hence, there is a sequence $\{t_n\}$ of positive real numbers such that $t_n\to+\infty$ and

$$(\theta(t_n),\chi_1(t_n),\chi_2(t_n),u(t_n))\longrightarrow(\theta_\infty,\chi_{1,\infty},\chi_{2,\infty},u_\infty)\quad\text{in }\ H\times H\times H\times\mathcal{V}. \tag{7.2}$$

For n and $t\ge 0$, we define

$$\theta_n(t):=\theta(t_n+t),\quad \chi_{j,n}(t):=\chi_j(t_n+t),\quad j=1,2,\quad u_n(t):=u(t_n+t).$$

We can introduce a pair of auxiliary functions $[\xi_{1,n},\xi_{2,n}]$ such that the functions $\theta_n,\chi_{1,n},\ \chi_{2,n},u_n,\sigma_n,\xi_{1,n},\xi_{2,n}$ solve relations

$$\gamma\begin{pmatrix}\chi_{1,n}\\ \chi_{2,n}\end{pmatrix}+\begin{pmatrix}\frac{\ell}{\theta_*}(\theta_n-\theta_*)\\ \sigma_n:\beta\end{pmatrix}+\begin{pmatrix}\xi_{1,n}\\ \xi_{2,n}\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}\qquad\text{a.e. in }\ \Omega\times(0,T), \tag{7.3}$$

$$\begin{pmatrix}\xi_{1,n}\\ \xi_{2,n}\end{pmatrix}\in\partial I_{\mathcal{K}}(\chi_1,\chi_2)\qquad\text{a.e. in }\ \Omega\times(0,T), \tag{7.4}$$

as well as the relations (3.1)–(3.3), for all $T\in(0,+\infty)$. However, note that in (3.1)–(3.2) F and G have to be replaced by $F_n:=F(\cdot+t_n)$ and $G_n:=G(\cdot+t_n)$, respectively. We also point out the initial condition $\theta_n(\cdot,0)=\theta(\cdot,t_n)$ almost everywhere in Ω.

Owing to Lemma 7.1 it is not difficult to prove some estimates for the functions $\theta_n,\chi_{1,n},\ \chi_{2,n},\ u_n,\xi_{1,n}$, and $\xi_{2,n}$ which are uniform with respect to n. The proof of the next result is omitted since it is analogous to the proof of Lemma 5.3.

LEMMA 7.2. *Let $T>0$. Under the above assumptions, letting $\xi_{1,n}$ and $\xi_{2,n}$ be as in* (7.3)–(7.4)*, there exists a positive constant c_7 with the same dependencies of c_6 such that*

$$\begin{aligned}&\|\theta_n\|_{H^1(0,T;H)\cap L^\infty(0,T;V)}+\|u_n\|_{H^1(0,T;\mathcal{V})}\\ &\quad+\sum_{j=1}^{2}\|\chi_{j,n}\|_{H^1(0,T;H)}+\sum_{j=1}^{2}\|\xi_{j,n}\|_{H^1(0,T;H)}\le c_7.\end{aligned} \tag{7.5}$$

Another consequence of Lemma 7.1 is to allow the identification of the limit of $\theta_n, \chi_{1,n}, \chi_{2,n}$, and u_n as $n \to +\infty$. More precisely, we have the following.

LEMMA 7.3. *Under the above assumptions, for every $T > 0$ there holds*

$$\theta_n \longrightarrow \theta_\infty \quad \text{strongly in } H^1(0,T;H), \tag{7.6}$$

$$\chi_{j,n} \longrightarrow \chi_{j,\infty} \quad \text{strongly in } H^1(0,T;H), \quad j=1,2, \tag{7.7}$$

$$u_n \longrightarrow u_\infty \quad \text{strongly in } H^1(0,T;\mathcal{V}). \tag{7.8}$$

Proof. Taking into account (7.1) it is straightforward to check that

$$\int_0^T \|\theta_{n,t}\|^2 = \int_{t_n}^{t_n+T} \|\theta_t\|^2 \longrightarrow 0 \quad \text{as } n \longrightarrow +\infty. \tag{7.9}$$

An analogous computation applies to $\chi_{1,n,t}, \chi_{2,n,t}$, and $u_{n,t}$, as well. Hence, we easily deduce that

$$\begin{aligned}\|\theta_n(t) - \theta_\infty\| &\le \|\theta_n(t) - \theta_n(0)\| + \|\theta(t_n) - \theta_\infty\| \\ &\le T^{1/2}\|\theta_{n,t}\|_{L^2(0,T;H)} + \|\theta(t_n) - \theta_\infty\|.\end{aligned}$$

Owing to (7.2) and (7.9), the right-hand side of the latter inequality goes to zero as $n \to +\infty$. Hence, (7.6) is proved. A similar argument ensures that (7.7)–(7.8) hold true. $\square$

After these preliminaries, we may prove Theorem 4.3 by passing to the limit as $n \to +\infty$ in (3.1)–(3.3) for the quadruple $(\theta_n, u_n, \chi_{1,n}, \chi_{2,n})$ and data $[F_n, G_n]$ and in relations (7.3)–(7.4). Thanks to the above lemmas and well-known compactness results we find a subsequence (not relabeled) of $\theta_n, \chi_{1,n}, \chi_{2,n}$, and u_n and a pair $(\xi_{1,\infty}, \xi_{2,\infty})$ such that, in addition to (7.6)–(7.8), the following convergences hold:

$$\theta_n \longrightarrow \theta_\infty \quad \text{weakly star in } L^\infty(0,T;V), \tag{7.10}$$

$$\xi_{j,n} \longrightarrow \xi_{j,\infty} \quad \text{weakly in } H^1(0,T;H), \quad j=1,2. \tag{7.11}$$

The convergences proved above and (A4) are sufficient in order to pass to the limit in (3.1)–(3.3). In particular, it turns out that $\theta_\infty = \theta_e$ almost everywhere in Ω. As far as relations (7.3)–(7.4) are concerned, we observe that we also have

$$\begin{pmatrix} \frac{\ell}{\theta_*}(\theta_n - \theta_*) \\ \sigma_n : \beta \end{pmatrix} \longrightarrow \begin{pmatrix} \frac{\ell}{\theta_*}(\theta_\infty - \theta_*) \\ \sigma_\infty : \beta \end{pmatrix} \quad \text{strongly in } (C([0,T];H))^2.$$

Hence, we just need to identify the limit of $[\xi_{1,n}, \xi_{2,n}]$. Indeed, from (7.7) and (7.11), one easily infers that

$$(\xi_{i,n}, \chi_{i,n}) \longrightarrow (\xi_{i,\infty}, \chi_{i,\infty}) \quad \text{a.e. in } (0,T) \text{ for } i=1,2.$$

The classical theory of maximal monotone operators (see, e.g., [7, Prop. 2.5, p. 27]) then entails that $[\xi_{1,\infty}, \xi_{2,\infty}] \in \partial I_K(\chi_{1,\infty}, \chi_{2,\infty})$ almost everywhere in Ω, and we have finally proved (4.2)–(4.5). In order to conclude the proof of Theorem 4.3 we provide the following stronger result.

LEMMA 7.4. *Let the external temperatures $\theta_{e,1}, \theta_{e,2}$ be given and let $(\chi_{1,i}, \chi_{2,i}, u_i) \in H \times H \times \mathcal{V}$ fulfill*

$$\begin{aligned} a(u_i, v) + \big(\mathbb{A}\beta\chi_{2,i}, \varepsilon(v)\big) = 0 \quad \forall v \in \mathcal{V}, \\ \mathbb{A}\big(\varepsilon(u_i) + \beta\chi_{2,i}\big) = \sigma_i \quad \textit{a.e. in } \Omega, \\ \gamma\begin{pmatrix}\chi_{1,i} \\ \chi_{2,i}\end{pmatrix} + \partial I_K \begin{pmatrix}\chi_{1,i} \\ \chi_{2,i}\end{pmatrix} \ni \begin{pmatrix} \frac{\ell}{\theta_*}(\theta_* - \theta_{e,i}) \\ -\sigma_i : \beta \end{pmatrix} \quad \textit{a.e. in } \Omega \end{aligned}$$

for $i=1,2$. *Then, there exists a positive constant* c_8 *depending just on* ℓ, θ_*, γ, $\mathbb{A}^{\frac{1}{2}}\beta$, *and* $c_{\mathcal{V}}$ *such that*

$$\sum_{j=1}^{2}\|\chi_{j,1}-\chi_{j,2}\|+\|u_1-u_2\|_{\mathcal{V}}\le c_8\,|\theta_{e,1}-\theta_{e,2}|. \tag{7.12}$$

Once again the proof of Lemma 7.4 may be easily obtained by adapting the argument of subsection 5.1. A consequence of the above continuous dependence result is that, since $(\theta_\infty,\chi_{1,\infty},\chi_{2,\infty},u_\infty)$ are uniquely determined, we readily check that the ω-limit set reduces to a point and the whole trajectory $(\theta(t),\chi_1(t),\chi_2(t),u(t))$ converges to $(\theta_\infty,\chi_{1,\infty},\chi_{2,\infty},u_\infty)$ as $t\to+\infty$. In particular, this concludes the proof of Theorem 4.3.

8. Lower bound for the temperature. Let us now turn to the proof of Theorem 4.4. This argument is very close to that of [11] and will be just sketched, referring to the latter paper for details. We will start by checking (4.7) in the dissipative case $\mu>0$. In this situation we claim that

$$\left(\gamma\chi_1+\frac{\ell}{\theta_*}(\theta-\theta_*)\right)^+\ge-\mu\chi_{1,t}\quad\text{a.e. in }\ Q, \tag{8.1}$$

where we used the standard notation for the positive part. Indeed, $\chi_{1,t}=0$ almost everywhere on the measurable set $\{\chi_1=1\}$. On the other hand, for almost every $(x,t)\in\{\chi_1<1\}$, one readily checks from (3.4) that (see [11])

$$\left(\mu\chi_{1,t}+\gamma\chi_1+\frac{\ell}{\theta_*}(\theta-\theta_*)\right)(x,t)\ge 0.$$

Let us now consider

$$\underline{\theta}:=\inf\{\inf\theta_0,\,\theta_e,\,\theta_*,\,\theta_d\}\in\mathbb{R}$$

(the case $\underline{\theta}=-\infty$ being obvious), choose $\varphi=-(\theta-\underline{\theta})^-\in V$ in (3.1), and take the integral on $(0,t)$ for $t\in(0,T]$, obtaining

$$\begin{aligned}\frac{c_s}{2}\|(\theta-\underline{\theta})^-(t)\|^2+k\int_0^t\|\nabla((\theta-\underline{\theta})^-)\|^2-h\int_0^t(\theta-\theta_e,(\theta-\underline{\theta})^-)_\Gamma\\ =-\int_0^t\langle F,(\theta-\underline{\theta})^-\rangle-\ell\int_0^t(\chi_{1,t},(\theta-\underline{\theta})^-).\end{aligned}\tag{8.2}$$

Owing to (A5) and (8.1) one gets that the above right-hand side may be controlled as follows:

$$\begin{aligned}&-\int_0^t\langle F,(\theta-\underline{\theta})^-\rangle-\ell\int_0^t(\chi_{1,t},(\theta-\underline{\theta})^-)\\ &\le\frac{\ell}{\mu}\int_0^t\left(\left(\gamma\chi_1+\frac{\ell}{\theta_*}(\theta-\theta_*)\right)^+,(\theta-\underline{\theta})^-\right)\\ &\le\frac{\ell}{\mu}\int_0^t\left(\left(\gamma+\frac{\ell}{\theta_*}(\underline{\theta}-\theta_*)\right)^+,(\theta-\underline{\theta})^-\right)=0,\end{aligned}$$

since $\underline{\theta}\le\theta_d$ and (4.6) holds . Hence, looking back to (8.2) and considering that $\underline{\theta}\le\theta_e$ as well, we readily check that $\theta\ge\underline{\theta}$ almost everywhere in Q. The proof of the lower bound for the temperature in the nondissipative case $\mu=0$ simply follows by approximation owing to (6.5).

Acknowledgments. The kind hospitality of Laboratoire de Mécanique et Génie Civil de Montpellier is acknowledged. Moreover, the author is gratefully indebted to the referee for some interesting remarks on the model.

REFERENCES

[1] M. Achenbach and I. Müller, *Models for shape memory,* J. Phys. C4 Suppl., 12 (1982), pp. 163–167.

[2] S. Antman, J. L. Ericksen, D. Kinderlehrer, and I. Müller, eds., *Metastability and Incompletely Posed Problems,* Springer-Verlag, New York, 1987.

[3] F. Auricchio, R. L. Taylor, and J. Lubliner, *Shape-memory alloys: Macromodelling and numerical simulations of the superelastic behavior,* Comput. Methods Appl. Mech. Engrg., 146 (1997), pp. 281–312.

[4] J. M. Ball and R. D. James, *Fine phase mixtures as minimizers of energy,* Arch. Ration. Mech. Anal., 100 (1987), pp. 13–52.

[5] J. M. Ball and D. Schryvers, *The formation of macrotwins in NiAl martensite,* in Mechanics of Martensitic Phase Transformation in Solids (Hong Kong, 2001), Solid Mech. Appl. 101, Kluwer Academic, Dordrecht, 2002, pp. 27–36.

[6] C. Bouvet, S. Calloch, and Ch. Lexcellent, *A phenomenological model for pseudoelasticity of shape memory alloys under multiaxial proportional and nonproportional loadings,* Eur. J. Mech. A Solids, 23 (2004), pp. 37–61.

[7] H. Brezis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert,* North Holland Math. Stud. 5, North-Holland, Amsterdam, 1973.

[8] P. Colli, *An existence result for a thermo-mechanical model of shape memory alloys,* Adv. Math. Sci. Appl., 1 (1992), pp. 83–97.

[9] P. Colli, M. Frémond, and A. Visintin, *Thermo-mechanical evolution of shape memory alloys,* Quart. Appl. Math., 48 (1990), pp. 31–47.

[10] P. Colli and J. Sprekels, *Global existence for a three-dimensional model for the thermodynamical evolution of shape memory alloys,* Nonlinear Anal., 18 (1992), pp. 873–888.

[11] P. Colli and J. Sprekels, *Positivity of temperature in the general Frémond model for shape memory alloys,* Contin. Mech. Thermodyn., 5 (1993), pp. 255–264.

[12] P. Colli and J. Sprekels, *Global solution to the full one-dimensional Frémond model for shape memory alloys,* Math. Methods Appl. Sci., 18 (1995), pp. 371–385

[13] Y. Y. Chu and L. C. Zhao, eds., *Shape memory materials and their applications,* in Materials Science Forum 394–395, Proceedings of the International Conference on Shape Memory and Superelastic Technologies and Shape Memory Materials (SMST-SMM 2001), Kunming, China, 2001, Trans Tech Publications, Switzerland, 2002.

[14] A. Chrysochoos, M. Löbel, and O. Maisonneuve, *Couplages thermo-mécaniques du comportament pseudoĺastique d'alliages Cu-Zn-Al et Ni-Ti,* C. R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers Sci. Terre, 320 (1994), pp. 217–223.

[15] S. Demoulini, *Weak solutions for a class of nonlinear systems of viscoelasticity,* Arch. Ration. Mech. Anal., 155 (2000), pp. 299–334.

[16] T. Duerig and A. R. Pelton eds., *Proceedings of the International Conference on Shape Memory and Superelastic Technologies,* Asilomar, CA, ASM International, 2003.

[17] G. Duvaut and J.-L. Lions, *Inequalities in Mechanics and Physics,* Springer-Verlag, Berlin, 1976.

[18] F. Falk, *Martensitic domain boundaries in shape-memory alloys as solitary waves,* J. Phys., C4 Suppl., 12 (1982), pp. 3–15.

[19] M. Frémond, *Matériaux à mémoire de forme,* C. R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers Sci. Terre, 304 (1987), pp. 239–244.

[20] M. Frémond, *Non-smooth Thermomechanics,* Springer-Verlag, Berlin, 2002.

[21] M. Frémond and S. Miyazaki, *Shape Memory Alloys,* CISM Courses and Lectures 351, Springer-Verlag, New York, 1996.

[22] G. Friesecke and G. Dolzmann, *Implicit time discretization and global existence for a quasi-linear evolution equation with nonconvex energy,* SIAM J. Math. Anal., 28 (1997), pp. 363–380.

[23] P. Germain, *Mécanique des milieux continues,* Masson, Paris, 1973.

[24] K. H. Hoffmann, M. Niezgódka and S. Zheng, *Existence and uniqueness to an extended model of the dynamical developments in shape memory alloys,* Nonlinear Anal., 15 (1990), pp. 977–990.

[25] T. J. R Hughes, T. Kato, and J. E. Marsden, *Well-posed quasi-linear second-order hyper-*

bolic systems with applications to nonlinear elastodynamics and general relativity, Arch. Ration. Mech. Anal., 63 (1976), pp. 273–294.

[26] S. Jiang and R. Racke, *Evolution Equations in Thermoelasticity,* Chapman Hall/CRC Monogr. Surv. Pure Appl. Math. 112, Chapman & Hall/CRC, Boca Raton, FL, 2000.

[27] J.-J. Moreau, *Sur les lois de frottement, de viscosité et plasticité,* C. R. Acad. Sci. Paris Sér. II Méc. Phys. Chim. Sci. Univers Sci. Terre, 271 (1970), pp. 608–611.

[28] R. Nochetto, G. Savaré, and C. Verdi, *Error control of nonlinear evolution equations,* C. R. Acad. Sci. Paris Sér. I Math., 326 (1998), pp. 1437–1442.

[29] R. Nochetto, G. Savaré, and C. Verdi, *A posteriori error estimates for variable time-step discretization of nonlinear evolution equations,* Comm. Pure Appl. Math., 53 (2000), pp. 525–589.

[30] A. R. Pelton, D. Hodgson, and T. Duerig, eds., *Proceedings of the Second International Conference on Shape Memory and Superelastic Technologies,* Asilomar, CA, SMST Publication, Menlo Park, CA, 1997.

[31] R. Peyroux, A. Chrysochoos, Ch. Licht, and M. Löbel, *Phenomenological constitutive equations for numerical simulations of SMA's structures. Effect of thermomechanical couplings,* J. Phys. C4 Suppl., 6 (1996), pp. 347–356.

[32] H. Pham, *Analyse thermomécanique d'un alliage à mémoire de forme de type Cu-Zn-Al,* Ph.D. thesis, Université des Sciences et des Techniques du Languedoc, Montpellier, 1994.

[33] B. Raniecki, Ch. Lexcellent, and K. Tanaka, *Thermodynamic models of pseudoelastic behaviour of shape memory alloys,* Arch. Mech. (Arch. Mech. Stos.), 44 (1992), pp. 261–284.

[34] U. Stefanelli, *Error control of a nonlinear evolution problem related to phase transitions,* Numer. Funct. Anal. Optim., 20 (1999), pp. 585–608.

[35] U. Stefanelli, *Analysis of a variable time-step discretization of the three-dimensional Frémond model for shape memory alloys,* Math. Comp., 71 (2002), pp. 1413–1435.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 156–188

EXISTENCE AND UNIQUENESS OF THE MILD SOLUTION FOR THE 1D VLASOV–POISSON INITIAL-BOUNDARY VALUE PROBLEM*

M. BOSTAN†

Abstract. We prove the existence and uniqueness of the mild solution for the 1D Vlasov–Poisson system with initial-boundary conditions by using iterated approximations. The same arguments yield existence and uniqueness for the free space or space periodic system. The major difficulty is the treatment of the boundary conditions. The main idea consists of splitting the velocities range by introducing critical velocities corresponding to each boundary. One of the crucial points is to estimate the critical velocity change in term of relative field. A result concerning the continuity of the mild solution upon the initial-boundary conditions is presented as well.

Key words. Vlasov–Poisson equations, Vlasov–Maxwell equations, weak/mild formulation

AMS subject classifications. 35Q99, 35L50

DOI. 10.1137/S0036141003434649

1. Introduction.

Many studies in the physics of charged particles are modeled by kinetic equations (Vlasov, Boltzmann, etc.) coupled with electromagnetic equations (Poisson, Maxwell). A few application domains are semiconductors, particle accelerators, electron guns, etc.

Various results have been obtained for the free space systems. Weak solutions for the Vlasov–Poisson system were constructed by Arseneev [1] and Horst and Hunze [16]. The existence of classical solutions has been studied in two and three dimensions by Ukai and Okabe [21], Horst [15], Batt [2], and Pfaffelmoser [18]. Classical solutions for the Vlasov–Poisson equations with small initial data have been constructed by Bardos and Degond [3]. The propagation of the velocity moments for the Vlasov–Poisson system in three dimensions has been studied by Lions and Perthame in [17]. They prove also a uniqueness result under a Lipschitz continuity assumption on the initial data. Another uniqueness result has been obtained by Robert for bounded, compactly supported initial data, [20]. A uniqueness result for solutions with bounded variation was obtained by Guo, Shu, and Zhou [14].

The existence of weak solutions for the Vlasov–Maxwell system in three dimensions was shown by DiPerna and Lions [9]. The relativistic Vlasov–Maxwell system was studied by Glassey and Schaeffer [10]. In one dimension, the existence and uniqueness have been obtained by Cooper and Klimas [7].

The boundary value problem has been studied as well. The existence of weak solutions for the Vlasov–Poisson initial-boundary value problem in three dimensions is a result of Abdallah [4]. The existence of weak solutions for the three-dimensional Vlasov–Maxwell initial-boundary value problem has been analyzed by Guo [12]. The stationary one-dimensional Vlasov–Poisson system has been studied by Greengard and Raviart [11]. An asymptotic analysis of the Vlasov–Poisson system has been performed by Degond and Raviart [8] in the case of the plane diode. The stationary Vlasov–Maxwell system in three dimensions was analyzed by Poupaud [19]. The

*Received by the editors September 18, 2003; accepted for publication (in revised form) October 15, 2004; published electronically August 31, 2005.
http://www.siam.org/journals/sima/37-1/43464.html

†Université de Franche-Comté, 16 route de Gray F-25030 Besançon Cedex, France (mbostan@math.univ-fcomte.fr).

regularity of the solutions for the Vlasov–Maxwell system in a half line has been studied by Guo [13]. Results for the time periodic case can be found in [6] for the Vlasov–Poisson system and in [5] for the Vlasov–Maxwell system.

In this paper we study the existence and uniqueness of the mild solution for the Vlasov–Poisson initial-boundary value problem in one dimension:

$$\begin{aligned}
\partial_t f + v\cdot\partial_x f + E(t,x)\cdot\partial_v f = 0,\ (t,x,v) \in]0,T[\times]0,1[\times\mathbb{R}_v,\\
f(t=0,x,v) = f_0(x,v),\ (x,v)\in]0,1[\times\mathbb{R}_v,\\
f(t,x,v) = g(t,x,v),\ (t,x,v)\in]0,T[\times\Sigma^-,\\
E(t,x) = -\partial_x U,\ \partial_x E = -\partial_x^2 U = \rho(t,x) := \int_{\mathbb{R}_v} f(t,x,v)dv,\ (t,x)\in]0,T[\times]0,1[,\\
U(t,x=0) = U_0(t),\ U(t,x=1) = U_1(t),\ t\in]0,T[.
\end{aligned}$$

The function $f(t,x,v)$ represents the particles distribution depending on the time t, the position x, and the velocity v. The electric field $E(t,x)$ derives from an electrostatic potential U verifying the Poisson equation with the charge density $\rho(t,x) := \int_{\mathbb{R}_v} f(t,x,v)dv$. Here Σ^- is the subset of $\Sigma = \{0,1\}\times\mathbb{R}_v$ corresponding to the incoming velocities:

$$\Sigma^- = \{(0,v)\mid v>0\}\cup\{(1,v)\mid v<0\} = \Sigma_0^-\cup\Sigma_1^-.$$

Similarly, we define also $\Sigma^+ = \{(0,v)\mid v<0\}\cup\{(1,v)\mid v>0\} = \Sigma_0^+\cup\Sigma_1^+$, which corresponds to the outgoing velocities and $\Sigma^0 = \{(0,0),(1,0)\}$. With the notation $g|_{]0,T[\times\Sigma_0^-} = g_0, g|_{]0,T[\times\Sigma_1^-} = g_1$, the boundary condition is written

$$f(t,x=0,v>0) = g_0(t,v>0),\quad f(t,x=1,v<0) = g_1(t,v<0),\quad t\in]0,T[.$$

The existence of a weak solution for the Vlasov–Poisson initial-boundary value problem has been obtained in previous works; in [4] weak solutions of finite total (kinetic and electric) energy are constructed in dimension d, $d\le 3$, by assuming initial-boundary conditions of finite kinetic, respectively flux of kinetic, energy:

$$\int_0^1\int_{\mathbb{R}_v} f_0(x,v)|v|^2dxdv + \sup_{0\le t\le T}\left\{\int_{v>0} v|v|^2 g_0(t,v)dv - \int_{v<0} v|v|^2 g_1(t,v)dv\right\} < +\infty,$$

and $|v|^\lambda f_0 \in L^\infty(]0,1[\times\mathbb{R}_v), |v|^\lambda g_0 \in L^\infty(]0,T[\times\mathbb{R}_v^+), |v|^\lambda g_1 \in L^\infty(]0,T[\times\mathbb{R}_v^-)$ for some $\lambda > d+1$. The main goal of this paper is to establish the existence and uniqueness of the mild solution (or solution by characteristics) in one dimension under a less restrictive hypothesis, say, for initial-boundary conditions of finite charge. As is usual when studying coupled equations, we search the solutions as fixed points for some nonlinear application. For the 1D Vlasov–Poisson system this application is written, for example, as $\mathcal{F}: B_R(X_T)\to B_R(X_T)$, where

$$\mathcal{F}E(t,x) = \int_0^x \rho_E(t,y)dy - \int_0^1 (1-y)\rho_E(t,y)dy - U_1(t) + U_0(t),$$
$$(t,x)\in]0,T[\times]0,1[,$$

where $\rho_E(t,x) = \int_{\mathbb{R}_v} f_E(t,x,v)dv$ and f_E solves the linear Vlasov problem associated with the field E and $B_R(X_T)$ is the ball of radius R of some space X_T. Naturally, in order to construct solutions by characteristics, which are written as

$$\frac{d}{ds}X(s;t,x,v) = V(s;t,x,v), \quad \frac{d}{ds}V(s;t,x,v) = E(s, X(s;t,x,v)),$$
$$s_{in}(t,x,v) \le s \le s_{out}(t,x,v),$$

the space X_T to be considered is $L^\infty(]0,T[;W^{1,\infty}(]0,1[))$. Here s_{in}/s_{out} represent the entry/exit time of the characteristics in the domain $]0,1[$ (see the next section for exact definitions). Since by construction $\partial_x \mathcal{F}E = \rho_E$ (conforming to the Poisson equation), it is clear that $B_R(X_T)$ is preserved by $\mathcal{F}$, provided that the charge density remains uniformly bounded in $L^\infty(]0,T[\times]0,1[)$. Therefore the natural hypotheses are

$$\int_{\mathbb{R}_v} \sup_{0<x<1} f_0(x,v)dv + \int_{v>0} \sup_{0<t<T} g_0(t,v)dv + \int_{v<0} \sup_{0<t<T} g_1(t,v)dv < +\infty$$

and

$$\max\{\|f_0\|_{L^\infty(]0,1[\times\mathbb{R}_v)}, \|g_0\|_{L^\infty(]0,T[\times\mathbb{R}_v^+)}, \|g_1\|_{L^\infty(]0,T[\times\mathbb{R}_v^-)}\} < +\infty.$$

We intend to show the existence of a unique fixed point for $\mathcal{F}$ by using the iterated approximations method, which requires us to estimate $\mathcal{F}A - \mathcal{F}B$ in terms of $A - B$ for A, B different fields of X_T. This can be done by using the mild formulation of the Vlasov problem. Indeed, by using the continuity equation $\partial_t \rho_E + \partial_x j_E = 0$, $\mathcal{F}E$ can be represented also in term of the current density. Or estimate $\int_0^t j_A(s,x)ds - \int_0^t j_B(s,x)ds$ in $L^\infty(]0,1[)$ reduces to a duality calculation by taking the product by L^1 functions φ:

$$\begin{aligned}\left\langle \int_0^t (j_A(s,\cdot) - j_B(s,\cdot))ds, \varphi(\cdot) \right\rangle &= \int_0^t \int_0^1 \int_{\mathbb{R}_v} (f_A(s,x,v) - f_B(s,x,v)) v\varphi(x) ds dx dv \\ &= \int_0^t \int_{v>0} v g_0(\tau, v) \int_{X_B(s_{out}^0)}^{X_A(s_{out}^0)} \varphi(u) du d\tau dv \\ &\quad - \int_0^t \int_{v<0} v g_1(\tau, v) \int_{X_B(s_{out}^1)}^{X_A(s_{out}^1)} \varphi(u) du d\tau dv \\ &\quad + \int_0^1 \int_{\mathbb{R}_v} f_0(x,v) \int_{X_B(s_{out}^i)}^{X_A(s_{out}^i)} \varphi(u) du dx dv,\end{aligned}$$

where $s_{out}^0 = s_{out}(\tau,0,v), s_{out}^i = s_{out}(0,x,v), s_{out}^1 = s_{out}(\tau,1,v)$ represent the exit times of the characteristics (see the next sections for the exact definitions). Note that for large velocities the integrand of the left boundary term vanishes since

$$X_A(s_{out}(\tau,0,v)) = X_B(s_{out}(\tau,0,v)) = 1.$$

This suggests the definition of some critical velocities $v^0(t;\tau,0), v^1(t;\tau,0)$ such that

$$s_{out}(\tau,0,v) < t, \ X(s_{out}(\tau,0,v);\tau,0,v) = 0, \ 0 < v < v^0(t;\tau,0),$$
$$s_{out}(\tau,0,v) = t, \ 0 < X(s_{out}(\tau,0,v);\tau,0,v) < 1, \ v^0(t;\tau,0) < v < v^1(t;\tau,0),$$
$$s_{out}(\tau,0,v) < t, \ X(s_{out}(\tau,0,v);\tau,0,v) = 1, \ v > v^1(t;\tau,0).$$

Similar definitions hold for the right boundary term. One of the key points of our analysis consists of estimating the relative critical velocity. For nondecreasing fields with respect to x, we have

$$|v_A^k(t;\tau,k) - v_B^k(t;\tau,k)| \leq \int_\tau^t \|A(s) - B(s)\|_{L^\infty(]0,1[)} ds,\ k = 0, 1,$$

and finally one gets

$$\|\mathcal{F}A(t) - \mathcal{F}B(t)\|_{L^\infty(]0,1[)} \leq C \int_0^t \|A(\tau) - B(\tau)\|_{L^\infty(]0,1[)} d\tau,$$

where C depends only on the $L^\infty(]0,T[;W^{1,\infty}(]0,1[))$ norms of A, B and the initial-boundary conditions. We prove the following existence and uniqueness result.

THEOREM. *Assume that there are $n_0, h_0, h_1 : [0,+\infty[\to [0,+\infty[$ bounded nonincreasing functions such that $f_0(x,v) \leq n_0(|v|)$ $\forall (x,v) \in]0,1[\times \mathbb{R}_v$, $g_0(t,v) \leq h_0(v)$ $\forall (t,v) \in]0,T[\times \mathbb{R}_v^+$, $g_1(t,v) \leq h_1(-v)$ $\forall (t,v) \in]0,T[\times \mathbb{R}_v^-$, and*

$$\int_{\mathbb{R}_v} n_0(|v|)dv + \int_{v>0} h_0(v)dv + \int_{v<0} h_1(-v)dv < +\infty,$$

$$\max\{\|n_0\|_{L^\infty(\mathbb{R}_v^+)}, \|h_0\|_{L^\infty(\mathbb{R}_v^+)}, \|h_1\|_{L^\infty(\mathbb{R}_v^+)}, \|U_1 - U_0\|_{L^\infty(]0,T[)}\} < +\infty.$$

Then there is a unique mild solution for the 1D *Vlasov–Poisson initial-boundary value problem.*

The estimate of the relative critical velocity, which is used for the treatment of the boundary terms, relies on some comparison results for characteristics associated with nondecreasing fields, presented in section 4. This is why, when studying the Vlasov–Poisson initial-boundary value problem, we consider only one species of charged particles. All the definitions concerning the weak/mild formulations for the Vlasov or Vlasov–Poisson problem are recalled in sections 2 and 3. The main result on the existence and uniqueness of the mild solution, as well as a continuity result upon the initial-boundary conditions, is developed in section 5. The same method applies when studying the free or periodic space problem. Moreover, in this cases there are no boundary terms and thus the analysis on critical velocities not need to be used. This time the existence and uniqueness result can be obtained for general electric fields (not necessarily nondecreasing in space) which allows us to treat systems with two species of charged particles (globally neutral plasma). Statements and sketches of proofs can be found in sections 6 and 7.

2. The Vlasov equation. The equation that models the transport of charged particles is called the Vlasov equation. In one dimension, if the particles move only under the action of an electric field this equation is written

$$\partial_t f + v \cdot \partial_x f + E(t,x) \cdot \partial_v f = 0,\ (t,x,v) \in]0,T[\times]0,1[\times \mathbb{R}_v. \tag{2.1}$$

Here $E(t,x)$ is a given electric field which derives from a potential $U(t,x)$:

$$E(t,x) = -\partial_x U,\ (t,x) \in]0,T[\times]0,1[.$$

The initial-boundary conditions for the particle distribution are given by

$$f(t=0,x,v) = f_0(x,v),\ (x,v) \in]0,1[\times \mathbb{R}_v, \tag{2.2}$$

(2.3) $\quad f(t,x=0,v>0)=g_0(t,v>0),\quad f(t,x=1,v<0)=g_1(t,v<0),\quad t\in\,]0,T[.$

Now let us briefly recall the definitions of weak and mild solutions for the Vlasov problem formed by (2.1), (2.2), and (2.3).

2.1. Weak solutions for the Vlasov–Poisson problem.

DEFINITION 2.1. *Assume that* $E\in L^\infty(]0,T[\times]0,1[)$, $f_0\in L^1_{loc}(]0,1[\times\mathbb{R}_v)$, $vg_0\in L^1_{loc}(]0,T[\times\mathbb{R}_v^+)$, $vg_1\in L^1_{loc}(]0,T[\times\mathbb{R}_v^-)$. *We say that* $f\in L^1_{loc}\,(]0,T[\times]0,1[\times\mathbb{R}_v)$ *is a weak solution for the Vlasov problem* (2.1), (2.2), (2.3) *iff*

$$
\begin{aligned}
&-\int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v} f(t,x,v)\cdot(\partial_t\varphi+v\cdot\partial_x\varphi+E(t,x)\cdot\partial_v\varphi)dtdxdv\\
&\quad=\int_0^1\!\!\int_{\mathbb{R}_v} f_0(x,v)\varphi(0,x,v)dxdv+\int_0^T\!\!\int_{v>0} vg_0(t,v)\varphi(t,0,v)dtdv\\
&\qquad-\int_0^T\!\!\int_{v<0} vg_1(t,v)\varphi(t,1,v)dtdv
\end{aligned}
$$

for all test function $\varphi\in\mathcal{T}_w$, *where*

$$
\begin{aligned}
\mathcal{T}_w=\{&\varphi\in W^{1,\infty}(]0,T[\times]0,1[\times\mathbb{R}_v)\mid \varphi|_{]0,T[\times\Sigma^+}=\varphi(T,\cdot,\cdot)=0,\\
&\exists R:\mathrm{supp}(\varphi)\subset[0,T]\times[0,1]\times B_R\}.
\end{aligned}
$$

2.2. Mild solutions for the Vlasov problem. We need to consider also some special solutions of (2.1), (2.2), (2.3), which are called mild solutions or solutions by characteristics. These solutions require more regularity on the electric field and they are particular cases of weak solutions. Assume that $E\in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$, and for $(t,x,v)\in\{[0,T[\times]0,1[\times\mathbb{R}_v\}\cup\{]0,T[\times\Sigma^-\}$ let us denote by $(X(s;t,x,v),V(s;t,x,v))$ the unique solution of the ordinary differential system of equations

(2.4)

$$
\begin{aligned}
&\frac{d}{ds}X(s;t,x,v)=V(s;t,x,v),\\
&\frac{d}{ds}V(s;t,x,v)=E(s,X(s;t,x,v)),\qquad s_{in}\le s\le s_{out},
\end{aligned}
$$

which verify the conditions

$$X(s=t;t,x,v)=x,\quad V(s=t;t,x,v)=v.$$

Here $s_{in}=s_{in}(t,x,v)$ (resp., $s_{out}=s_{out}(t,x,v)$) represents the incoming (resp., outgoing) time of the characteristics in the domain $]0,1[$ defined by

(2.5) $$s_{in}(t,x,v)=\max\{0,\sup\{0\le s\le t\ :\ X(s;t,x,v)\in\{0,1\}\}\},$$

and

(2.6) $$s_{out}(t,x,v)=\min\{T,\inf\{T\ge s\ge t\ :\ X(s;t,x,v)\in\{0,1\}\}\}.$$

The total travel time through the domain (lifetime) is written $\tau(t,x,v)=s_{out}(t,x,v)-s_{in}(t,x,v)\le T$. Now we replace in Definition 2.1 the function $\partial_t\varphi+v\cdot\partial_x\varphi+E(t,x)\cdot\partial_v\varphi$ with ψ, which gives, after integration,

$$\varphi(t,x,v)=-\int_t^{s_{out}(t,x,v)}\psi(s,X(s;t,x,v),V(s;t,x,v))ds,$$

and we define the mild solution as follows.

DEFINITION 2.2. *Assume that* $E \in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$, $f_0 \in L^1_{loc}(]0,1[\times\mathbb{R}_v)$, $vg_0 \in L^1_{loc}(]0,T[\times\mathbb{R}_v^+)$, $vg_1 \in L^1_{loc}(]0,T[\times\mathbb{R}_v^-)$. *We say that* $f \in L^1_{loc}(]0,T[\times]0,1[\times\mathbb{R}_v)$ *is a mild solution for the Vlasov problem* (2.1), (2.2), (2.3) *iff*

$$\begin{aligned}
&\int_0^T\!\!\int_0^1\!\!\int_{\mathbb{R}_v} f(t,x,v)\psi(t,x,v)dtdxdv\\
&\quad= \int_0^1\!\!\int_{\mathbb{R}_v} f_0(x,v)\int_0^{s_{out}(0,x,v)} \psi(s,X(s;0,x,v),V(s;0,x,v))dsdxdv\\
&\qquad+\int_0^T\!\!\int_{v>0} vg_0(t,v)\int_t^{s_{out}(t,0,v)}\psi(s,X(s;t,0,v),V(s;t,0,v))dsdtdv\\
&\qquad-\int_0^T\!\!\int_{v<0} vg_1(t,v)\int_t^{s_{out}(t,1,v)}\psi(s,X(s;t,1,v),V(s;t,1,v))dsdtdv,
\end{aligned}$$

for all test functions $\psi \in \mathcal{T}_m$, *where*

$$\mathcal{T}_m = \{\psi \in L^\infty(]0,T[\times]0,1[\times\mathbb{R}_v) \mid \exists R>0 : \mathrm{supp}(\psi) \subset [0,T]\times[0,1]\times B_R\}.$$

In order to simplify the formulas we shall use the following notation:

$$\begin{aligned}
(X(s),V(s)) &= (X(s;t,x,v),V(s;t,x,v)), (X^0(s),V^0(s))\\
&= (X(s;t,0,v),V(s;t,0,v)),\\
(X^1(s),V^1(s)) &= (X(s;t,1,v),V(s;t,1,v)), (X^i(s),V^i(s))\\
&= (X(s;0,x,v),V(s;0,x,v))
\end{aligned}$$

and

$$\begin{aligned}
s_{in} = s_{in}(t,x,v),\quad s_{out} = s_{out}(t,x,v),\quad s^0_{out} = s_{out}(t,0,v),\\
s^1_{out} = s_{out}(t,1,v),\quad s^i_{out} = s_{out}(0,x,v).
\end{aligned}$$

Remark 2.3. It is well known that the mild solution is unique and is given by $f(t,x,v) = g_k(s_{in},V(s_{in}))$ if $s_{in}(t,x,v) > 0$, $X(s_{in}(t,x,v);t,x,v) = k, k = 0,1$, $f(t,x,v) = f_0(X(s_{in}),V(s_{in}))$ if $s_{in}(t,x,v) = 0$.

Note that every mild solution is also a weak solution. Moreover, the existence of a weak solution for the Vlasov problem with bounded initial-boundary conditions f_0, g_0, $g_1 \in L^\infty$ follows by regularization of the electric field with respect to x by convolution with $\zeta_\varepsilon(\cdot) = \frac{1}{\varepsilon}\zeta\left(\frac{\cdot}{\varepsilon}\right)$, $\zeta \in C_0^\infty$, $\mathrm{supp}(\zeta) = [-1,1], \zeta \geq 0$, $\int_{\mathbb{R}}\zeta(u)du = 1$, and by passing to the limit for $\varepsilon \searrow 0$ in the weak formulation of f^ε, the mild solution associated with $E^\varepsilon = E \star \zeta_\varepsilon$.

3. The Vlasov–Poisson system. The self-consistent electric field solves the Poisson equation

$$\partial_x E = -\partial_x^2 U = \rho(t,x) := \int_{\mathbb{R}_v} f(t,x,v)dv,\ (t,x) \in]0,T[\times]0,1[\tag{3.1}$$

with the boundary conditions

$$U(t,x=0) = U_0(t),\ \ U(t,x=1) = U_1(t),\ \ t \in]0,T[. \tag{3.2}$$

The system formed by (2.1), (2.2), (2.3), (3.1), (3.2) is called the Vlasov–Poisson initial-boundary value problem in one dimension. Obviously, the electric field is written

$$E(t,x) = \int_0^x \rho(t,y)dy - \int_0^1 (1-y)\rho(t,y)dy - U_1(t) \\ + U_0(t),\ (t,x) \in]0,T[\times]0,1[, \tag{3.3}$$

and therefore we can give the following definitions.

DEFINITION 3.1. *Assume that* $f_0 \in L^1_{loc}(]0,1[\times\mathbb{R}_v)$, $vg_0 \in L^1_{loc}(]0,T[\times\mathbb{R}_v^+)$, $vg_1 \in L^1_{loc}(]0,T[\times\mathbb{R}_v^-)$, $U_1 - U_0 \in L^\infty(]0,T[)$. *We say that* $(f,E) \in L^1(]0,T[\times]0,1[\times\mathbb{R}_v)\times L^\infty(]0,T[\times]0,1[)$ (*resp.*, $(f,E) \in L^1(]0,T[\times]0,1[\times\mathbb{R}_v)\times L^\infty(]0,T[;W^{1,\infty}(]0,1[))$) *is a weak* (*resp., mild*) *solution for the Vlasov–Poisson problem iff f is a weak* (*resp., mild*) *solution for the Vlasov problem* (2.1), (2.2), (2.3) *corresponding to the electric field* (3.3) *given by the Poisson problem.*

4. Characteristics. The main tool of our analysis is the mild formulation of the Vlasov problem. In order to estimate the charge and current densities, we need more informations about the characteristics. We present here some properties of the characteristics associated with regular, nondecreasing with respect to x, fields.

PROPOSITION 4.1. *Assume that* $E \in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$ *is nondecreasing with respect to x and that* $(X_1(s),V_1(s))$, $(X_2(s),V_2(s))$ *are two characteristics such that there is* $s_1 < s_2$ *verifying* $X_1(s_i) = X_2(s_i)$, $i = 1,2$. *Then the characteristics coincide:* $(X_1(s),V_1(s)) = (X_2(s),V_2(s))\ \forall s$.

Proof. The conclusion follows easily after multiplication of the equation $\frac{d^2}{ds^2}(X_1(s) - X_2(s)) = E(s,X_1(s)) - E(s,X_2(s))$ by $X_1(s) - X_2(s)$ and integration by parts on $[s_1,s_2]$. ☐

PROPOSITION 4.2. *Assume that* $E \in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$ *is nondecreasing with respect to x. If* $v_1 < v_2$, *then we have*

$$X(s;t,x,v_1) < X(s;t,x,v_2),\ V(s;t,x,v_1) < V(s;t,x,v_2) \\ \forall s \in]t, s_{out}(t,x,v_1)] \cap]t, s_{out}(t,x,v_2)]$$

and

$$X(s;t,x,v_1) > X(s;t,x,v_2),\ V(s;t,x,v_1) < V(s;t,x,v_2) \\ \forall s \in [s_{in}(t,x,v_1), t[\cap [s_{in}(t,x,v_2)[.$$

Proof. Suppose that there is $s \in [s_{in}(t,x,v_1), s_{out}(t,x,v_1)] \cap [s_{in}(t,x,v_2), s_{out}(t,x,v_2)]$, $s \neq t$, such that $X(s;t,x,v_1) = X(s;t,x,v_2)$. Since $X(t;t,x,v_1) = X(t;t,x,v_2) = x$, by Proposition 4.1 it follows that the characteristics coincide, and thus $v_1 = v_2$, which is in contradiction with the hypothesis. Therefore $X(s;t,x,v_1) - X(s;t,x,v_2)$ has constant sign on the intervals $[s_{in}(t,x,v_1), t[\cap [s_{in}(t,x,v_2), t[$ and $]t, s_{out}(t,x,v_1)] \cap]t, s_{out}(t,x,v_2)]$. On the other hand, we have

$$\frac{d}{ds}(X(s;t,x,v_1) - X(s;t,x,v_2))|_{s=t} = v_1 - v_2 < 0,$$

and therefore $X(s;t,x,v_1) - X(s;t,x,v_2)$ is decreasing locally in $s = t$. We deduce that

$$X(s;t,x,v_1) > X(s;t,x,v_2),\ s \in [s_{in}(t,x,v_1), t[\cap [s_{in}(t,x,v_2), t[,$$

and

$$X(s;t,x,v_1) < X(s;t,x,v_2),\ s \in]t, s_{out}(t,x,v_1)] \cap]t, s_{out}(t,x,v_2)].$$

By using the characteristics equations, one gets

$$\frac{d}{ds}(V(s;t,x,v_1) - V(s;t,x,v_2)) = E(s, X(s;t,x,v_1)) - E(s, X(s;t,x,v_2)),$$

and thus $V(s;t,x,v_1) - V(s;t,x,v_2)$ is nondecreasing on $[s_{in}(t,x,v_1), t[\cap [s_{in}(t,x,v_2), t[$ and nonincreasing on $]t, s_{out}(t,x,v_1)] \cap]t, s_{out}(t,x,v_2)]$. We deduce that

$$V(s;t,x,v_1) - V(s;t,x,v_2) \le v_1 - v_2 < 0,$$
$$s \in [s_{in}(t,x,v_1), s_{out}(t,x,v_1)] \cap [s_{in}(t,x,v_2), s_{out}(t,x,v_2)]. \qquad \square$$

When using the mild formulation of the Vlasov problem it is important to distinguish the characteristics with respect to the exit point. This justifies the following definitions: for $(t,x) \in \{[0,T[\times]0,1[\} \cup \{]0,T[\times\{0,1\}\}$ we denote by $\mathcal{V}^0, \mathcal{V}^1, \mathcal{V}^T$ the subsets of $\mathbb{R}_v$ given by

$$\mathcal{V}^0(T;t,x) := \{v \in \mathbb{R}_v : s_{out}(t,x,v) < T,\ X(s_{out}(t,x,v);t,x,v) = 0\}, \tag{4.1}$$

$$\mathcal{V}^1(T;t,x) := \{v \in \mathbb{R}_v : s_{out}(t,x,v) < T,\ X(s_{out}(t,x,v);t,x,v) = 1\}, \tag{4.2}$$

$$\mathcal{V}^T(T;t,x) := \{v \in \mathbb{R}_v : s_{out}(t,x,v) = T,\ 0 < X(T;t,x,v) < 1\}. \tag{4.3}$$

Note that when E is bounded there is R large enough such that $]-\infty, -R[\subset \mathcal{V}^0(T;t,x)$ and $]R, +\infty[\subset \mathcal{V}^1(T;t,x)$, and thus $\mathcal{V}^0(T;t,x) \neq \emptyset, \mathcal{V}^1(T;t,x) \neq \emptyset$. By the definition $\mathcal{V}^0(T;t,x) \cap \mathcal{V}^1(T;t,x) = \emptyset$ and $\mathcal{V}^T(T;t,x) \cap \{\mathcal{V}^0(T;t,x) \cup \mathcal{V}^1(T;t,x)\} = \emptyset$.

PROPOSITION 4.3. *Assume that $E \in L^\infty(]0,T[; W^{1,\infty}(]0,1[))$ is nondecreasing with respect to x. Then we have that*

(1) *if $v_2 \in \mathcal{V}^0(T;t,x)$, then $v_1 \in \mathcal{V}^0(T;t,x)$ $\forall v_1 < v_2$;*
(2) *if $v_1 \in \mathcal{V}^1(T;t,x)$, then $v_2 \in \mathcal{V}^1(T;t,x)$ $\forall v_2 > v_1$;*
(3) *if $v_1 \in \mathcal{V}^0(T;t,x)$, $v_2 \in \mathcal{V}^1(T;t,x)$, then $v_1 < v_2$.*

Proof. (1) Suppose that $s_{out}(t,x,v_1) \ge s_{out}(t,x,v_2)$. By Proposition 4.2 we deduce that

$$X(s;t,x,v_1) < X(s;t,x,v_2)\ \forall s \in]t, s_{out}(t,x,v_1)] \cap]t, s_{out}(t,x,v_2)] =]t, s_{out}(t,x,v_2)].$$

In particular, for $s = s_{out}(t,x,v_2)$ we find that

$$0 \le X(s_{out}(t,x,v_2);t,x,v_1) < X(s_{out}(t,x,v_2);t,x,v_2) = 0,$$

which is not possible. Finally, we have $s_{out}(t,x,v_1) < s_{out}(t,x,v_2) < T$ and

$$X(s_{out}(t,x,v_1);t,x,v_1) < X(s_{out}(t,x,v_1);t,x,v_2) < 1.$$

We deduce that $X(s_{out}(t,x,v_1);t,x,v_1) = 0$ or $v_1 \in \mathcal{V}^0(T;t,x)$.

(2) Similarly, if $v_1 \in \mathcal{V}^1(T;t,x)$ and $v_1 < v_2$ we have $s_{out}(t,x,v_2) < s_{out}(t,x,v_1) < T$ (otherwise $1 = X(s_{out}(t,x,v_1);t,x,v_1) < X(s_{out}(t,x,v_1);t,x,v_2)$) and $0 < X(s_{out}(t,x,v_2);t,x,v_1) < X(s_{out}(t,x,v_2);t,x,v_2)$. We deduce that $X(s_{out}(t,x,v_2);t,x,v_2) = 1$ and $v_2 \in \mathcal{V}^1(T;t,x)$.

(3) Suppose that $v_1 \ge v_2$. Since $v_1 \in \mathcal{V}^0(T;t,x)$, by (1) it follows that $v_2 \in \mathcal{V}^0(T;t,x) \cap \mathcal{V}^1(T;t,x) = \emptyset$. Therefore we have $v_1 < v_2$. $\square$

We introduce the critical velocities $v^0(T;t,x), v^1(T;t,x)$ given by

$$v^0(T;t,x) := \sup \mathcal{V}^0(T;t,x),\quad v^1(T;t,x) := \inf \mathcal{V}^1(T;t,x). \tag{4.4}$$

Obviously, we have $-\infty < v^0(T;t,x) \le v^1(T;t,x) < +\infty$.

PROPOSITION 4.4. *Assume that* $E \in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$ *is nondecreasing with respect to* x. *We have*

(1) $]-\infty, v^0(T;t,x)[\subset \mathcal{V}^0(T;t,x) \subset]-\infty, v^0(T;t,x)]$;

(2) $]v^1(T;t,x), +\infty[\subset \mathcal{V}^1(T;t,x) \subset [v^1(T;t,x), +\infty[$;

(3) $]v^0(T;t,x), v^1(T;t,x)[\subset \mathcal{V}^T(T;t,x) \subset [v^0(T;t,x), v^1(T;t,x)]$.

Proof. From Proposition 4.3 and the definitions of v^0, v^1 we deduce (1) and (2). On the other hand, $\mathcal{V}^T(T;t,x) \subset \mathbb{R}_v - \{\mathcal{V}^0(T;t,x) \cup \mathcal{V}^1(T;t,x)\} \subset \mathbb{R}_v - \{]-\infty, v^0(T;t;x)[\cup]v^1(T;t,x), +\infty[\} = [v^0(T;t,x), v^1(T;t,x)]$. Let us prove that $]v^0, v^1[\subset \mathcal{V}^T$. Consider $v^0 < v < v^1$ if $v^0 < v^1$. Suppose that $s_{out}(t,x,v) < T$ with $X(s_{out}(t,x,v);t,x,v) = 0$ or $v \in \mathcal{V}^0(T;t,x)$. By Proposition 4.3 we deduce that $\tilde{v} \in \mathcal{V}^0(T;t,x)\ \forall v^0 < \tilde{v} < v$, which is in contradiction with $\tilde{v} > v^0 = \sup \mathcal{V}^0(T;t,x)$. The same arguments apply for $s_{out}(t,x,v) < T, X(s_{out}(t,x,v);t,x,v) = 1$ by taking $v < \tilde{v} < v^1$. We have that $s_{out}(t,x,v) = T\ \forall v^0 < v < v^1$. Suppose now that $X(T;t,x,v) = 0$. If we take $v^0 < \tilde{v} < v$, we deduce that $s_{out}(t,x,\tilde{v}) = T$ and by Proposition 4.2 we find that $0 \le X(T;t,x,\tilde{v}) < X(T;t,x,v) = 0$. Similarly, we can show that $X(T;t,x,v) = 1$ is not possible. Finally, we deduce that $X(T;t,x,v) \in]0,1[\forall v^0 < v < v^1$, and thus $]v^0, v^1[\subset \mathcal{V}^T$. $\square$

Let us consider two fields A, B. In order to prove the uniqueness of the mild solution for the Vlasov–Poisson problem, it will be useful to estimate the change of critical velocity $|v_A^k - v_B^k|, k = 0, 1$, with respect to the relative field $A - B$. For this we need to introduce the notion of sub-/supercharacteristics as follows.

DEFINITION 4.5. *Assume that* $E \in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$ *is nondecreasing with respect to* x. *We say that* $(X(s), V(s))$ *is a subcharacteristic* (*resp., supercharacteristic*) *iff* X *is twice differentiable with respect to* s *and*

$$\frac{dX}{ds} = V(s), \quad \frac{dV}{ds} \le E(s, X(s)), \quad s_{in} \le s \le s_{out}$$

$$\left(resp., \frac{dX}{ds} = V(s), \quad \frac{dV}{ds} \ge E(s, X(s)), \quad s_{in} \le s \le s_{out} \right)$$

with the same definitions for s_{in}, s_{out} *as before.*

We have the following comparison result.

PROPOSITION 4.6 (forward comparison). *Assume that* $E \in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$ *is nondecreasing with respect to* x. *Consider* $(\underline{X}(s), \underline{V}(s)), (\overline{X}(s), \overline{V}(s))$ *to be a subcharacteristic, resp., a supercharacteristic, such that* $\underline{X}(t) \le \overline{X}(t),\ \underline{V}(t) \le \overline{V}(t)$. *Then we have*

$$\underline{X}(s) \le \overline{X}(s) \ \ \underline{V}(s) \le \overline{V}(s) \ \ \forall s \in [t, \underline{s}_{out}] \cap [t, \overline{s}_{out}].$$

Proof. We can extend the field E to $]0,T[\times\mathbb{R}_x$ by $\tilde{E}(t,x) = E(t,0), x < 0$, and $\tilde{E}(t,x) = E(t,1), x > 1$. We have $\|\tilde{E}\|_{L^\infty(]0,T[;W^{1,\infty}(]0,1[))} \le \|E\|_{L^\infty(]0,T[;W^{1,\infty}(]0,1[))}$ and $\tilde{E}$ is nondecreasing with respect to x. Consider $(x,v) \in \mathbb{R}_x \times \mathbb{R}_v$ such that $\underline{X}(t) \le x \le \overline{X}(t),\ \underline{V}(t) \le v \le \overline{V}(t)$. Denote by $(X(s;t,x,v), V(s;t,x,v))$ the characteristic associated with the field $\tilde{E}$,

$$\frac{dX}{ds} = V(s), \quad \frac{dV}{ds} = \tilde{E}(s, X(s)),$$

with the conditions $X(s=t;t,x,v)=x, \quad V(s=t;t,x,v)=v$. We show that $\underline{X}(s) \le X(s) \le \overline{X}(s),\ \underline{V}(s) \le V(s) \le \overline{V}(s)\ \forall s \in [t, \underline{s}_{out}] \cap [t, \overline{s}_{out}]$. For this we can use the iterated approximations method. For example, in order to prove that $\underline{X} \le X, \underline{V} \le V$ we consider as first approximation $X^0 = \underline{X}, V^0 = \underline{V}$ and we define $X^{n+1}(s) = x + \int_t^s V^n(\tau)d\tau,\ V^{n+1}(s) = v + \int_t^s \tilde{E}(\tau, X^n(\tau))d\tau\ \forall s \in [t, \underline{s}_{out}]\ \forall n \ge 0$. We check easily that $X^n(s) \ge \underline{X}(s), V^n(s) \ge \underline{V}(s)\ \forall s \in [t, \underline{s}_{out}]$ and by passing to the limit for $n \to +\infty$ we find that $X(s) \ge \underline{X}(s),\ V(s) \ge \underline{V}(s)\ \forall s \in [t, \underline{s}_{out}]$. In the same way, by taking as initial approximation $(X^0, V^0) = (\overline{X}, \overline{V})$ we prove that $X(s) \le \overline{X}(s), V(s) \le \overline{V}(s) \forall s \in [t, \overline{s}_{out}]$. Finally, we have

$$\underline{X}(s) \le X(s) \le \overline{X}(s), \quad \underline{V}(s) \le V(s) \le \overline{V}(s) \quad \forall s \in [t, \underline{s}_{out}] \cap [t, \overline{s}_{out}]. \qquad \square$$

Remark 4.7. In fact, since $0 \le \underline{X}(s), \overline{X}(s) \le 1\ \forall\ t \le s \le \min\{\underline{s}_{out}, \overline{s}_{out}\}$ it follows that $0 \le X(s) \le 1\ \forall\ t \le s \le \min\{\underline{s}_{out}, \overline{s}_{out}\}$, and therefore (X, V) coincide with the characteristic associated with the field E. Moreover, $s_{out}(t, x, v) \ge \min\{\underline{s}_{out}, \overline{s}_{out}\}$.

Now we are ready to prove a result of continuous dependence of the critical velocities with respect to the electric field. We have the following lemma.

LEMMA 4.8 (critical velocity change). *Assume that $A, B \in L^\infty(]0, T[; W^{1,\infty}(]0, 1[))$ are nondecreasing with respect to x. Then $\forall (t, x) \in [0, T[\times [0, 1]$ we have the following inequality:*

$$|v_A^k(T; t, x) - v_B^k(T; t, x)| \le \int_t^T \|A(s) - B(s)\|_{L^\infty(]0,1[)} ds, \ \ k = 0, 1. \tag{4.5}$$

Proof. Denote $m = \|A - B\|_{L^1(]t,T[; L^\infty(]0,1[))}$. Let us prove, for example, that $|v_A^0 - v_B^0| \le m$. Suppose that $v_A^0 - v_B^0 > m$. Therefore there is $v > v_B^0$ such that $\tilde{v} = v + m < v_A^0$, and thus we deduce from Proposition 4.4 that $X_B(s_{out}^B(t, x, v); t, x, v) > 0$, $X_A(s_{out}^A(t, x, \tilde{v}); t, x, \tilde{v}) = 0,\ s_{out}^A(t, x, \tilde{v}) < T$. Consider the solution (X_C, V_C) of the following system of ordinary differential equations:

$$\frac{dX_C}{ds} = V_C(s), \quad \frac{dV_C}{ds} = B(s, X_A(s)), \ \ t \le s \le s_{out}^C(t, x, v),$$

with the conditions $X_C(t) = x, V_C(t) = v$. With the notation

$$(X_A(s), V_A(s)) = (X_A(s; t, x, \tilde{v}), V_A(s; t, x, \tilde{v})),\ t \le s \le s_{out}^A(t, x, \tilde{v}),$$

and

$$(X_B(s), V_B(s)) = (X_B(s; t, x, v), V_B(s; t, x, v)),\ t \le s \le s_{out}^B(t, x, v),$$

we have also

$$\frac{dX_A}{ds} = V_A(s), \quad \frac{dV_A}{ds} = A(s, X_A(s)), \ \ t \le s \le s_{out}^A(t, x, \tilde{v}),$$

with $X_A(t) = x, V_A(t) = \tilde{v}$, and

$$\frac{dX_B}{ds} = V_B(s), \quad \frac{dV_B}{ds} = B(s, X_B(s)), \ \ t \le s \le s_{out}^B(t, x, v),$$

with $X_B(t) = x, V_B(t) = v$. We deduce that

$$\begin{aligned}\frac{d}{ds}(X_A - X_C) = V_A(s) - V_C(s), \frac{d}{ds}&(V_A - V_C) \\ &= (A - B)(s, X_A(s)), t \le s \le \min\{s_{out}^A, s_{out}^C\},\end{aligned}$$

and $X_A(t) - X_C(t) = 0,\ V_A(t) - V_C(t) = \tilde{v} - v = m$. We have

$$|V_A(s) - V_C(s) - V_A(t) + V_C(t)| \leq \int_t^s \|A(\tau) - B(\tau)\|_{L^\infty(]0,1[)} d\tau$$
$$\leq m, t \leq s \leq \min\{s_{out}^A, s_{out}^C\},$$

and thus $V_A(s) - V_C(s) \geq V_A(t) - V_C(t) - m = 0,\ \ t \leq s \leq \min\{s_{out}^A, s_{out}^C\}$. Moreover, since $X_A(t) = X_C(t) = x$, it follows that $X_A(s) \geq X_C(s),\ t \leq s \leq \min\{s_{out}^A, s_{out}^C\}$. If we suppose that $s_{out}^A < s_{out}^C$, we deduce that $X_C(s_{out}^A; t, x, v) \leq X_A(s_{out}^A; t, x, \tilde{v}) = 0$, and thus we have $s_{out}^C \leq s_{out}^A$, which is in contradiction with the previous supposition. Therefore we have $s_{out}^C \leq s_{out}^A < T$. In particular, $X_C(s_{out}^C; t, x, v) \in \{0, 1\}$ and $X_C(s_{out}^C; t, x, v) \leq X_A(s_{out}^C; t, x, \tilde{v})$. Note also that $X_A(s_{out}^C; t, x, \tilde{v}) = 1$ implies that $s_{out}^A \leq s_{out}^C$, and thus it follows that $s_{out}^A = s_{out}^C < T$, which is not possible because $X_A(s_{out}^A; t, x, \tilde{v}) = 0$ and $X_A(s_{out}^C; t, x, \tilde{v}) = 1$. We obtain that $X_C(s_{out}^C; t, x, v) \leq X_A(s_{out}^C; t, x, \tilde{v}) < 1$ and we deduce that $X_C(s_{out}^C; t, x, v) = 0$. On the other hand,

$$\frac{d^2}{ds^2} X_C = B(s, X_A(s)) \geq B(s, X_C(s)),\ t \leq s \leq s_{out}^C,$$

and

$$\frac{d^2}{ds^2} X_B = B(s, X_B(s)),\ t \leq s \leq s_{out}^B.$$

Note that $X_C(t) = X_B(t) = x$ and $V_C(t) = V_B(t) = v$. Thus by applying the forward comparison (see Proposition 4.6) we deduce that $X_C(s) \geq X_B(s), V_C(s) \geq V_B(s),\ t \leq s \leq \min\{s_{out}^B, s_{out}^C\}$. If we suppose that $s_{out}^C < s_{out}^B$, we deduce that

$$0 = X_C(s_{out}^C; t, x, v) \geq X_B(s_{out}^C; t, x, v),$$

and thus we have $s_{out}^B \leq s_{out}^C$, which is in contradiction with the previous supposition. Therefore we have $s_{out}^B \leq s_{out}^C \leq s_{out}^A < T$ and

$$X_B(s) \leq X_C(s) \leq X_A(s),\ V_B(s) \leq V_C(s) \leq V_A(s),\ \ t \leq s \leq s_{out}^B.$$

Since $v > v_B^0$ and $s_{out}^B < T$ we have $X_B(s_{out}^B; t, x, v) = 1$. Now, by taking $s = s_{out}^B$ in the previous inequality we obtain

$$1 = X_B(s_{out}^B; t, x, v) \leq X_A(s_{out}^B; t, x, \tilde{v}),$$

which implies that $X_A(s_{out}^B; t, x, \tilde{v}) = 1$ and $s_{out}^A \leq s_{out}^B$ or $s_{out}^A = s_{out}^B$. As before we obtain a contradiction because $X_A(s_{out}^A; t, x, \tilde{v}) = 0$ and $X_A(s_{out}^B; t, x, \tilde{v}) = 1$. Finally, we have proved that the supposition $v_A^0 - v_B^0 > m$ is false and thus $v_A^0 - v_B^0 \leq m$. By changing A to B we obtain also that $v_B^0 - v_A^0 \leq m$ or $|v_A^0 - v_B^0| \leq m$. The same arguments apply for the critical velocities v_A^1, v_B^1. ∎

We end this section with some usual calculations concerning the continuity of the characteristics with respect to the field.

PROPOSITION 4.9. *Assume that* $A, B \in L^\infty(]0, T[; W^{1,\infty}(]0,1[))$ *and consider* $(t, x, v) \in \{[0, T[\times]0, 1[\times\mathbb{R}_v\} \cup \{]0, T[\times\Sigma^-\}$. *Then for* $s \in [s_{in}^A(t, x, v), s_{out}^A(t, x, v)] \cap [s_{in}^B(t, x, v), s_{out}^B(t, x, v)]$ *we have*

$$|X_A(s; t, x, v) - X_B(s; t, x, v)| + |V_A(s; t, x, v) - V_B(s; t, x, v)|$$
$$\leq \left| \int_t^s \|A(\tau) - B(\tau)\|_{L^\infty(]0,1[)} d\tau \right| \cdot \exp\left(\left| \int_t^s (1 + \|\partial_x B(\tau)\|_{L^\infty(]0,1[)}) d\tau \right| \right).$$

5. Existence and uniqueness of the mild solution. In this section we intend to prove the existence and the uniqueness of the mild solution for the Vlasov–Poisson initial-boundary value problem in one dimension by using the iterated approximations method. We consider the application $\mathcal{F}$ defined for a regular electric field $E \in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$ as follows:

$$E \to f_E \to \rho_E = \int_{\mathbb{R}_v} f_E(t,x,v)dv \to E_1 = \mathcal{F}(E), \tag{5.1}$$

where f_E is the mild solution of the Vlasov problem associated with the field E and E_1 is the Poisson electric field corresponding to the charge density ρ_E . Before analyzing the application $\mathcal{F}$ let us introduce some notation. If $u : [0,+\infty[\to [0,+\infty[$ is a bounded nonincreasing real function and $R > 0$, we denote by $u^R : [-R,+\infty[\to [0,+\infty[$ the function given by $u^R(t) = u(0)$ if $-R \le t \le R$ and $u^R(t) = u(t-R)$ if $t > R$. If we assume that u belongs to $L^1(\mathbb{R}^+)$, and therefore,

$$\|u^R\|_{L^1(-R,+\infty)} = 2R\|u\|_{L^\infty(\mathbb{R}^+)} + \|u\|_{L^1(\mathbb{R}^+)}.$$

5.1. Estimate of $\mathcal{F}$. We assume that the initial-boundary conditions verify the following hypothesis denoted by (H): there are $n_0, h_0, h_1 : [0,+\infty[\to [0,+\infty[$ bounded, nonincreasing functions such that

$$(H)\quad \begin{aligned} &f_0(x,v) \le n_0(|v|), \quad (x,v) \in]0,1[\times \mathbb{R}_v,\\ &g_0(t,v) \le h_0(v), \quad (t,v) \in]0,T[\times \mathbb{R}_v^+,\\ &g_1(t,v) \le h_1(-v), \quad (t,v) \in]0,T[\times \mathbb{R}_v^-, \end{aligned}$$

and

$$(H_0)\quad M_0 := \int_{\mathbb{R}_v} n_0(|v|)dv + \int_{v>0} h_0(v)dv + \int_{v<0} h_1(-v)dv < +\infty,$$

$$(H_\infty)\quad M_\infty := \max\{\|n_0\|_{L^\infty(\mathbb{R}_v^+)}, \|h_0\|_{L^\infty(\mathbb{R}_v^+)}, \|h_1\|_{L^\infty(\mathbb{R}_v^+)}\} < +\infty.$$

Under the previous hypothesis we can prove the following proposition.

Proposition 5.1. *Assume that f_0, g_0, g_1 satisfy the hypotheses (H), (H_0), (H_∞) and $U_0 - U_1 \in L^\infty(]0,T[)$. Then for every $E \in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$ we have $f_E \in L^\infty(]0,T[;L^1(]0,1[\times\mathbb{R}_v))$, $\rho_E \in L^\infty(]0,T[;L^1(]0,1[)) \cap L^\infty(]0,T[\times]0,1[)$, $\mathcal{F}E \in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$. Moreover, the following estimates hold:*

$$\begin{aligned} \|f_E\|_{L^\infty(]0,t[;L^1(]0,1[\times\mathbb{R}_v))} &= \|\rho_E\|_{L^\infty(]0,t[;L^1(]0,1[))}\\ &\le 6\cdot M_\infty \int_0^t \|E(\tau)\|_{L^\infty(]0,1[)}d\tau + M_0,\\ \|\rho_E\|_{L^\infty(]0,t[\times]0,1[)} &= \|\partial_x \mathcal{F}E\|_{L^\infty(]0,t[\times]0,1[)}\\ &\le 6\cdot M_\infty \int_0^t \|E(\tau)\|_{L^\infty(]0,1[)}d\tau + M_0,\\ \|\mathcal{F}E\|_{L^\infty(]0,t[;W^{1,\infty}(]0,1[))} &\le 12\cdot M_\infty \int_0^t \|E(s)\|_{L^\infty(]0,1[)}ds\\ &\quad + 2M_0 + \|U_0 - U_1\|_{L^\infty(]0,t[)}, \end{aligned}$$

$$\lim_{R_1\to+\infty} \int_{|v|>R_1} f_E(t,x,v)dv = 0 \text{ uniformly with respect to } (t,x) \in]0,T[\times]0,1[,$$

and the mild formulation of the Vlasov problem holds for test functions $\psi \in L^\infty(]0,T[\times]0,1[\times\mathbb{R}_v)$.

Proof. By Remark 2.3 we have

$$\begin{aligned}\rho_E(t,x) &= \int_{\mathbb{R}_v} f_E(t,x,v)dv = \int_{\mathbb{R}_v} f_0(X(0;t,x,v),V(0;t,x,v))\mathbf{1}_{\{s_{in}(t,x,v)=0\}}dv \\ &\quad + \sum_{k=0}^{1}\int_{\mathbb{R}_v} g_k(s_{in}(t,x,v),V(s_{in}(t,x,v);t,x,v))\mathbf{1}_{\{s_{in}(t,x,v)>0\}} \\ &\quad \times \mathbf{1}_{\{X(s_{in}(t,x,v);t,x,v)=k\}}dv \\ &= \mathcal{I}^i + \mathcal{I}^0 + \mathcal{I}^1.\end{aligned}$$

Let us estimate the first integral $\mathcal{I}^i$. For this, we consider $R = \int_0^t \|E(\tau)\|_{L^\infty(]0,1[)}d\tau$ and remark that $|V(0;t,x,v)| \geq |v| - R$, which implies that $n_0(|V(0;t,x,v)|) \leq n_0^R(|v|)$. By using the hypothesis (H), we find

$$\begin{aligned}\mathcal{I}^i &\leq \int_{\mathbb{R}_v} n_0(|V(0;t,x,v)|)\mathbf{1}_{\{s_{in}(t,x,v)=0\}}dv \\ &\leq \int_{\mathbb{R}_v} n_0^R(|v|)dv = 2R\|n_0\|_{L^\infty(\mathbb{R}_v^+)} + 2\cdot\|n_0\|_{L^1(\mathbb{R}_v^+)}.\end{aligned}$$

In the same way, by writing $v \geq V(s_{in}(t,x,v);t,x,v) - R \geq -R$ when $X(s_{in}(t,x,v);t,x,v) = 0$ and $v \leq V(s_{in}(t,x,v);t,x,v) + R \leq R$ when $X(s_{in}(t,x,v);t,x,v) = 1$, one gets

$$\begin{aligned}\mathcal{I}^0 + \mathcal{I}^1 &\leq \int_{v>-R} h_0^R(v)dv + \int_{v<R} h_1^R(-v)dv \\ &\leq 2\cdot R\cdot(\|h_0\|_{L^\infty(\mathbb{R}_v^+)} + \|h_1\|_{L^\infty(\mathbb{R}_v^+)}) + \|h_0\|_{L^1(\mathbb{R}_v^+)} + \|h_1\|_{L^1(\mathbb{R}_v^+)}.\end{aligned}$$

Finally, we deduce that

$$\rho_E(t,x) \leq 6\cdot M_\infty \int_0^t \|E(\tau)\|_{L^\infty(]0,1[)}d\tau + M_0, (t,x) \in]0,T[\times]0,1[,$$

and therefore

$$\begin{aligned}|\mathcal{F}E(t,x)| &= \left|\int_0^x \rho_E(t,y)dy - \int_0^1 (1-y)\rho_E(t,y)dy - U_1(t) + U_0(t)\right| \\ &\leq \|\rho_E\|_{L^\infty(]0,t[;L^1(]0,1[))} + \|U_0 - U_1\|_{L^\infty]0,t[}.\end{aligned}$$

We remark that in order to estimate the charge outside a ball of radius R_1, we can write, for example,

$$\begin{aligned}\mathcal{I}^i_{R_1} &= \int_{|v|>R_1} f_0(X(0;t,x,v),V(0;t,x,v))\mathbf{1}_{\{s_{in}(t,x,v)=0\}}dv \\ &\leq \int_{|v|>R_1} n_0^R(|v|)dv = \int_{|v|>R_1-R} n_0(|v|)dv\end{aligned}$$

for $R_1 > R$. Finally, one gets that

$$\int_{|v|>R_1} f_E(t,x,v)dv \leq \int_{|v|>R_1-R} n_0(|v|)dv + \int_{v>R_1-R} h_0(v)dv + \int_{v<-R_1+R} h_1(-v)dv \to 0$$

as $R_1 \to +\infty$ uniformly with respect to $(t,x) \in]0,T[\times]0,1[$. Consider now $\psi \in L^\infty(]0,T[\times]0,1[\times\mathbb{R}_v)$ and $\psi_{R_1} = \chi_{R_1}(v)\psi(t,x,v)$, where $\chi_{R_1}(\cdot) = \chi(\cdot/R_1)$ and $\chi \in C_c^1(\mathbb{R}), \chi(u) = 1, |u| \le 1, \chi(u) = 0, |u| \ge 2, 0 \le \chi(u) \le 1, 1 \le |u| \le 2$. Obviously, $\psi_{R_1} \in \mathcal{T}_m$, and thus

$$\begin{aligned}
&\int_0^T\int_0^1\int_{\mathbb{R}_v} f_E(t,x,v)\psi_{R_1}(t,x,v)dtdxdv \\
&\quad= \int_0^1\int_{\mathbb{R}_v} f_0(x,v)\int_0^{s_{out}(0,x,v)} \psi_{R_1}(s, X(s;0,x,v), V(s;0,x,v))dsdxdv \\
&\quad+ \int_0^T\int_{v>0} vg_0(t,v)\int_t^{s_{out}(t,0,v)} \psi_{R_1}(s, X(s;t,0,v), V(s;t,0,v))dsdtdv \\
&\quad- \int_0^T\int_{v<0} vg_1(t,v)\int_t^{s_{out}(t,1,v)} \psi_{R_1}(s, X(s;t,1,v), V(s;t,1,v))dsdtdv.
\end{aligned}$$

We have

$$\begin{aligned}
\left|\int_0^T\int_0^1\int_{\mathbb{R}_v} f_E\psi_{R_1}dtdxdv - \int_0^T\int_0^1\int_{\mathbb{R}_v} f_E\psi dtdxdv\right| &\le \int_0^T\int_0^1\int_{\mathbb{R}_v} f_E(1-\chi_{R_1}(v))|\psi|dtdxdv \\
&\le \|\psi\|_{L^\infty}\int_0^T\int_0^1\int_{|v|>R_1} f_E dtdxdv \to 0 \\
&\quad \text{as } R_1 \to +\infty.
\end{aligned}$$

In order to apply the dominated convergence theorem of Lebesgue, observe that

$$|f_0(x,v)\int_0^{s_{out}^i} \psi_{R_1}(s, X^i(s), V^i(s))ds| \le f_0(x,v)\|\psi\|_{L^\infty}T \in L^1(]0,1[\times\mathbb{R}_v).$$

Note also that for $R = \|E\|_{L^1(]0,T[;L^\infty(]0,1[))}$ we have

$$\begin{aligned}
|vg_0(t,v)\int_t^{s_{out}^0}\psi_{R_1}(s, X^0(s), V^0(s))ds| &\le 2Rg_0(t,v)T\|\psi\|_{L^\infty}\mathbf{1}_{\{0<v\le 2R\}} \\
&\quad + vg_0(t,v)\|\psi\|_{L^\infty}\frac{1}{v-R}\mathbf{1}_{\{v>2R\}} \\
&\le 2RT\|\psi\|_{L^\infty}g_0(t,v)\mathbf{1}_{\{0<v\le 2R\}} \\
&\quad + 2\|\psi\|_{L^\infty}g_0(t,v)\mathbf{1}_{\{v>2R\}} \in L^1(]0,T[\times\mathbb{R}_v^+)
\end{aligned}$$

since $V^0(s) \ge v - R$ and $s_{out}^0 - t \le \frac{1}{v-R}$ for $v > R$. The same arguments apply for the right boundary term. Finally, by passing $R_1 \to +\infty$ we deduce that the mild formulation holds for every $\psi \in L^\infty(]0,T[\times]0,1[\times\mathbb{R}_v)$. ∎

Remark 5.2. Consider $x(t) = (M_0 + \|U_0 - U_1\|_{L^\infty(]0,T[)})\exp(6\cdot M_\infty t)$ and

$$X_T = \{E \in L^\infty(]0,T[;W^{1,\infty}(]0,1[)) \mid \|E\|_{L^\infty(]0,t[\times]0,1[)} \le x(t)\forall\, 0 \le t \le T\}.$$

Then $\mathcal{F}X_T \subset X_T$ and

$$\|\mathcal{F}E\|_{L^\infty(]0,T[;W^{1,\infty}(]0,1[))} \le 2\cdot x(T) - \|U_0 - U_1\|_{L^\infty(]0,T[)}.$$

5.2. Estimate of $\mathcal{F}\quad - \mathcal{F}\quad$. The aim of this section is to estimate the L^∞ norm of $\mathcal{F}A - \mathcal{F}B$ with respect to the L^∞ norm of $A - B$. First, we perform our computations by introducing the current density $j_E(t,x) := \int_{\mathbb{R}_v} v f_E(t,x,v)dv$. This requires an additional hypothesis on the initial-boundary conditions. For the moment we assume also that

$$(H_1)\ \ M_1 := \int_{\mathbb{R}_v} n_0(|v|)|v|dv + \int_{v>0} h_0(v)vdv - \int_{v<0} h_1(-v)vdv < +\infty.$$

Later on we shall see that this hypothesis can be removed.

PROPOSITION 5.3. *Assume that f_0, g_0, g_1 satisfy $(H), (H_1), (H_\infty)$ and $U_0 - U_1 \in L^\infty(]0,T[)$. Then for every $E \in L^\infty(]0,T[; W^{1,\infty}(]0,1[)) f_E|v| \in L^\infty(]0,T[; L^1(]0,1[\times\mathbb{R}_v))$, $|j_E|(t,x) := \int_{\mathbb{R}_v} f_E(t,x,v)|v|dv \in L^\infty(]0,T[; L^1(]0,1[)) \cap L^\infty(]0,T[\times]0,1[)$, $\mathcal{F}E + U_1 - U_0 \in W^{1,\infty}(]0,T[\times]0,1[)$. Moreover, the following estimates hold:*

$$\max\{\|\,|j_E|\,\|_{L^\infty(]0,T[;L^1(]0,1[))}, \||j_E|\,\|_{L^\infty(]0,T[\times]0,1[)}\} \le 3\cdot M_\infty \left(\int_0^t \|E(s)\|_{L^\infty(]0,1[)}ds\right)^2 + M_0\int_0^t \|E(s)\|_{L^\infty(]0,1[)}ds + M_1,$$

$$\partial_t\{\mathcal{F}E + U_1 - U_0\} = -j_E(t,x) + \int_0^1 j_E(t,y)dy,\ (t,x) \in]0,T[\times]0,1[,$$

$$\lim_{R_1\to+\infty} \int_{|v|>R_1} |v| f_E(t,x,v)dv = 0 \ \ \textit{uniformly with respect to } (t,x) \in]0,T[\times]0,1[,$$

and the mild formulation of the Vlasov problem holds for every function ψ such that $|\psi(t,x,v)| \le C(1+|v|)$.

Proof. Exactly as before, we have

$$\begin{aligned}|j_E|(t,x) &= \int_{\mathbb{R}_v} |v| f_E(t,x,v)dv = \int_{\mathbb{R}_v} |v| f_0(X(0;t,x,v), V(0;t,x,v))\mathbf{1}_{\{s_{in}(t,x,v)=0\}} \\ &\quad + \sum_{k=0}^1 \int_{\mathbb{R}_v} |v| g_k(s_{in}(t,x,v), V(s_{in}(t,x,v);t,x,v))\mathbf{1}_{\{s_{in}(t,x,v)>0\}} \\ &\quad \times \mathbf{1}_{\{X(s_{in}(t,x,v);t,x,v)=k\}}dv \\ &= \mathcal{J}^i + \mathcal{J}^0 + \mathcal{J}^1.\end{aligned}$$

Consider $R = \int_0^t \|E(s)\|_{L^\infty(]0,1[)}ds$ and thus $|V(0;t,x,v)| \ge |v| - R$, which implies that

$$\mathcal{J}^i \le \int_{\mathbb{R}_v} |v| n_0^R(|v|)dv = R^2 n_0(0) + \int_{\mathbb{R}_v} |v| n_0(|v|)dv + R\int_{\mathbb{R}_v} n_0(|v|)dv.$$

The terms $\mathcal{J}^k, k \in \{0,1\}$ can be estimated in the same manner and, finally, one gets

$$|j_E|(t,x) \le 3\cdot R^2 M_\infty + RM_0 + M_1, \qquad (t,x) \in]0,T[\times]0,1[.$$

By performing the same computations on $\mathbb{R}_v - B_{R_1}$, we get that $\lim_{R_1\to+\infty} \int_{|v|>R_1} |v| f_E\, dv = 0$ uniformly with respect to $(t,x) \in]0,T[\times]0,1[$. In order to check that the mild formulation holds $\forall\psi$ such that $|\psi(t,x,v)| \le C(1+|v|)$, consider $\psi_{R_1} = \chi_{R_1}(v)\psi \in \mathcal{T}_m$. This time we have

$$\begin{aligned}\left|\int_0^T\int_0^1\int_{\mathbb{R}_v} f_E\psi_{R_1}dtdxdv - \int_0^T\int_0^1\int_{\mathbb{R}_v} f_E\psi dtdxdv\right| &\le \int_0^T\int_0^1\int_{\mathbb{R}_v} f_E(1-\chi_{R_1}(v))|\psi(t,x,v)|dtdxdv\\ &\le \int_0^T\int_0^1\int_{|v|>R_1} f_E\cdot C(1+|v|)dtdxdv \to 0,\\ &\quad \text{as } R_1\to+\infty.\end{aligned}$$

In order to pass to the limit in the other terms of the mild formulation for the test function ψ_{R_1}, take $R = \|E\|_{L^1(]0,T[;L^\infty(]0,1[))}$ and note that

$$\left|f_0(x,v)\int_0^{s^i_{out}}\psi_{R_1}(s,X^i(s),V^i(s))ds\right| \le f_0(x,v)\cdot T\cdot C(1+|v|+R) \in L^1(]0,1[\times\mathbb{R}_v),$$

$$\begin{aligned}\left|vg_k(t,v)\int_t^{s^k_{out}}\psi_{R_1}(s,X^k(s),V^k(s))ds\right| &\le 2Rg_k(t,v)\cdot T\cdot C(1+|v|+R)\mathbf{1}_{\{|v|\le 2R\}}\\ &\quad + |v|g_k(t,v)\frac{C(1+|v|+R)}{|v|-R}\mathbf{1}_{\{|v|>2R\}}\\ &\le 2R\cdot T\cdot C\cdot g_k(t,v)(1+|v|+R)\mathbf{1}_{\{|v|\le 2R\}}\\ &\quad + C\left(3+\frac{1}{R}\right)|v|g_k(t,v)\mathbf{1}_{\{|v|>2R\}}\\ &\qquad\qquad \in L^1(]0,T[\times\mathbb{R}_v^{\pm}).\end{aligned}$$

By passing to the limit in the mild formulation for $R_1 \to +\infty$ and using the dominated convergence theorem, our conclusion follows. Let us compute now the time derivative of $\mathcal{F}E + U_1 - U_0$. First of all, by using the mild formulation with the test function $\psi(t,x,v) = \partial_t\varphi + v\partial_x\varphi$, $\varphi \in C^1_c(]0,T[\times]0,1[)$ (note that $|\psi(t,x,v)| \le C(1+|v|)$) we deduce the continuity equation $\partial_t\rho_E + \partial_x j_E = 0$ in $\mathcal{D}'(]0,T[\times]0,1[)$. By direct computation, the continuity equation implies that

$$\partial_t\{\mathcal{F}E + U_1 - U_0\} = -j_E(t,x) + \int_0^1 j_E(t,y)dy \in L^\infty(]0,T[\times]0,1[).$$

Obviously $\partial_x\{\mathcal{F}E + U_1 - U_0\} = \rho_E \in L^\infty(]0,T[\times]0,1[)$ and thus we obtain that $\mathcal{F}E + U_1 - U_0 \in W^{1,\infty}(]0,T[\times]0,1[)$. □

Remark 5.4. We have

$$\begin{aligned}\mathcal{F}E(t,x) + U_1(t) - U_0(t) &= -\int_0^t j_E(s,x)ds + \int_0^t\int_0^1 j_E(s,y)dsdy + \mathcal{F}E(0,x)\\ &\quad + U_1(0) - U_0(0)\\ &= -\int_0^t j_E(s,x)ds + \int_0^t\int_0^1 j_E(s,y)dsdy\\ &\quad + \int_0^x\int_{\mathbb{R}_v} f_0(y,v)dydv - \int_0^1\int_{\mathbb{R}_v}(1-y)f_0(y,v)dydv.\end{aligned}$$

By using the formula given above we can estimate $\mathcal{F}A - \mathcal{F}B$. This will be done in the following two propositions. One of the key points is the critical velocity change result (see Lemma 4.8).

PROPOSITION 5.5. *Assume that $A, B \in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$ are nondecreasing with respect to x and the hypotheses $(H),(H_1),(H_\infty)$ hold. Then for $0 \le t \le T$ we have*

$$\left\| \int_0^t j_A(s,\cdot)ds - \int_0^t j_B(s,\cdot)ds \right\|_{L^\infty(]0,1[)} \le C \cdot \int_0^t \|A(s) - B(s)\|_{L^\infty(]0,1[)} ds,$$

where C is a constant depending only on $\|A\|_{L^1(]0,T[;W^{1,\infty}(]0,1[))}, \|B\|_{L^1(]0,T[;W^{1,\infty}(]0,1[))}$, T and the initial-boundary conditions.

Proof. Consider $\varphi \in L^1(]0,1[)$ bounded and let us estimate $\int_0^1\int_0^t (j_A(s,x) - j_B(s,x))\varphi(x)dxds$. By applying the mild formulation with $\psi(t,x,v) = \varphi(x)v$ (which is possible since $|\psi(t,x,v)| \le \|\varphi\|_{L^\infty}|v|$) we have

$$\begin{aligned}
&\int_0^1\int_0^t (j_A(s,x) - j_B(s,x))\varphi(x)dxds \\
&= \int_0^t\int_0^1\int_{\mathbb{R}_v} (f_A(s,x,v) - f_B(s,x,v))v\varphi(x)dsdxdv \\
&= \int_0^1\int_{\mathbb{R}_v} f_0(x,v)\left[\int_0^{s_A^i} V_A^i(\tau)\varphi(X_A^i(\tau))d\tau - \int_0^{s_B^i} V_B^i(\tau)\varphi(X_B^i(\tau))d\tau\right]dxdv \\
&\quad + \int_0^t\int_{v>0} vg_0(s,v)\left[\int_s^{s_A^0} V_A^0(\tau)\varphi(X_A^0(\tau))d\tau - \int_s^{s_B^0} V_B^0(\tau)\varphi(X_B^0(\tau))d\tau\right]dsdv \\
&\quad - \int_0^t\int_{v<0} vg_1(s,v)\left[\int_s^{s_A^1} V_A^1(\tau)\varphi(X_A^1(\tau))d\tau - \int_s^{s_B^1} V_B^1(\tau)\varphi(X_B^1(\tau))d\tau\right]dsdv \\
&= \int_0^1\int_{\mathbb{R}_v} f_0(x,v)\left[\int_x^{X_A^i(s_A^i)} \varphi(u)du - \int_x^{X_B^i(s_B^i)} \varphi(u)du\right]dxdv \\
&\quad + \int_0^t\int_{v>0} vg_0(s,v)\left[\int_0^{X_A^0(s_A^0)} \varphi(u)du - \int_0^{X_B^0(s_B^0)} \varphi(u)du\right]dsdv \\
&\quad - \int_0^t\int_{v<0} vg_1(s,v)\left[\int_1^{X_A^1(s_A^1)} \varphi(u)du - \int_1^{X_B^1(s_B^1)} \varphi(u)du\right]dsdv \\
&= \mathcal{I}_{AB}^i + \mathcal{I}_{AB}^0 + \mathcal{I}_{AB}^1.
\end{aligned}$$

We introduce the notation $\Phi_C^i = \int_x^{X_C^i(s_C^i)}\varphi(u)du$, $\Phi_C^k = \int_k^{X_C^k(s_C^k)}\varphi(u)du$, $k \in \{0,1\}$, $C \in \{A,B\}$. Here s_C^i, s_C^k represent the exit times associated with the domain $]0,t[\times]0,1[\times\mathbb{R}_v$, with $k \in \{0,1\}, C \in \{A,B\}$. The term $\mathcal{I}_{AB}^i$ is written

$$\begin{aligned}
\mathcal{I}_{AB}^i &= \int_0^1\int_{\mathbb{R}_v} f_0(x,v)[\Phi_A^i\mathbf{1}_{\{v<v_A^0\}} - \Phi_B^i\mathbf{1}_{\{v<v_B^0\}}]dxdv \\
&\quad + \int_0^1\int_{\mathbb{R}_v} f_0(x,v)[\Phi_A^i\mathbf{1}_{\{v_A^0<v<v_A^1\}} - \Phi_B^i\mathbf{1}_{\{v_B^0<v<v_B^1\}}]dxdv \\
&\quad + \int_0^1\int_{\mathbb{R}_v} f_0(x,v)[\Phi_A^i\mathbf{1}_{\{v>v_A^1\}} - \Phi_B^i\mathbf{1}_{\{v>v_B^1\}}]dxdv \\
&= \mathcal{I}_0^i + \mathcal{I}_t^i + \mathcal{I}_1^i,
\end{aligned}$$

where $v_C^k = v_C^k(t;0,x)$ are the critical velocities corresponding to the domain $]0,t[\times]0,1[$, the point $(0,x)$, and the field C, with $k = 0,1,\ C = A,B$. The first and third integrals are easy to estimate since for $v < v_A^0$ we have $X_A^i(s_A^i) = 0$ and thus $\Phi_A^i = \int_x^0 \varphi(u)du$; for $v > v_A^1$ we have $X_A^i(s_A^i) = 1$ and $\Phi_A^i = \int_x^1 \varphi(u)du$. By using the critical velocity change, we obtain

$$\begin{aligned}|\mathcal{I}_0^i| &\leq \|\varphi\|_{L^1(]0,1[)}\|f_0\|_{L^\infty(]0,T[\times]0,1[\times\mathbb{R}_v)}\int_0^1 |v_A^0(t;0,x) - v_B^0(t;0,x)|dx\\ &\leq \|\varphi\|_{L^1(]0,1[)}\|f_0\|_{L^\infty(]0,T[\times]0,1[\times\mathbb{R}_v)}\int_0^t \|A(s) - B(s)\|_{L^\infty(]0,1[)}ds\end{aligned}$$

and also

$$|\mathcal{I}_1^i| \leq \|\varphi\|_{L^1(]0,1[)}\|f_0\|_{L^\infty(]0,T[\times]0,1[\times\mathbb{R}_v)}\int_0^t \|A(s) - B(s)\|_{L^\infty(]0,1[)}ds.$$

Let us estimate now the second integral $\mathcal{I}_t^i$. We remark that when $v_A^0 < v < v_A^1$ we have $s_{out}^A(0,x,v) = t$ and thus $\Phi_A^i = \int_x^{X_A(t)}\varphi(u)du$. Similarly, $\Phi_B^i = \int_x^{X_B(t)}\varphi(u)du$ when $v_B^0 < v < v_B^1$. We can write

$$\begin{aligned}|\mathcal{I}_t^i| \leq &\left|\int_0^1\int_{\mathbb{R}_v} f_0(x,v)\Phi_A^i \mathbf{1}_{\{v_A^0 < v < \max\{v_A^0,v_B^0\}\}}dxdv\right|\\ &+\left|\int_0^1\int_{\mathbb{R}_v} f_0(x,v)\Phi_A^i \mathbf{1}_{\{\min\{v_A^1,v_B^1\} < v < v_A^1\}}dxdv\right|\\ &+\left|\int_0^1\int_{\mathbb{R}_v} f_0(x,v)\Phi_B^i \mathbf{1}_{\{v_B^0 < v < \max\{v_A^0,v_B^0\}\}}dxdv\right|\\ &+\left|\int_0^1\int_{\mathbb{R}_v} f_0(x,v)\Phi_B^i \mathbf{1}_{\{\min\{v_A^1,v_B^1\} < v < v_B^1\}}dxdv\right|\\ &+\left|\int_0^1\int_{\mathbb{R}_v} f_0(x,v)(\Phi_A^i - \Phi_B^i) \mathbf{1}_{\{\max\{v_A^0,v_B^0\} < v < \min\{v_A^1,v_B^1\}\}}dxdv\right|.\end{aligned}$$

By using Lemma 4.8 we deduce

$$\max\{|v_A^0 - \max\{v_A^0,v_B^0\}|, |v_B^0 - \max\{v_A^0,v_B^0\}|\} \leq \int_0^t \|A(s) - B(s)\|_{L^\infty(]0,1[)}ds$$

and

$$\max\{|v_A^1 - \min\{v_A^1,v_B^1\}|, |v_B^1 - \min\{v_A^1,v_B^1\}|\} \leq \int_0^t \|A(s) - B(s)\|_{L^\infty(]0,1[)}ds.$$

The first four terms can be estimated by $4\cdot\|\varphi\|_{L^1}\|f_0\|_{L^\infty}\int_0^t \|A(s) - B(s)\|_{L^\infty}ds$. When $\max\{v_A^0,v_B^0\} < v < \min\{v_A^1,v_B^1\}$ we have

$$\begin{aligned}|\Phi_A^i - \Phi_B^i| &= \left|\int_{X_B(t)}^{X_A(t)}\varphi(u)du\right|\\ &\leq \int_0^1 |\varphi(u)|\mathbf{1}_{\{|u - X_A(t)| \leq |X_A(t) - X_B(t)|\}}du.\end{aligned}$$

Therefore, by using Proposition 4.9, we write the last term of $\mathcal{I}_t^i$ as

$$|\mathcal{I}_5| = \left| \int_0^1 \int_{\mathbb{R}_v} f_0(x,v)(\Phi_A^i - \Phi_B^i) \mathbf{1}_{\{\max\{v_A^0, v_B^0\} < v < \min\{v_A^1, v_B^1\}\}} dxdv \right|$$

$$\leq \int_0^1 |\varphi(u)| \int_0^1 \int_{\mathbb{R}_v} f_0(x,v) \mathbf{1}_{\{v_A^0 < v < v_A^1\}} \mathbf{1}_{\{|u - X_A(t)| \leq C \int_0^t \|A(s) - B(s)\|_{L^\infty} ds\}} dxdvdu, \tag{5.2}$$

where $C = \exp(\int_0^t (1 + \|\partial_x B(s)\|_{L^\infty(]0,1[)}) ds)$. By the change of variables $y = X_A(t;0,x,v)$, $w = V_A(t;0,x,v)$ on $\{(x,v) \in]0,1[\times \mathbb{R}_v \ : \ v_A^0(t;0,x) < v < v_A^1(t;0,x)\}$, one gets

$$\begin{aligned}
|\mathcal{I}_5| &\leq \int_0^1 |\varphi(u)| \int_0^1 \int_{\mathbb{R}_w} f_0(X_A(0;t,y,w), V_A(0;t,y,w)) \mathbf{1}_{\{|u-y| \leq C \int_0^t \|A(s)-B(s)\|_{L^\infty} ds\}} dydwdu \\
&\leq \int_0^1 |\varphi(u)| \int_0^1 \int_{\mathbb{R}_w} n_0^R(|w|) \mathbf{1}_{\{|u-y| \leq C \int_0^t \|A(s)-B(s)\|_{L^\infty} ds\}} dydwdu \\
&\leq 2 \cdot C \cdot \int_0^t \|A(s) - B(s)\|_{L^\infty} ds (2 \cdot R \cdot \|n_0\|_{L^\infty(\mathbb{R}_v^+)} + 2 \cdot \|n_0\|_{L^1(\mathbb{R}_v^+)}) \|\varphi\|_{L^1(]0,1[)},
\end{aligned}$$

where as usual $R = \int_0^t \|A(s)\|_{L^\infty(]0,1[)} ds$. Finally, we proved that

$$\begin{aligned}
|\mathcal{I}_{AB}^i| &\leq \left\{ 6\|f_0\|_{L^\infty} + 4C \left(\int_0^t \|A(s)\|_{L^\infty} ds \|n_0\|_{L^\infty(\mathbb{R}_v^+)} + \|n_0\|_{L^1(\mathbb{R}_v^+)} \right) \right\} \\
&\quad \times \int_0^t \|A(s) - B(s)\|_{L^\infty} ds \cdot \|\varphi\|_{L^1} \\
&\leq C^i \cdot \int_0^t \|A(s) - B(s)\|_{L^\infty} ds \cdot \|\varphi\|_{L^1(]0,1[)}.
\end{aligned}$$

Let us analyze the term $\mathcal{I}_{AB}^0$. As before, we have

$$\begin{aligned}
\mathcal{I}_{AB}^0 &= \int_0^t \int_{\mathbb{R}_v} v g_0(s,v) [\Phi_A^0 \mathbf{1}_{\{0 < v < v_A^0\}} - \Phi_B^0 \mathbf{1}_{\{0 < v < v_B^0\}}] dsdv \\
&\quad + \int_0^t \int_{\mathbb{R}_v} v g_0(s,v) [\Phi_A^0 \mathbf{1}_{\{v_A^0 < v < v_A^1\}} - \Phi_B^0 \mathbf{1}_{\{v_B^0 < v < v_B^1\}}] dsdv \\
&\quad + \int_0^t \int_{\mathbb{R}_v} v g_0(s,v) [\Phi_A^0 \mathbf{1}_{\{v > v_A^1\}} - \Phi_B^0 \mathbf{1}_{\{v > v_B^1\}}] dsdv \\
&= \mathcal{I}_0^0 + \mathcal{I}_t^0 + \mathcal{I}_1^0.
\end{aligned}$$

Taking into account that for $0 < v < v_C^0(t;s,0)$ we have $X_C^0(s_{out,C}^0) = 0$, we deduce that $\Phi_C^0 = 0$ for $C = A, B$ and thus $\mathcal{I}_0^0 = 0$. By the other hand, for $v > v_C^1$ we have $X_C^0(s_{out,C}^0) = 1$ and thus $\Phi_C^0 = \int_0^1 \varphi(u) du$ for $C = A, B$. One gets

$$|\mathcal{I}_1^0| \leq \left| \int_0^t \int_{v_A^1}^{v_B^1} v g_0(s,v) \int_0^1 \varphi(u) dsdvdu \right| \leq t \cdot \|v g_0\|_{L^\infty(]0,T[\times \mathbb{R}_v^+)} |v_A^1 - v_B^1| \cdot \|\varphi\|_{L^1(]0,1[)}.$$

By applying Lemma 4.8 we have

$$|v_A^1(t;s,0) - v_B^1(t;s,0)| \leq \int_s^t \|A(\tau) - B(\tau)\|_{L^\infty(]0,1[)} d\tau,$$

and therefore

$$|\mathcal{I}_1^0| \leq t \cdot \|vg_0\|_{L^\infty(]0,T[\times\mathbb{R}_v^+)} \cdot \|\varphi\|_{L^1(]0,1[)} \cdot \int_0^t \|A(s)-B(s)\|_{L^\infty(]0,1[)}ds.$$

The term $\mathcal{I}_t^0$ is written

$$\begin{aligned}|\mathcal{I}_t^0| \leq & \left|\int_0^t\int_{\mathbb{R}_v} vg_0(s,v)\Phi_A^0\mathbf{1}_{\{v_A^0<v<\max\{v_A^0,v_B^0\}\}}dsdv\right| \\ &+\left|\int_0^t\int_{\mathbb{R}_v} vg_0(s,v)\Phi_A^0\mathbf{1}_{\{\min\{v_A^1,v_B^1\}<v<v_A^1\}}dsdv\right| \\ &+\left|\int_0^t\int_{\mathbb{R}_v} vg_0(s,v)\Phi_B^0\mathbf{1}_{\{v_B^0<v<\max\{v_A^0,v_B^0\}\}}dsdv\right| \\ &+\left|\int_0^t\int_{\mathbb{R}_v} vg_0(s,v)\Phi_B^0\mathbf{1}_{\{\min\{v_A^1,v_B^1\}<v<v_B^1\}}dsdv\right| \\ &+\left|\int_0^t\int_{\mathbb{R}_v} vg_0(s,v)(\Phi_A^0-\Phi_B^0)\mathbf{1}_{\{\max\{v_A^0,v_B^0\}<v<\min\{v_A^1,v_B^1\}\}}dsdv\right|.\end{aligned}$$

The first four terms can be estimated as before by

$$t \cdot \|vg_0\|_{L^\infty(]0,T[\times\mathbb{R}_v^+)} \cdot \|\varphi\|_{L^1(]0,1[)} \cdot \int_0^t \|A(s)-B(s)\|_{L^\infty(]0,1[)}ds.$$

Since for $\max\{v_A^0,v_B^0\} < v < \min\{v_A^1,v_B^1\}$ we have $\Phi_A^0 - \Phi_B^0 = \int_{X_B(t)}^{X_A(t)} \varphi(u)du$, the last term is written

$$\begin{aligned}|\mathcal{I}_5| \leq & \left|\int_0^t\int_{v>0} vg_0(s,v)\int_{X_B(t)}^{X_A(t)}\varphi(u)\mathbf{1}_{\{\max\{v_A^0,v_B^0\}<v<\min\{v_A^1,v_B^1\}\}}dsdvdu\right| \\ \leq & \int_0^1|\varphi(u)|\int_0^t\int_{v>0} vg_0(s,v)\mathbf{1}_{\{|u-X_A(t;s,0,v)|<|X_A(t;s,0,v)-X_B(t;s,0,v)|\}} \\ & \times\mathbf{1}_{\{\max\{v_A^0,v_B^0\}<v<\min\{v_A^1,v_B^1\}\}} \\ \leq & \int_0^1|\varphi(u)|\int_0^t\int_{v>0} vg_0(s,v)\mathbf{1}_{\{|u-X_A(t;s,0,v)|<C\cdot\int_0^t\|A(\tau)-B(\tau)\|_{L^\infty}d\tau\}}\mathbf{1}_{\{v_A^0<v<v_A^1\}}dudsdv.\end{aligned}$$

This time we perform the change of variables $(y,w) = S(s,v)$, with $y = X_A(t;s,0,v)$, $w = V_A(t;s,0,v)$ on the set $D = \{(s,v) \in]0,t[\times\mathbb{R}_v : v_A^0(t;s,0) < v < v_A^1(t;s,0)\}$. By standard computations one gets that

$$\left|\frac{\partial(y,w)}{\partial(s,v)}\right| = |v|,$$

and thus

$$\begin{aligned}|\mathcal{I}_5| \leq & \int_0^1|\varphi(u)|\int_0^1\int_{w>-R}\mathbf{1}_{\{(y,w)\in S(D)\}}g_0(s_{in}(t,y,w), \\ & \qquad \cdot V(s_{in}(t,y,w);t,y,w))\mathbf{1}_{\{|u-y|<C\cdot\int_0^t\|A(\tau)-B(\tau)\|_{L^\infty}d\tau\}} \\ \leq & \int_0^1|\varphi(u)|\int_0^1\int_{w>-R}h_0^R(w)\mathbf{1}_{\{|u-y|<C\cdot\int_0^t\|A(\tau)-B(\tau)\|_{L^\infty}d\tau\}}dydwdu \\ \leq & \, 2C\int_0^t\|A(\tau)-B(\tau)\|_{L^\infty}d\tau\cdot(2R\|h_0\|_{L^\infty(\mathbb{R}_v^+)}+\|h_0\|_{L^1(\mathbb{R}_v^+)})\cdot\|\varphi\|_{L^1(]0,1[)},\end{aligned}$$

where $R = \int_0^t \|A(\tau)\|_{L^\infty} d\tau$. Finally, one gets

$$\begin{aligned}|\mathcal{I}^0_{AB}| \le &\left\{5 \cdot t \cdot \|vh_0\|_{L^\infty(\mathbb{R}^+_v)} + 2\exp\left(\int_0^t (1 + \|\partial_x B(s)\|_{L^\infty(]0,1[)})ds\right)\right.\\ &\left.\cdot\left(2 \cdot \|h_0\|_{L^\infty(\mathbb{R}^+_v)} \cdot \int_0^t \|A(s)\|_{L^\infty(]0,1[)}ds + \|h_0\|_{L^1(\mathbb{R}^+_v)}\right)\right\}\\ &\cdot \int_0^t \|A(s) - B(s)\|_{L^\infty} ds \cdot \|\varphi\|_{L^1(]0,1[)}\\ \le\ & C^0 \cdot \int_0^t \|A(s) - B(s)\|_{L^\infty(]0,1[)} ds \cdot \|\varphi\|_{L^1(]0,1[)}.\end{aligned}$$

The same arguments apply for $\mathcal{I}^1_{AB}$, and we deduce that

$$\begin{aligned}\left|\int_0^1\left(\int_0^t j_A(s,x)ds - \int_0^t j_B(s,x)ds\right)\varphi(x)dx\right| &\le |\mathcal{I}^i_{AB}| + |\mathcal{I}^0_{AB}| + |\mathcal{I}^1_{AB}|\\ &\le (C^i + C^0 + C^1)\int_0^t \|A(s)\\ &\quad - B(s)\|_{L^\infty(]0,1[)} ds \cdot \|\varphi\|_{L^1}\end{aligned}$$

$\forall\, \varphi \in L^1(]0,1[)$ bounded, in particular $\forall\, \varphi \in C_0(]0,1[)$. Since $\int_0^t j_A(s,\cdot)ds - \int_0^t j_B(s,\cdot)ds$ belongs to $L^\infty(]0,1[)$ we deduce by density that the previous inequality holds $\forall\, \varphi \in L^1(]0,1[)$, and we have the estimate

$$\left\|\int_0^t j_A(s,\cdot)ds - \int_0^t j_B(s,\cdot)ds\right\|_{L^\infty(]0,1[)} \le C \cdot \int_0^t \|A(s) - B(s)\|_{L^\infty(]0,1[)} ds, \quad 0 \le t \le T,$$

with $C = C^i + C^0 + C^1$ a constant which depends on $\|A\|_{L^1(]0,T[;W^{1,\infty}(]0,1[))}$, $\|B\|_{L^1(]0,T[;W^{1,\infty}(]0,1[))}$, $\|n_0\|_{L^\infty}$, $\|h_0\|_{L^\infty}$, $\|h_1\|_{L^\infty}$, $\|n_0\|_{L^1}$, $\|h_0\|_{L^1}$, $\|h_1\|_{L^1}$ but not on $\|vn_0\|_{L^1}$, $\|vh_0\|_{L^1}$, $\|vh_1\|_{L^1}$ (note also that since h_k are nonincreasing we have $\|vh_k\|_{L^\infty(\mathbb{R}^+_v)} \le \|h_k\|_{L^1(\mathbb{R}^+_v)}$, $k = 0, 1$). $\square$

PROPOSITION 5.6. *Assume that* $A, B \in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$ *are nondecreasing with respect to* x *and that the hypotheses* (H), (H_1), (H_∞) *hold. Then* $\forall\, 0 \le t \le T$ *we have*

$$\|\mathcal{F}A(t) - \mathcal{F}B(t)\|_{L^\infty(]0,1[)} \le 2 \cdot C \cdot \int_0^t \|A(s) - B(s)\|_{L^\infty(]0,1[)} ds$$

with $C = C^i + C^0 + C^1$ *as before.*

Proof. By Remark 5.4 we have

$$\begin{aligned}|\mathcal{F}A(t,x) - \mathcal{F}B(t,x)| &\le \left|\int_0^t j_A(s,x)ds - \int_0^t j_B(s,x)ds\right| + \int_0^1\left|\int_0^t j_A(s,y)ds\right.\\ &\qquad\left. - \int_0^t j_B(s,y)ds\right| dy\\ &\le 2 \cdot \left\|\int_0^t j_A(s,\cdot)ds - \int_0^t j_B(s,\cdot)ds\right\|_{L^\infty(]0,1[)}\\ &\le 2 \cdot C \cdot \int_0^t \|A(s) - B(s)\|_{L^\infty(]0,1[)} ds, \quad 0 \le t \le T. \qquad \square\end{aligned}$$

5.3. Existence and uniqueness of the mild solution.

THEOREM 5.7. *Assume that the hypotheses* (H), (H_1), (H_∞) *hold and* $U_1 - U_0 \in L^\infty(]0,T[)$. *Then there is a unique mild solution* (f,E) *for the* 1D *Vlasov–Poisson initial-boundary value problem. Moreover, we have the estimates*

$$\|\rho_E\|_{L^\infty(]0,T[\,\times\,]0,1[)} \le B(\exp(TA)-1)+C,$$

$$\||j_E|\|_{L^\infty(]0,T[\,\times\,]0,1[)} \le \frac{B^2}{2A}(\exp(TA)-1)^2 + \frac{BC}{A}(\exp(TA)-1)+M_1,$$

$$\|E\|_{L^\infty(]0,T[;W^{1,\infty}(]0,1[))} \le 2B\exp(TA)+C-B,$$

where $A = 6\cdot M_\infty$, $B = M_0 + \|U_1-U_0\|_{L^\infty(]0,T[)}$, $C = M_0$.

Proof. Consider $X_T = \{E \in L^\infty(]0,T[;W^{1,\infty}(]0,1[)) \,|\, \|\partial_x E(t)\|_{L^\infty(]0,1[)} \le B\exp(tA)+C-B,\ \|E(t)\|_{L^\infty(]0,1[)} \le B\exp(tA),\ 0\le t\le T\}$. By Proposition 5.1 and Remark 5.2 we know that $\mathcal{F}: X_T \to X_T$ is well defined, and by Proposition 5.6 there is a constant $C_1 = C_1(M_0, M_\infty, \|U_0-U_1\|_{L^\infty(]0,T[)}, T)$ such that

$$\|\mathcal{F}A(t)-\mathcal{F}B(t)\|_{L^\infty(]0,1[)} \le C_1\cdot\int_0^t \|A(s)-B(s)\|_{L^\infty(]0,1[)}ds,\ \ A,B\in X_T.$$

We deduce that $\mathcal{F}$ has a unique fixed point $E\in X_T$, and therefore (f_E,E) is the unique mild solution of the 1D Vlasov–Poisson initial-boundary value problem. The estimation $|j_E|$ follows by Proposition 5.3. ∎

5.4. Existence and uniqueness of the mild solution in the general case. In this section we study the existence and uniqueness of the mild solution when assuming only the hypotheses (H), (H_0), (H_∞). In order to do this we need only prove that Proposition 5.6 still holds under the above hypotheses. For $\alpha>0$ let us consider the initial-boundary conditions given by

$$\begin{aligned} f_0^\alpha(x,v) &= \frac{f_0(x,v)}{1+\alpha|v|},\ \ (x,v)\in\,]0,1[\,\times\mathbb{R}_v,\\ g_0^\alpha(t,v) &= \frac{g_0(t,v)}{1+\alpha v},\ \ (t,v)\in\,]0,T[\,\times\mathbb{R}_v^+,\\ g_1^\alpha(x,v) &= \frac{g_1(t,v)}{1-\alpha v},\ \ (t,v)\in\,]0,T[\,\times\mathbb{R}_v^-. \end{aligned}$$

It is easy to check that if (H), (H_0), (H_∞) hold, then the same hypotheses (H^α), (H_0^α), (H_∞^α), corresponding to the initial-boundary conditions $f_0^\alpha, g_0^\alpha, g_1^\alpha$, hold with the functions $n_0^\alpha(v) := \frac{n_0(v)}{1+\alpha v}$, $h_k^\alpha(v) := \frac{h_k(v)}{1+\alpha v}$, $v\in\mathbb{R}_v^+$, $k=0,1$, and we have $M_0^\alpha \le M_0 < +\infty$, $M_\infty^\alpha \le M_\infty < +\infty$. Moreover, note also that (H_1^α) is satisfied with $M_1^\alpha \le \frac{M_0}{\alpha} < +\infty$. Since $n_0, h_0, h_1 \in L^1(\mathbb{R}_v^+)$ are nonincreasing we check easily that $n_0^\alpha, h_0^\alpha, h_1^\alpha$ are nonincreasing and

$$\|vh_k^\alpha\|_{L^\infty(\mathbb{R}_v^+)} \le \|vh_k\|_{L^\infty(\mathbb{R}_v^+)} \le \|h_k\|_{L^1(\mathbb{R}_v^+)},\ k=0,1,\alpha>0.$$

PROPOSITION 5.8. *Assume that* $A,B\in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$ *are nondecreasing with respect to* x *and that* $(H),(H_0),(H_\infty)$ *hold. Then* $\forall\, 0\le t\le T$ *we have*

$$\|\mathcal{F}A(t)-\mathcal{F}B(t)\|_{L^\infty(]0,1[)} \le C\cdot\int_0^t \|A(s)-B(s)\|_{L^\infty(]0,1[)}ds,$$

where C *depends only on* $\|A\|_{L^1(]0,T[;W^{1,\infty}(]0,1[))}, \|B\|_{L^1(]0,T[;W^{1,\infty}(]0,1[))}, M_0, M_\infty, T$.

Proof. By Proposition 5.6 we have

$$\|\mathcal{F}^\alpha A(t) - \mathcal{F}^\alpha B(t)\|_{L^\infty(]0,1[)} \le C^\alpha \cdot \int_0^t \|A(s) - B(s)\|_{L^\infty(]0,1[)} ds, \tag{5.3}$$

where $\mathcal{F}^\alpha$ corresponds to the initial-boundary conditions $f_0^\alpha, g_0^\alpha, g_1^\alpha$. $(C^\alpha)_{\alpha>0}$ is bounded since we have

$$\begin{aligned} C^\alpha &= C(\|A\|_{L^1(]0,T[;W^{1,\infty}(]0,1[))}, \|B\|_{L^1(]0,T[;W^{1,\infty}(]0,1[))}, M_0^\alpha, M_\infty^\alpha, T) \\ &\le C(\|A\|_{L^1(]0,T[;W^{1,\infty}(]0,1[))}, \|B\|_{L^1(]0,T[;W^{1,\infty}(]0,1[))}, M_0, M_\infty, T). \end{aligned}$$

The conclusion follows by passing to the limit in inequality (5.3) for $\alpha \to 0$ and by using the monotone convergence theorem. □

Now we can state the existence and uniqueness result in the general case.

THEOREM 5.9. *Assume that the hypotheses* $(H), (H_0), (H_\infty)$ *hold and* $U_1 - U_0 \in L^\infty(]0,T[)$. *Then there is a unique mild solution of the* 1D *Vlasov–Poisson initial-boundary value problem* (f_E, E), *which verifies the estimates*

$$\begin{aligned} \|\partial_x E\|_{L^\infty} = \|\rho_E\|_{L^\infty} &\le (M_0 + \|U_1 - U_0\|_{L^\infty}) \exp(6 \cdot T M_\infty) - \|U_1 - U_0\|_{L^\infty}, \\ \|E\|_{L^\infty} &\le (M_0 + \|U_1 - U_0\|_{L^\infty}) \exp(6 \cdot T M_\infty). \end{aligned}$$

5.5. Continuity upon the initial-boundary conditions. The goal of this section is to estimate the difference between two mild solutions $(f^k, E^k), k = 1,2$, with respect to the initial-boundary conditions. Consider two sets of initial-boundary conditions $f_0^k, g_0^k, g_1^k, U_0^k - U_1^k \in L^\infty$ verifying the hypotheses $(H^k), (H_0^k), (H_\infty^k)$, $k = 1,2$. We define the applications $\mathcal{F}^k$ as before. We have for $t \in [0,T]$

$$\|\partial_x \mathcal{F}^k E(t)\|_{L^\infty} = \|\rho_E^k\|_{L^\infty} \le 6 \cdot M_\infty^k \int_0^t \|E(s)\|_{L^\infty} ds + M_0^k$$

and

$$\|\mathcal{F}^k E(t)\|_{L^\infty} \le 6 \cdot M_\infty^k \int_0^t \|E(s)\|_{L^\infty} ds + M_0^k + |U_0^k(t) - U_1^k(t)|.$$

First of all let us assume the hypotheses (H), (H_1), and (H_∞). We have the following proposition.

PROPOSITION 5.10. *Assume that* $E \in L^\infty(]0,T[;W^{1,\infty}(]0,1[))$ *is nondecreasing with respect to* x *and that the hypotheses* $(H^k), (H_1^k), (H_\infty^k)$ *hold. We suppose also that the functions*

$$(H_{\mathrm{i}}) \;\; l_k(v) = \sup_{0\le t\le T} |g_k^1(t, (-1)^k v) - g_k^2(t, (-1)^k v)|, \; k = 0,1,$$

are nonincreasing with respect to $v \in \mathbb{R}_v^+$, *or*

$$(H_{\mathrm{ii}}) \;\; \int_0^T \int_{v>0} v|g_0^1(t,v) - g_0^2(t,v)| dt dv - \int_0^T \int_{v<0} v|g_1^1(t,v) - g_1^2(t,v)| dt dv < +\infty.$$

Then $\forall\, 0 \le t \le T$ *we have*

$$\begin{aligned} &\|\mathcal{F}^1 E(t) - \mathcal{F}^2 E(t)\|_{L^\infty} \le C(\|E\|_{L^1(]0,t[;L^\infty(]0,1[))}) \\ &\cdot \left(\|f_0^1 - f_0^2\|_{L^1(]0,1[\times\mathbb{R}_v)} + \sum_{k=0}^1 (\|l_k\|_{L^1} + \|l_k\|_{L^\infty}) \right) + |U_1^1(t) - U_0^1(t) - U_1^2(t) + U_0^2(t)| \end{aligned}$$

in case (i) *or*

$$\|\mathcal{F}^1E(t)-\mathcal{F}^2E(t)\|_{L^\infty}\le 2\bigg(\|f_0^1-f_0^2\|_{L^1(]0,1[\times\mathbb{R}_v)}+\sum_{k=0}^{1}\|v(g_k^1-g_k^2)\|_{L^1(]0,t[\times\mathbb{R}_v^+)}\bigg)$$
$$+|U_1^1(t)-U_0^1(t)-U_1^2(t)+U_0^2(t)|$$

in case (ii).

Proof. Consider $\varphi\in L^1(]0,1[)$ and calculate

$$\begin{aligned}
&\int_0^1\!\!\int_0^t(j_E^1(s,x)-j_E^2(s,x))\varphi(x)dxds\\
&\quad=\int_0^t\!\!\int_0^1\!\!\int_{\mathbb{R}_v}(f_E^1-f_E^2)v\varphi(x)dsdxdv\\
&\quad=\int_0^1\!\!\int_{\mathbb{R}_v}(f_0^1(x,v)-f_0^2(x,v))\int_0^{s^i_{out}}V^i(s)\varphi(X^i(s))dxdvds\\
&\qquad+\sum_{k=0}^{1}(-1)^k\int_0^t\!\!\int_{(-1)^kv>0}v(g_k^1-g_k^2)\int_s^{s^k_{out}}V^k(\tau)\varphi(X^k(\tau))dsdvd\tau\\
&\quad=\int_0^1\!\!\int_{\mathbb{R}_v}(f_0^1(x,v)-f_0^2(x,v))\int_x^{X^i(s^i_{out})}\varphi(u)dxdvdu\\
&\qquad+\sum_{k=0}^{1}(-1)^k\int_0^t\!\!\int_{(-1)^kv>0}v(g_k^1(s,v)-g_k^2(s,v))\int_k^{X^k(s^k_{out})}\varphi(u)dsdvdu\\
&\quad=\mathcal{I}^i+\Sigma_{k=0}^1\mathcal{I}^k.
\end{aligned}$$

Obviously we have

$$|\mathcal{I}^i|\le\|f_0^1-f_0^2\|_{L^1(]0,1[\times\mathbb{R}_v)}\cdot\|\varphi\|_{L^1(]0,1[)}.$$

On the other hand, with the notation $\Phi^k=\int_k^{X^k(s^k_{out})}\varphi(u)du$, we have

$$\begin{aligned}
\mathcal{I}^0&=\int_0^t\!\!\int_{\mathbb{R}_v}v(g_0^1(s,v)-g_0^2(s,v))dsdv\Phi^0\mathbf{1}_{\{0<v<v_E^0\}}\\
&\quad+\int_0^t\!\!\int_{\mathbb{R}_v}v(g_0^1(s,v)-g_0^2(s,v))dsdv\Phi^0\mathbf{1}_{\{v_E^0<v<v_E^1\}}\\
&\quad+\int_0^t\!\!\int_{\mathbb{R}_v}v(g_0^1(s,v)-g_0^2(s,v))dsdv\Phi^0\mathbf{1}_{\{v>v_E^1\}}\\
&=\mathcal{I}_0^0+\mathcal{I}_t^0+\mathcal{I}_1^0,
\end{aligned}$$

where $v_E^k = v_E^k(t;s,k)$ are the critical velocities corresponding to the domain $]0,t[\times]0,1[$, the point (s,k), and the field E. Calculate now

$$
\begin{aligned}
&\int_0^1 \left(\int_0^1 \int_0^t (j_E^1(s,y) - j_E^2(s,y)) ds dy \right) \varphi(x) dx \\
&= \int_0^t \int_0^1 \int_{\mathbb{R}_v} v(f_E^1(s,y,v) - f_E^2(s,y,v)) ds dy dv \cdot \int_0^1 \varphi(u) du \\
&= \int_0^1 \varphi(u) du \cdot \Big\{ \int_0^1 \int_{\mathbb{R}_v} (f_0^1 - f_0^2) \int_0^{s^i_{out}} V^i(s) dx dv ds \\
&\qquad + \sum_{k=0}^1 (-1)^k \int_0^t \int_{(-1)^k v > 0} v(g_k^1 - g_k^2) \int_s^{s^k_{out}} V^k(\tau) ds dv d\tau \Big\} \\
&= \int_0^1 \varphi(u) du \cdot \Big\{ \int_0^1 \int_{\mathbb{R}_v} (f_0^1 - f_0^2)(X^i(s^i_{out}) - x) dx dv \\
&\qquad + \sum_{k=0}^1 (-1)^k \int_0^t \int_{(-1)^k v > 0} v(g_k^1 - g_k^2)(X^k(s^k_{out}) - k) ds dv \Big\} \\
&= \mathcal{J}^i + \sum_{k=0}^1 \mathcal{J}^k.
\end{aligned}
$$

Obviously we have $|\mathcal{J}^i| \le \|f_0^1 - f_0^2\|_{L^1(]0,1[\times\mathbb{R}_v)} \cdot \|\varphi\|_{L^1(]0,1[)}$. On the other hand, we have

$$
\begin{aligned}
\mathcal{J}^0 &= \int_0^t \int_{\mathbb{R}_v} v(g_0^1 - g_0^2)(X^0(s^0_{out}) - 0) \mathbf{1}_{\{0<v<v_E^0\}} ds dv \int_0^1 \varphi(u) du \\
&\quad + \int_0^t \int_{\mathbb{R}_v} v(g_0^1 - g_0^2)(X^0(s^0_{out}) - 0) \mathbf{1}_{\{v_E^0<v<v_E^1\}} ds dv \int_0^1 \varphi(u) du \\
&\quad + \int_0^t \int_{\mathbb{R}_v} v(g_0^1 - g_0^2)(X^0(s^0_{out}) - 0) \mathbf{1}_{\{v>v_E^1\}} ds dv \int_0^1 \varphi(u) du \\
&= \mathcal{J}_0^0 + \mathcal{J}_t^0 + \mathcal{J}_1^0.
\end{aligned}
$$

For $0 < v < v_E^0$ we have $X^0(s^0_{out}) = 0$ and thus $\mathcal{I}_0^0 = \mathcal{J}_0^0$. For $v > v_E^1$ we have $X^0(s^0_{out}) = 1$ and thus $\mathcal{I}_1^0 = \mathcal{J}_1^0$. In order to evaluate $\mathcal{I}_t^0$ and $\mathcal{J}_t^0$ we can perform the change of variables $(y,w) = S(s,v)$,

$$
y = X^0(t;s,0,v), \quad w = V^0(t;s,0,v), \quad \left| \frac{\partial(y,w)}{\partial(s,v)} \right| = |v|,
$$

on $D = \{(s,v) \in]0,t[\times\mathbb{R}_v^+ \mid v_E^0(t;s,0) < v < v_E^1(t;s,0)\}$. In case (i) one gets

$$
\begin{aligned}
|\mathcal{I}_t^0| &\le \int_0^1 |\varphi(u)| du \cdot \int_0^t \int_{\mathbb{R}_v} v |g_0^1(s,v) - g_0^2(s,v)| \mathbf{1}_{\{v_E^0<v<v_E^1\}} ds dv \\
&= \int_0^1 |\varphi(u)| du \cdot \int_0^1 \int_{w>-R} |g_0^1 - g_0^2|(s^0_{in}(t,y,w), V^0(s^0_{in}(t,y,w);t,y,w)) \mathbf{1}_{S(D)} dy dw \\
&\le \int_{w>-R} l_0^R(w) dw \cdot \|\varphi\|_{L^1(]0,1[)} = (2R\|l_0\|_{L^\infty(\mathbb{R}_v^+)} + \|l_0\|_{L^1(\mathbb{R}_v^+)}) \cdot \|\varphi\|_{L^1(]0,1[)},
\end{aligned}
$$

where $R = \int_0^t \|E(s)\|_{L^\infty(]0,1[)} ds$. In a similar manner we find that

$$\begin{aligned}|\mathcal{J}_t^0| &\leq \int_0^t \int_{\mathbb{R}_v} v|g_0^1 - g_0^2| \mathbf{1}_{\{v_E^0 < v < v_E^1\}} ds dv \cdot \|\varphi\|_{L^1(]0,1[)} \\ &\leq \int_{w > -R} h_0^R(w) dw \cdot \|\varphi\|_{L^1(]0,1[)}.\end{aligned}$$

Finally, one gets that

$$\begin{aligned}|\mathcal{I} - \mathcal{J}| &= \left| \int_0^1 \left(\int_0^t (j_E^1(s,x) - j_E^2(s,x)) ds - \int_0^1 \int_0^t (j_E^1(s,y) - j_E^2(s,y)) dy ds \right) \varphi(x) dx \right| \\ &= |\mathcal{I}^i + \mathcal{I}^0 + \mathcal{I}^1 - \mathcal{J}^i - \mathcal{J}^0 - \mathcal{J}^1| \\ &\leq |\mathcal{I}^i| + |\mathcal{J}^i| + |\mathcal{I}_t^0| + |\mathcal{J}_t^0| + |\mathcal{I}_t^1| + |\mathcal{J}_t^1| \\ &\leq C(R) \left(\|f_0^1 - f_0^2\|_{L^1} + \sum_{k=0}^1 (\|l_k\|_{L^\infty} + \|l_k\|_{L^1}) \right) \cdot \|\varphi\|_{L^1(]0,1[)},\end{aligned}$$

and the conclusion follows in case (i) by using Remark 5.4. For case (ii) it is sufficient to remark that

$$\max\{|\mathcal{I}_t^k|, |\mathcal{J}_t^k|\} \leq \int_0^t \int_{(-1)^k v > 0} (-1)^k v |g_k^1(s,v) - g_k^2(s,v)| ds dv \cdot \|\varphi\|_{L^1(]0,1[)}, \ k = 0, 1. \quad \square$$

Remark 5.11. The conclusion of Proposition 5.10 still holds if we replace hypothesis (H_1^k) with (H_0^k), $k = 0, 1$ (proceed as in the proof of Proposition 5.8).

PROPOSITION 5.12. *Assume that $E^1, E^2 \in L^\infty(]0,T[; W^{1,\infty}(]0,1[))$ are nondecreasing with respect to x and that $(H^k), (H_0^k), (H_\infty^k)$ hold. We suppose also that (H_{i}) or (H_{ii}) is verified. Then $\forall\ 0 \leq t \leq T$ we have*

$$\|\mathcal{F}^1 E^1(t) - \mathcal{F}^2 E^2(t)\|_{L^\infty(]0,1[)} \leq C_1 + C_2 \int_0^t \|E^1(s) - E^2(s)\|_{L^\infty(]0,1[)} ds,$$

where $C_2 = C_2(\|E^k\|_{L^1(]0,T[; W^{1,\infty}(]0,1[))}, M_0^k, M_\infty^k, T)$ and

$$\begin{aligned}C_1 &= C_1(\|E^k\|_{L^1(]0,T[; L^\infty(]0,1[))}) \left(\|f_0^1 - f_0^2\|_{L^1} + \Sigma_{k=0}^1 (\|l_k\|_{L^1} + \|l_k\|_{L^\infty}) \right) \\ &\quad + |U_1^1 - U_0^1 - U_1^2 + U_0^2|(t)\end{aligned}$$

in case (i) *or*

$$\begin{aligned}C_1 = 2 &\left(\|f_0^1 - f_0^2\|_{L^1} + \|v(g_0^1 - g_0^2)\|_{L^1} + \|v(g_1^1 - g_1^2)\|_{L^1} \right) \\ &+ |U_1^1 - U_0^1 - U_1^2 + U_0^2|(t)\end{aligned}$$

in case (ii).

Proof. We can write

$$\|\mathcal{F}^1 E^1(t) - \mathcal{F}^2 E^2(t)\|_{L^\infty} \leq \|\mathcal{F}^1 E^1(t) - \mathcal{F}^1 E^2(t)\|_{L^\infty} + \|\mathcal{F}^1 E^2(t) - \mathcal{F}^2 E^2(t)\|_{L^\infty}.$$

By using Proposition 5.5 we find

$$\|\mathcal{F}^1 E^1(t) - \mathcal{F}^1 E^2(t)\|_{L^\infty(]0,1[)} \leq C_2 \int_0^t \|E^1(s) - E^2(s)\|_{L^\infty(]0,1[)} ds,$$

where C_2 depends on $\|E^k\|_{L^1(]0,T[;W^{1,\infty}(]0,1[))}, M_0^1, M_\infty^1, T$. The conclusion follows by Proposition 5.10 and Remark 5.11. ☐

THEOREM 5.13. *Assume that* $f_0^k, g_0^k, g_1^k, U_1^k - U_0^k \in L^\infty(]0,T[)$, $k = 1, 2$, *are two sets of initial-boundary conditions verifying the hypotheses* $(H^k), (H_0^k), (H_\infty^k)$, *and* $(H_{\rm i})$ *or* $(H_{\rm ii})$. *Denote by* (f^k, E^k), $k = 1, 2$, *the corresponding unique mild solutions. Then we have* $\forall\, 0 \le t \le T$

$$\|E^1(t) - E^2(t)\|_{L^\infty(]0,1[)} \le C\left\{ \|f_0^1 - f_0^2\|_{L^1} + \sum_{k=0}^{1}(\|l_k\|_{L^1} + \|l_k\|_{L^\infty}) \right. \\ \left. + |U_1^1 - U_0^1 - U_1^2 + U_0^2|(t) \right\}$$

in case (i) *or*

$$\|E^1(t) - E^2(t)\|_{L^\infty(]0,1[)} \le C\left\{ \|f_0^1 - f_0^2\|_{L^1(]0,1[\times\mathbb{R}_v)} + \sum_{k=0}^{1} \|v(g_k^1 - g_k^2)\|_{L^1(]0,T[\times\mathbb{R}_v^\pm)} \right. \\ \left. + |U_1^1 - U_0^1 - U_1^2 + U_0^2|(t) \right\}$$

in case (ii), *where* C *is a constant depending on* $M_0^k, M_\infty^k, \|U_0^k - U_1^k\|_{L^\infty}, T$.

Proof. Since (f^k, E^k) are mild solutions, we have $\mathcal{F}^k E^k = E^k$, E^k are nondecreasing with respect to x, and we know that

$$\|E^k\|_{L^\infty(]0,T[;W^{1,\infty}(]0,1[))} \le C(M_0^k, M_\infty^k, \|U_1^k - U_0^k\|_{L^\infty(]0,T[)}, T).$$

By Proposition 5.12 we have $\forall\, 0 \le t \le T$

$$\begin{aligned}\|E^1(t) - E^2(t)\|_{L^\infty(]0,1[)} &= \|\mathcal{F}^1 E^1(t) - \mathcal{F}^2 E^2(t)\|_{L^\infty(]0,1[)} \\ &\le C_1 + C_2 \int_0^t \|E^1(s) - E^2(s)\|_{L^\infty(]0,1[)} ds,\end{aligned}$$

with C_1, C_2 as before. The conclusion of the theorem follows by using the Gronwall lemma. ☐

6. The 1D Vlasov–Maxwell system. In this section we study the 1D Vlasov–Maxwell system with initial conditions by adapting the method used previously. Since the proofs are quite similar we only sketch them. Moreover, as explained in the introduction, in this case we can consider different species of particles. Recall that results on the existence and uniqueness have already been obtained by Cooper and Klimas [7]. Let us introduce the equations

$$\partial_t f^\pm + v \cdot \partial_x f^\pm \pm E \cdot \partial_v f^\pm = 0, \quad (t, x, v) \in]0, T[\times \mathbb{R}_x \times \mathbb{R}_v, \tag{6.1}$$

$$\begin{aligned}\partial_t E &= -j(t, x) := -j^+ + j^- \\ &= -\int_{\mathbb{R}_v} v(f^+(t, x, v) - f^-(t, x, v)) dv, \quad (t, x) \in]0, T[\times \mathbb{R}_x,\end{aligned} \tag{6.2}$$

with the initial conditions

$$f^\pm(t = 0, x, v) = f_0^\pm(x, v), \quad (x, v) \in \mathbb{R}_x \times \mathbb{R}_v, \tag{6.3}$$

$$E(t = 0, x) = E_0(x) = \int \rho_0(y) dy, \quad x \in \mathbb{R}_x, \tag{6.4}$$

where $\rho_0 = \rho_0^+ - \rho_0^- = \int_{\mathbb{R}_v}(f_0^+ - f_0^-)dv$ and $\int \rho_0(y)dy$ denotes an arbitrary primitive of ρ_0. Assume that $E \in L^\infty(]0,T[;W^{1,\infty}(\mathbb{R}_x)), f_0^\pm \in L^1_{loc}(\mathbb{R}_x \times \mathbb{R}_v)$. We denote by $(X^\pm(s), V^\pm(s))$ the characteristics associated with $\pm E$. As usual we say that $f^\pm \in L^1_{loc}(]0,T[\times \mathbb{R}_x \times \mathbb{R}_v)$ is a mild solution for the Vlasov problem (6.1), (6.3) iff

$$\int_0^T\int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f^\pm(t,x,v)\psi(t,x,v)dtdxdv = \int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f_0^\pm \int_0^T \psi(s, X^\pm(s;0,x,v), \\ \times V^\pm(s;0,x,v))dxdvds$$

for all test functions $\psi \in L^\infty(]0,T[\times \mathbb{R}_x \times \mathbb{R}_v)$ compactly supported in $[0,T]\times\mathbb{R}_x\times\mathbb{R}_v$. Assume now that $f_0^\pm \in L^1(\mathbb{R}_x \times \mathbb{R}_v)$. We say that $(f^\pm, E) \in L^1(]0,T[\times\mathbb{R}_x\times\mathbb{R}_v) \times L^\infty(]0,T[;W^{1,\infty}(\mathbb{R}_x))$ is a mild solution of the 1D Vlasov–Maxwell problem iff $f^\pm$ is a mild solution for the Vlasov problem (6.1), (6.3) corresponding to the electric field $\pm E$ such that

$$\begin{aligned}\int_{\mathbb{R}_x} E(t,x)\varphi(x)dx = &- \int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f_0^+(x,v)\int_x^{X^+(t;0,x,v)} \varphi(u)dudxdv \\ &+ \int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f_0^-(x,v)\int_x^{X^-(t;0,x,v)} \varphi(u)dudxdv \\ &+ \int_{\mathbb{R}_x} E_0(x)\varphi(x)dx \ \ \forall \varphi \in L^1(\mathbb{R}_x).\end{aligned}$$

Remark 6.1. Note that the previous formula defines a unique function $E \in L^\infty(]0,T[\times\mathbb{R}_x)$. This definition can be derived formally from (6.2) by using the mild formulation

$$\begin{aligned}\int_{\mathbb{R}_x} E(t,x)\varphi(x) &= -\int_0^t\int_{\mathbb{R}_x}\int_{\mathbb{R}_v} v(f^+(s,x,v) - f^-(s,x,v))\varphi(x)dsdxdv + \int_{\mathbb{R}_x} E_0(x)\varphi(x)dx \\ &= -\int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f_0^+(x,v)\int_0^t V^+(s)\varphi(X^+(s))dsdxdv + \int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f_0^-(x,v) \\ &\quad \times \int_0^t V^-(s)\varphi(X^-(s))dsdxdv + \int_{\mathbb{R}_x} E_0(x)\varphi(x)dx \\ &= -\int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f_0^+(x,v)\int_x^{X^+(t;0,x,v)} \varphi(u)dudxdv \\ &\quad + \int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f_0^-(x,v)\int_x^{X^-(t;0,x,v)} \varphi(u)dudxdv + \int_{\mathbb{R}_x} E_0(x)\varphi(x)dx.\end{aligned}$$

As before we define the application $\mathcal{F}$ for $E \in L^\infty(]0,T[;W^{1,\infty}(\mathbb{R}_x))$ by

$$E \to f_E^\pm \to E_1(t) = \mathcal{F}E(t) = E_0 - \int_0^t\int_{\mathbb{R}_v} v(f_E^+ - f_E^-)dsdv,$$

where $f_E^\pm$ are the mild solutions of the Vlasov problem (6.1), (6.3) associated with the field $\pm E$, E_0 is given by (6.4), and $-\int_0^t\int_{\mathbb{R}_v} v(f_E^+ - f_E^-)dsdv$ is defined as in Remark 6.1.

6.1. Estimate of $\mathcal{F}$. We assume that there is $n_0^\pm : [0,+\infty[\to [0,+\infty[$ nonincreasing, such that

$$(H^\pm)\ f_0^\pm(x,v) \le n_0^\pm(|v|),\ \ (x,v)\in\mathbb{R}_x\times\mathbb{R}_v,$$
$$(H_0^\pm)\ M_0^\pm := \int_{\mathbb{R}_v} n_0^\pm(|v|)dv < +\infty,$$
$$(H_\infty^\pm)\ M_\infty^\pm := \|n_0^\pm\|_{L^\infty(\mathbb{R}_v^+)} < +\infty,$$
$$(H_{\rho_0})\ M_{\rho_0} := \sup_{x\in\mathbb{R}_x}\left|\int_0^x(\rho_0^+(y)-\rho_0^-(y))dy\right| < +\infty.$$

Proposition 6.2. *Assume that $f_0^\pm\in L^1(\mathbb{R}_x\times\mathbb{R}_v)$ satisfy $(H^\pm),(H_0^\pm),(H_\infty^\pm)$. Then for every $E\in L^\infty(]0,T[;W^{1,\infty}(\mathbb{R}_x))$ we have $f_E^\pm\in L^\infty(]0,T[;L^1(\mathbb{R}_x\times\mathbb{R}_v))$, $\rho_E^\pm\in L^\infty(]0,T[\times\mathbb{R}_x)$, $\mathcal{F}E\in L^\infty(]0,T[;W^{1,\infty}(\mathbb{R}_x))$. Moreover, the following estimates hold:*

$$\|f_E^\pm\|_{L^\infty(]0,T[;L^1(\mathbb{R}_x\times\mathbb{R}_v))} = \|\rho_E^\pm\|_{L^\infty(]0,T[;L^1(\mathbb{R}_x))} = \int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f_0^\pm(x,v)dxdv,$$
$$\|\rho_E^\pm\|_{L^\infty(]0,T[\times\mathbb{R}_x)} \le 2M_\infty^\pm\int_0^t\|E(s)\|_{L^\infty(\mathbb{R}_x)}ds + M_0^\pm,$$
$$\|\mathcal{F}E\|_{L^\infty(]0,T[\times\mathbb{R}_x)} \le C + M_{\rho_0} + \|f_0^+\|_{L^1(\mathbb{R}_x\times\mathbb{R}_v)} + \|f_0^-\|_{L^1(\mathbb{R}_x\times\mathbb{R}_v)},$$
$$\|\partial_x\mathcal{F}E\|_{L^\infty(]0,T[\times\mathbb{R}_x)} \le 2(M_\infty^+ + M_\infty^-)\int_0^t\|E(s)\|_{L^\infty(\mathbb{R}_x)}ds + M_0^+ + M_0^-,$$
$$\lim_{R_1\to+\infty}\int_{|v|>R_1} f_E^\pm(t,x,v)dv = 0 \text{ uniformly with respect to } (t,x)\in]0,T[\times\mathbb{R}_x,$$

and the mild formulation of the Vlasov problem holds $\forall\,\psi\in L^\infty(]0,T[\times\mathbb{R}_x\times\mathbb{R}_v)$.

Proof. We have

$$\begin{aligned}\rho_E^\pm(t,x) &= \int_{\mathbb{R}_v} f_E^\pm dv = \int_{\mathbb{R}_v} f_0^\pm(X^\pm(0;t,x,v),V^\pm(0;t,x,v))dv\\ &\le \int_{\mathbb{R}_v} n_0^\pm(|V^\pm(0;t,x,v)|)dv \le \int_{\mathbb{R}_v} n_0^{\pm,R}(|v|)dv\\ &= 2RM_\infty^\pm + M_0^\pm,\end{aligned}$$

where $R=\int_0^t\|E(s)\|_{L^\infty(\mathbb{R}_x)}ds$. By the definition of $\mathcal{F}E$, taking into account that $E_0(x)=C+\int_0^x\rho_0(y)dy$, we deduce that

$$\|\mathcal{F}E(t)\|_{L^\infty(\mathbb{R}_x)} \le C + M_{\rho_0} + \|f_0^+\|_{ofL^1(\mathbb{R}_x\times\mathbb{R}_v)} + \|f_0^-\|_{L^1(\mathbb{R}_x\times\mathbb{R}_v)},\ \ 0\le t\le T.$$

By using the definition of $\mathcal{F}E(t)$ and the mild formulation we check that $\partial_x\mathcal{F}E(t)=\rho(t)$ in $\mathcal{D}'(\mathbb{R}_x)$, $0\le t\le T$, and we deduce that $\|\partial_x\mathcal{F}E\|_{L^\infty}\le\|\rho_E^+\|_{L^\infty}+\|\rho_E^-\|_{L^\infty}\le 2R(M_\infty^+ + M_\infty^-)+M_0^+ + M_0^-$. The last two assertions follow by standard calculations as was done for the Vlasov–Poisson problem. ∎

Remark 6.3. If we note $X_T=\{E\in L^\infty(]0,T[;W^{1,\infty}(\mathbb{R}_x))\mid \|E\|_{L^\infty(]0,T[\times\mathbb{R}_x)}\le \|E_0\|_{L^\infty(\mathbb{R}_x)}+\|f_0^+\|_{L^1}+\|f_0^-\|_{L^1}\}$, then $\mathcal{F}(X_T)\subset X_T$ and

$$\|\partial_x\mathcal{F}E\|_{L^\infty(]0,T[\times\mathbb{R}_x)} \le 2(M_\infty^+ + M_\infty^-)\cdot T\cdot(\|E_0\|_{L^\infty(\mathbb{R}_x)}+\|f_0^+\|_{L^1}+\|f_0^-\|_{L^1})+M_0^+ + M_0^-.$$

6.2. Estimate of $\mathcal{F} \quad - \mathcal{F} \quad$.

PROPOSITION 6.4. *Assume that* $A, B \in L^\infty(]0,T[; W^{1,\infty}(\mathbb{R}_x))$ *and* $f_0^\pm \in L^1(\mathbb{R}_x \times \mathbb{R}_v)$ *verify the hypotheses* $(H^\pm), (H_0^\pm), (H_\infty^\pm)$. *Then* $\forall\, 0 \le t \le T$ *we have*

$$\|\mathcal{F}A(t) - \mathcal{F}B(t)\|_{L^\infty(\mathbb{R}_x)} \le C \int_0^t \|A(s) - B(s)\|_{L^\infty(\mathbb{R}_x)} ds,$$

with C *a constant depending on* $\|A\|_{L^1(]0,T[;W^{1,\infty}(\mathbb{R}_x))}, \|B\|_{L^1(]0,T[;W^{1,\infty}(\mathbb{R}_x))}, M_0^\pm$, $M_\infty^\pm, T$.

Proof. Take $\varphi \in L^1(\mathbb{R}_x)$ and calculate

$$\begin{aligned}
&\left| \int_{\mathbb{R}_x} (\mathcal{F}A(t,x) - \mathcal{F}B(t,x))\varphi(x)dx \right| \\
&\quad = \left| - \int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f_0^+ \int_{X_B^+(t)}^{X_A^+(t)} \varphi(u)dudxdv + \int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f_0^- \int_{X_B^-(t)}^{X_A^-(t)} \varphi(u)dudxdv \right| \\
&\quad \le \sum_{k=\pm} \int_{\mathbb{R}_u} |\varphi(u)| \int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f_0^k(x,v) \mathbf{1}_{\{|u - X_A^k(t)| < |X_B^k(t) - X_A^k(t)|\}} dudxdv \\
&\quad \le \sum_{k=\pm} \int |\varphi(u)| \int\int f_0^k(X_A^k(0;t,y,w), V_A^k(0;t,y,w)) \mathbf{1}_{\{|u-y| \le C \cdot R\}} \\
&\quad \le \|\varphi\|_{L^1(\mathbb{R}_x)} 2CR \left(M_0^+ + M_0^- + 2\int_0^t \|A(s)\|_{L^\infty} ds (M_\infty^+ + M_\infty^-) \right),
\end{aligned}$$

where $C = \exp(\int_0^t (1 + \|\partial_x B(s)\|_{L^\infty(\mathbb{R}_x)})ds)$ and $R = \int_0^t \|A(s) - B(s)\|_{L^\infty(\mathbb{R}_x)}$ ds. $\square$

We can prove the following theorem by using the iterated approximations method.

THEOREM 6.5. *Assume that* $f_0^\pm \in L^1(\mathbb{R}_x \times \mathbb{R}_v)$ *verify the hypotheses* $(H^\pm), (H_0^\pm)$, $(H_\infty^\pm)$. *Then, for a fixed choice of primitive in* (6.4)*, there is a unique mild solution for the* 1D *Vlasov–Maxwell initial value problem.*

Remark 6.6. If in addition we assume that $|v|^p f_0^\pm \in L^1(\mathbb{R}_x \times \mathbb{R}_v)$ and

$$(H_p^\pm) \;\; M_p^\pm := \int_{\mathbb{R}_v} |v|^p n_0^\pm(|v|)dv < +\infty,$$

for some integer $p \ge 1$ we can prove that

$$|v|^p f^\pm \in L^\infty(]0,T[; L^1(\mathbb{R}_x \times \mathbb{R}_v)), \;\; \int_{\mathbb{R}_v} |v|^p f^\pm(t,x,v)dv \in L^\infty(]0,T[\times \mathbb{R}_x).$$

In particular $j^\pm = \int_{\mathbb{R}_v} v f^\pm dv \in L^\infty(]0,T[\times\mathbb{R}_x)$ and $\partial_t E = -j$, $\lim_{R_1 \to +\infty} \int_{|v|>R_1} |v|^p f^\pm dv = 0$ uniformly with respect to $(t,x) \in]0,T[\times \mathbb{R}_x$ and the mild formulation of the Vlasov problem holds for all functions $|\psi(t,x,v)| \le C(1 + |v|^p)$.

Proof. Multiplying the Vlasov equation by $|v|^p$, we get

$$\frac{d}{dt} \int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f^\pm(t,x,v)|v|^p dxdv = \pm \int_{\mathbb{R}_x}\int_{\mathbb{R}_v} E f^\pm p|v|^{p-2} v dxdv.$$

Therefore we deduce that

$$\begin{aligned}
\int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f^\pm(t,x,v)|v|^p dxdv \le \int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f_0^\pm(x,v)|v|^p dxdv + p\|E\|_{L^\infty(]0,T[\times\mathbb{R}_x)} \\
\cdot \int_0^t \int_{\mathbb{R}_x}\int_{\mathbb{R}_v} f^\pm |v|^{p-1} dxdvds,
\end{aligned}$$

and the conclusion follows by induction on p. On the other hand,

$$\begin{aligned}\int_{\mathbb{R}_v} |v|^p f^{\pm}(t,x,v)dv &= \int_{\mathbb{R}_v} |v|^p f_0^{\pm}(X^{\pm}(0;t,x,v),V^{\pm}(0;t,x,v))dv \\ &\le \int_{\mathbb{R}_v} |v|^p n_0^{\pm,R}(|v|)dv \\ &\le C(R)(\|n_0^{\pm}\|_{L^\infty(\mathbb{R}_v^+)} + \||v|^p n_0^{\pm}(|v|)\|_{L^1(\mathbb{R}_v)}),\end{aligned}$$

with $R = \int_0^t \|E(s)\|_{L^\infty(\mathbb{R}_x)}ds$. In order to verify that $\partial_t E = -j$ in $\mathcal{D}'(]0,T[\times\mathbb{R}_x)$, take $\varphi \in C_0^1(]0,T[\times\mathbb{R}_x)$ and use the mild formulation with the test function $\psi(t,x,v) = v\varphi(t,x)$. $\square$

7. The periodic 1D Vlasov–Poisson problem. In this section we analyze the space periodic 1D Vlasov–Poisson problem:

$$\partial_t f^{\pm} + v\cdot\partial_x f^{\pm} \pm E\cdot\partial_v f^{\pm} = 0,\ \ (t,x,v)\in]0,T[\times]0,1[\times\mathbb{R}_v, \tag{7.1}$$

(7.2)

$$\partial_x E = \rho(t,x) := \rho^+ - \rho^- = \int_{\mathbb{R}_v} (f^+(t,x,v) - f^-(t,x,v))dv,\ \ (t,x)\in]0,T[\times]0,1[,$$

with the space periodic initial conditions

$$f^{\pm}(t=0,x,v) = f_0^{\pm}(x,v),\ (x,v)\in]0,1[\times\mathbb{R}_v. \tag{7.3}$$

The electric field derives from a space periodic potential and thus $\int_0^1 E(t,x)dx = 0$. In this case the Poisson field can be written as

$$E(t,x) = \int_0^x \rho(t,y)dy - \int_0^1 (1-y)\rho(t,y)dy,\ x\in[0,1],\ t\in[0,T]. \tag{7.4}$$

We introduce the mild formulation as before by taking space periodic test functions. This time it is convenient to define the application $\mathcal{F}$ for 1-periodic with respect to x fields $E \in L^\infty(]0,T[\times\mathbb{R}_x)$ by

$$E \to f_E^{\pm} \to \rho_E^{\pm} \to \int_0^x \rho_E(t,y)dy - \int_0^1 (1-y)\rho_E(t,y)dy = \mathcal{F}E.$$

Remark 7.1. $\mathcal{F}E$ is 1-periodic in x iff $\int_0^1 \rho_E(t,y)dy = 0,\ 0\le t\le T$ and therefore, by the conservation of the total charge, iff $\int_0^1\int_{\mathbb{R}_v} f_0^+(x,v)dxdv = \int_0^1\int_{\mathbb{R}_v} f_0^-(x,v)dxdv$.

7.1. Estimate of $\mathcal{F}$. We assume that $f_0^{\pm}$ verify the hypotheses $(H^{\pm})$, $(H_0^{\pm})$, $(H_\infty^{\pm})$. We suppose also that the neutrality condition holds:

$$(N)\ \ \int_0^1\int_{\mathbb{R}_v} f_0^+(x,v)dxdv = \int_0^1\int_{\mathbb{R}_v} f_0^-(x,v)dxdv.$$

Proposition 7.2. *Assume that $f_0^\pm$ are 1-periodic in x and satisfy $(H^\pm), (H_0^\pm)$, $(H_\infty^\pm)$, and (N). Then for every $E \in L^\infty(]0,T[; W^{1,\infty}(\mathbb{R}_x))$ 1-periodic in x we have*

$$\|\rho_E^\pm\|_{L^\infty(]0,T[\,\times\mathbb{R}_x)} \le 2M_\infty^\pm \int_0^t \|E(s)\|_{L^\infty(\mathbb{R}_x)} ds + M_0^\pm,$$

$$\|\mathcal{F}E\|_{L^\infty(]0,T[\,\times\mathbb{R}_x)} \le \int_0^1\!\!\int_{\mathbb{R}_v} f_0^+(x,v)dxdv + \int_0^1\!\!\int_{\mathbb{R}_v} f_0^-(x,v)dxdv \le M_0^+ + M_0^-,$$

$$\|\partial_x\mathcal{F}E\|_{L^\infty(]0,T[\,\times\mathbb{R}_x)} \le 2(M_\infty^+ + M_\infty^-)\int_0^t \|E(s)\|_{L^\infty(\mathbb{R}_x)} ds + M_0^+ + M_0^-.$$

Moreover, $\lim_{R_1\to+\infty}\int_{|v|>R_1} f_E^\pm(t,x,v)dv = 0$ uniformly with respect to $(t,x) \in\,]0,T[\,\times\mathbb{R}_x$ and the mild formulation of the Vlasov problem holds for all functions $\psi \in L^\infty(]0,T[\,\times\mathbb{R}_x\times\mathbb{R}_v)$ 1-periodic in x.

7.2. Estimate of $\mathcal{F}$ $-\,\mathcal{F}$.

Proposition 7.3. *Assume that $A, B \in L^\infty(]0,T[; W^{1,\infty}(\mathbb{R}_x))$ are 1-periodic in x and the hypotheses $(H^\pm), (H_0^\pm), (H_\infty^\pm), (N)$ hold. Then $\forall\, 0 \le t \le T$ we have*

$$\|\mathcal{F}A(t) - \mathcal{F}B(t)\|_{L^\infty(\mathbb{R}_x)} \le C\int_0^t \|A(s)-B(s)\|_{L^\infty(\mathbb{R}_x)} ds,$$

where the constant C depends on $\|A\|_{L^1(]0,T[;W^{1,\infty}(\mathbb{R}_x))}, \|B\|_{L^1(]0,T[;W^{1,\infty}(\mathbb{R}_x))}, M_0^\pm$, $M_\infty^\pm, T$.

Proof. Take $\varphi \in L^1_{loc}(\mathbb{R}_x)$ and calculate

$$\begin{aligned}
\mathcal{I}_1^\pm &= \left|\int_0^1 \varphi(x)\int_0^x (\rho_A^\pm(t,y) - \rho_B^\pm(t,y))dydx\right| \\
&= \Bigg|\iint (f_0^\pm(X_A^\pm(0;t,y,v), V_A^\pm(0;t,y,v)) - f_0^\pm(X_B^\pm(0;t,y,v), V_B^\pm(0;t,y,v))) \\
&\qquad \cdot \int_y^1 \varphi(x)dxdydv\Bigg| \\
&= \left|\iint f_0^\pm(\xi,\eta)\int_{X_A^\pm(t;0,\xi,\eta)}^{X_B^\pm(t;0,\xi,\eta)} \varphi(x)dxd\xi d\eta\right| \\
&\le \int_0^1 |\varphi(u)| \iint f_0^\pm(\xi,\eta)\mathbf{1}_{\{|u-X_A^\pm(t)|<|X_A^\pm(t)-X_B^\pm(t)|\}} d\xi d\eta du \\
&\le \int_0^1 |\varphi(u)| \iint f_0^\pm(X_A^\pm(0;t,y,w), V_A^\pm(0;t,y,w))\mathbf{1}_{\{|u-y|<CR\}} dydw \\
&\le 2CR\left(2\int_0^t \|A(s)\|_{L^\infty} ds \cdot M_\infty^\pm + M_0^\pm\right)\cdot\|\varphi\|_{L^1(]0,1[)},
\end{aligned}$$

where $C = \exp(\int_0^t (1 + \|\partial_x B(s)\|_{L^\infty})ds)$ and $R = \int_0^t \|A(s) - B(s)\|_{L^\infty} ds$. In order to estimate $\mathcal{I}_2^\pm = |\int_0^1 (1-y)(\rho_A^\pm(t,y) - \rho_B^\pm(t,y))dy|$ take $\varphi \equiv 1$ in the previous computation. ∎

Finally we obtain the existence and uniqueness of the space periodic mild solution.

Theorem 7.4. *Assume that $f_0^\pm$ are 1-periodic in x and satisfy the hypotheses $(H^\pm), (H_0^\pm)$, $(H_\infty^\pm)$, (N). Then there is a unique mild solution for the space periodic*

1D *Vlasov–Poisson problem. Moreover, we have the estimates*

$$\|\rho^{\pm}\|_{L^\infty(]0,T[\,\times\mathbb{R}_x)} \le 2M_\infty^{\pm}\cdot T\cdot(M_0^+ + M_0^-) + M_0^{\pm},$$
$$\|E\|_{L^\infty(]0,T[\,\times\mathbb{R}_x)} \le M_0^+ + M_0^-.$$

REFERENCES

[1] A. Arseneev, *Global existence of a weak solution of the Vlasov system of equations*, U.R.S.S. Comput. Math. Phys., 15 (1975), pp. 131–143.

[2] J. Batt, *Global symmetric solutions of the initial value problem of stellar dynamics*, J. Differential Equations, 25 (1977), pp. 342–364.

[3] C. Bardos and P. Degond, *Global existence for the Vlasov–Poisson equation in* 3 *space variables with small initial data*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 2 (1985), pp. 101–118.

[4] N. Ben Abdallah, *Weak solutions of the initial-boundary value problem for the Vlasov-Poisson system*, Math. Methods Appl. Sci., 17 (1994), pp. 451–476.

[5] M. Bostan, *Solutions périodiques en temps des équations de Vlasov-Maxwell*, C. R. Acad. Sci. Paris, 339 (2004), pp. 451–456.

[6] M. Bostan, *Permanent regimes for the* 1*D Vlasov–Poisson system with boundary conditions*, SIAM J. Math. Anal., 35 (2003), pp. 922–948.

[7] J. Cooper and A. Klimas, *Boundary value problem for the Vlasov–Maxwell equation in one dimension*, J. Math. Anal. Appl., 75 (1980), pp. 306–329.

[8] P. Degond and P.-A. Raviart, *An asymptotic analysis of the one-dimensional Vlasov-Poisson system: The Child-Langmuir law*, Asymptotic Anal., 4 (1991) pp. 187–214.

[9] R. J. DiPerna and P. L. Lions, *Global weak solutions of Vlasov–Maxwell system*, Comm. Pure Appl. Math., 42 (1989), pp. 729–757.

[10] R. Glassey and J. Schaeffer, *The two and one-half dimensional relativistic Vlasov–Maxwell system*, Comm. Math. Phys., 185 (1997), pp. 257–284.

[11] C. Greengard and P.-A. Raviart, *A boundary value problem for the stationary Vlasov–Poisson equations: The plane diode*, Comm. Pure Appl. Math., 43 (1990), pp. 473–507.

[12] Y. Guo, *Global weak solutions of the Vlasov–Maxwell system with boundary conditions*, Comm. Math. Phys., 154 (1993), pp. 245–263.

[13] Y. Guo, *Singular solutions of the Vlasov-Maxwell system on a half line*, Arch. Rational Mech. Anal., 131 (1995), pp. 241–304.

[14] Y. Guo, C.-W. Shu, and T. Zhou, *The dynamics of a plane diode*, SIAM J. Math. Anal., 35 (2004), pp. 1617–1635.

[15] E. Horst, *On the classical solutions of the initial value problem for the unmodified nonlinear Vlasov equation*, Math. Methods Appl. Sci., 3 (1981), pp. 229–248.

[16] E. Horst and R. Hunze, *Weak solutions of the initial value problem for the unmodified nonlinear Vlasov equation*, Math. Methods Appl. Sci., 6 (1984), pp. 262–279.

[17] P.-L. Lions and B. Perthame, *Propagation of moments and regularity for the* 3*-dimensional Vlasov–Poisson system*, Invent. Math., 105 (1991), pp. 415–430.

[18] K. Pfaffelmoser, *Global classical solutions of the Vlasov-Poisson system in three dimensions for general initial data*, J. Differential Equations, 95 (1992), pp. 281–303.

[19] F. Poupaud, *Boundary value problems for the stationary Vlasov–Maxwell system*, Forum Math., 4 (1992), pp. 499–527.

[20] R. Robert, *Unicité de la solution faible à support compact de l'équation de Vlasov–Poisson*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 873–877.

[21] T. Ukai and S. Okabe, *On the classical solution in the large time of the two dimensional Vlasov equations*, Osaka J. Math., 15 (1978), pp. 245–261.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 189–199

THE NONLINEAR SCHRÖDINGER EQUATION WITH A STRONGLY ANISOTROPIC HARMONIC POTENTIAL*

NAOUFEL BEN ABDALLAH†, FLORIAN MÉHATS†, CHRISTIAN SCHMEISER‡, AND RADA M. WEISHÄUPL§

Abstract. The nonlinear Schrödinger equation with general nonlinearity of polynomial growth and harmonic confining potential is considered. More precisely, the confining potential is strongly anisotropic; i.e., the trap frequencies in different directions are of different orders of magnitude. The limit as the ratio of trap frequencies tends to zero is carried out. A concentration of mass on the ground state of the dominating harmonic oscillator is shown to be propagated, and the lower-dimensional modulation wave function again satisfies a nonlinear Schrödinger equation. The main tools of the analysis are energy and Strichartz estimates, as well as two anisotropic Sobolev inequalities. As an application, the dimension reduction of the three-dimensional Gross–Pitaevskii equation is discussed, which models the dynamics of Bose–Einstein condensates.

Key words. energy estimates, Strichartz estimates, anisotropic Sobolev embedding, Gross–Pitaevskii equation, Bose–Einstein condensate

AMS subject classification. 35Q55

DOI. 10.1137/040614554

1. Introduction and main result. The main goal of the present work is the analysis of the space dimension reduction of the $(n+d)$-dimensional nonlinear Schrödinger equation with external confining potential. We are interested in the case where the external potential is anisotropic and strongly confining in d directions. This work follows the approach used in [3] for analyzing a dimension reduction (from dimension 3 to dimension 2) for the Schrödinger–Poisson system and where asymptotics for strong partial confinement was introduced. In other words, we deal with the asymptotic behavior of solutions of the $(n+d)$-dimensional Schrödinger equation

$$i\psi_t = -\frac{1}{2}\Delta\psi + V^\varepsilon(x,z)\psi + f(\delta|\psi|)\psi\,, \tag{1}$$

$$\psi_I(x,z) = \psi(0,x,z)\,, \tag{2}$$

$$V^\varepsilon(x,z) = \frac{|x|^2}{2} + \frac{|z|^2}{2\varepsilon^4}\,, \qquad x\in R^n\,, \quad z\in R^d\,,$$

*Received by the editors September 7, 2004; accepted for publication (in revised form) February 1, 2005; published electronically August 31, 2005. This work was supported by the European IHP network RNT2 2001 349, "Hyperbolic and kinetic equations: Asymptotics, numerics, applications."
http://www.siam.org/journals/sima/37-1/61455.html

†MIP, Laboratoire CNRS (UMR 5640), Université Paul Sabatier,118, route de Narbonne, 31062 Toulouse Cedex 04, France (naoufel@mip.ups-tlse.fr, mehats@mip.ups-tlse.fr). The work of the first and second authors was supported by the ACI Nouvelles Interfaces des Mathématiques ACINIM 176-2004, "MOQUA," and funded by the French ministry of research, as well as the ACI Jeunes chercheurs JC1035 "Modèles dispersifs vectoriels pour le transport à l'échelle nanométrique."

‡Institut für Analysis und Scientific Computing, TU Wien, Wiedner Hauptstraße 8–10, 1040 Wien, Austria, and Johann Radon Institute for Computational and Applied Mathematics, Altenbergerstraße 69, 4040 Linz, Austria (schmeise@deana.math.tuwien.ac.at). The work of this author was supported by the Ph.D. program "Differential Equations" funded by the Austrian Science Fund, project W8, and through project P16174-N05, Austrian Science Fund.

§Institut für Mathematik, Universität Wien, Nordbergstraße 15, 1090 Wien, Austria (rada.weishaeupl@univie.ac.at). The work of this author was supported by the Ph.D. program "Differential Equations" funded by the Austrian Science Fund, project W, and by the "Wittgenstein 2000" Award of P. Markowich.

as $\varepsilon \to 0$. Here V^ε is the trapping harmonic potential and $\Delta = \Delta_x + \Delta_z$ is the Laplacian in R^{n+d} and ψ satisfies the normalization condition

$$\int_{R^{n+d}} |\psi|^2 dx dz = 1\,, \tag{3}$$

which is preserved by (1). Since the sign of the function f is not specified, we are dealing with both focusing and defocusing nonlinearities. Performing the limit $\varepsilon \to 0$ in this system will enable us to write a reduced model involving a nonlinear Schrödinger equation in dimension n. In section 4, an application to the dynamics of Bose–Einstein condensates is presented; we justify mathematically the effective models which can be found in the physics literature [13]. In this context, as was remarked in [1], the use of such approximate models significantly reduces the complexity of numerical simulations.

In order to balance the kinetic and potential energy terms in the z-direction, we introduce the rescaling $z \to \varepsilon z$. In order to keep the wavefunction normalized we have to rescale by $\psi \to \varepsilon^{-d/2}\psi^\varepsilon(t,x,z)$. As we want to balance the nonlinearity with the terms of order 1, we choose $\delta = \varepsilon^{d/2}$; thus we consider weak nonlinearities. The rescaled problem reads

$$\begin{aligned} &i\psi_t^\varepsilon = H^\perp \psi^\varepsilon + \frac{1}{\varepsilon^2} H\psi^\varepsilon + f(|\psi^\varepsilon|)\psi^\varepsilon\,, \\ &\psi^\varepsilon(t=0,x,z) = \psi_I(x,z)\,, \end{aligned} \tag{4}$$

with $H^\perp = -\frac{1}{2}\Delta_x + \frac{|x|^2}{2}$ and $H = -\frac{1}{2}\Delta_z + \frac{|z|^2}{2}$, harmonic oscillators in the x- and z-directions, respectively.

We introduce a new time scale $\tau = t/\varepsilon^2$, so that we have the fast oscillations in z corresponding to the fast time scale τ. If we let $\varepsilon \to 0$, we formally obtain the equation

$$i\Psi_\tau = H\Psi\,,$$

which we can solve explicitly in terms of the spectral decomposition of H:

$$\Psi = \sum_{k\geq 0} \phi_k e^{-i\mu_k \tau} \omega_k(z)\,.$$

Here $(\mu_k, \omega_k(z))_{k\geq 0}$ are the eigenvalues and normalized (with respect to $L^2(R^d)$) eigenfunctions of H, and $(\phi_k)_{(k\geq 0)}$ are coefficients independent of τ and z. The eigenvalue problem can be solved explicitly with the eigenvalues $\mu_k = k + \frac{d}{2}$ (see [18, Theorem 8.4]). The eigenfunctions are products of a Gaussian with Hermite polynomials, and, in particular, the ground state eigenfunction is given by

$$\omega_0(z) = \left(\frac{1}{\pi}\right)^{d/4} e^{-\frac{|z|^2}{2}}\,.$$

By modulation, thus introducing the slow variables x and t, we would have ϕ_k depending on (t,x). This motivates us to expand ψ^ε with respect to the eigenstates of H:

$$\psi^\varepsilon(t,x,z) = \sum_{k\geq 0} e^{-i\mu_k t/\varepsilon^2} \phi_k^\varepsilon(t,x)\omega_k(z)\,. \tag{5}$$

Actually, our aim is to determine and justify approximations of the form

$$\psi^\varepsilon(t,x,z) \approx \varphi(t,x)e^{-i\mu_0 t/\varepsilon^2}\omega_0(z)\,, \tag{6}$$

i.e., modulations of the ground state, under an assumption of well-prepared initial data (see (11) below). A formal analysis indicates that the general case, where the transport occurs on several modes, is more complicated and might involve coupling terms between the limiting n-dimensional Schrödinger equations (this is not the case for the Schrödinger–Poisson system [3], where the nonlinearity is weaker).

The projection Π onto the eigenspace generated by the groundstate $\omega_0(z)$ is given by

$$\Pi\psi^\varepsilon(t,x,z) = e^{-i\mu_0 t/\varepsilon^2}\phi^\varepsilon(t,x)\omega_0(z)$$

with

$$\phi^\varepsilon(x,t) := e^{i\mu_0 t/\varepsilon^2}\int_{R^d}\psi^\varepsilon(t,x,z)\omega_0(z)dz\,. \tag{7}$$

It is obvious that the projection has the following properties:

$$\partial_t\Pi = \Pi\partial_t\,, \qquad \Pi H^\perp = H^\perp\Pi\,, \qquad \Pi H = \mu_0\Pi\,.$$

By projecting (4) we obtain

$$i\phi^\varepsilon_t = H^\perp\phi^\varepsilon + e^{i\mu_0 t/\varepsilon^2}\int_{R^d} f(|\psi^\varepsilon|)\psi^\varepsilon\omega_0 dz\,. \tag{8}$$

The nonlinearity can be written as

$$\begin{aligned} & e^{i\mu_0 t/\varepsilon^2}\int_{R^d} f(|\psi^\varepsilon|)\psi^\varepsilon\omega_0 dz = \overline{f}(|\phi^\varepsilon|)\phi^\varepsilon + h^\varepsilon \\ & \text{with} \quad \overline{f}(|\phi|) = \int_{R^d} f(|\phi|\omega_0)\omega_0^2 dz \\ & \text{and} \quad h^\varepsilon = e^{i\mu_0 t/\varepsilon^2}\int_{R^d}[f(|\psi^\varepsilon|)\psi^\varepsilon\omega_0 - f(|\phi^\varepsilon|\omega_0)e^{-i\mu_0 t/\varepsilon^2}\phi^\varepsilon\omega_0^2]dz\,. \end{aligned} \tag{9}$$

Then the formal limit of (8) as $\varepsilon \to 0$ is the n-dimensional Schrödinger equation

$$i\varphi_t = H^\perp\varphi + \overline{f}(|\varphi|)\varphi\,. \tag{10}$$

When the initial data for the full problem (4) are chosen compatible with the ansatz (6), i.e.,

$$\psi_I(x,z) = \varphi_I(x)\omega_0(z)\,, \tag{11}$$

then appropriate initial conditions for the solution of (10) are

$$\varphi(0,x) = \varphi_I(x)\,. \tag{12}$$

The main result of this work is a justification of the limit problem (10), (12) under the following assumptions on the initial data and on the nonlinearity.

Assumption 1. The function φ_I satisfies

$$\int_{R^n} \left(|\nabla_x \varphi_I(x)|^2 + |x\varphi_I(x)|^2\right) dx < \infty, \quad \int_{R^n} |\varphi_I(x)|^2 dx = 1\,.$$

Assumption 2. The nonlinearity f satisfies

$$|f(|u|)u - f(|v|)v)| \le C(|u|^\alpha + |v|^\alpha)|u-v|\,,$$

where either $f \ge 0$ (defocusing case) and $0 \le \alpha < \frac{4}{n+d-2}$, or $0 \le \alpha < \min\{\frac{4}{n+d-2}, \frac{4}{n}\}$. Additionally, $\alpha \le \frac{2}{n-2}$ if $n > 2$.

Remark. The assumptions are sufficient for proving existence and uniqueness of local solutions of both the full problem (4), (11) and the limit problem (10), (12) (see [6, 15, 5]). Note that the property of f required in Assumption 2 carries over to $\overline{f}$. In the repulsive case, global existence is a straightforward consequence of energy conservation (see section 2). Without sign assumptions on the nonlinearity, the additional requirement $\alpha < 4/n$ leads to global solvability of the limit problem [15]. Here, however, it is used for proving ε-independent estimates for the full problem on finite time intervals.

THEOREM 1. *Let Assumptions* 1 *and* 2 *be satisfied and let* ψ^ε *and* φ *be the unique solutions of* (4), (11) *and* (10), (12), *respectively. Then for every* $T < \infty$ *there exists a constant* c_T *such that*

$$\sup_{t\in(0,T)} \|\psi^\varepsilon(t,\cdot,\cdot)) - e^{-i\mu_0 t/\varepsilon^2}\varphi(t,\cdot)\omega_0\|_{L^2(R^{n+d})} \le c_T\,\varepsilon\,.$$

The rest of the paper is organized as follows. In the following section, conservation of energy is used to derive uniform estimates of H^1-norms of the solution of the $(n+d)$-dimensional problem and its ground state contribution. Whereas for repulsive nonlinearities these results follow directly from the energy conservation, in the general case the nonlinearity needs to be controlled by an anisotropic generalization of the Gagliardo–Nirenberg inequality. Also the difference between the full solution and its projection to the ground state is shown to be small. In section 3, the difference between the ground state contribution and its formal limit is estimated. The main tools are Strichartz estimates [6, 10, 17] and an anisotropic Sobolev inequality.

Section 4 deals with an application, the Gross–Pitaevskii equation, which has a cubic nonlinearity and models the dynamics of Bose–Einstein condensates. In this case, dimension reduction means obtaining disk-shaped or cigar-shaped condensates. Finally, in the appendix the anisotropic Sobolev embedding and the anisotropic Gagliardo–Nirenberg inequality are proved.

2. Uniform estimates. In this section we derive some ε-independent estimates from energy conservation. The energy is defined by

$$E^\varepsilon[\psi^\varepsilon(t)] := \left\langle H^\perp\psi^\varepsilon(t), \psi^\varepsilon(t)\right\rangle + \frac{1}{\varepsilon^2}\left\langle H\psi^\varepsilon(t), \psi^\varepsilon(t)\right\rangle + 2\mathcal{F}[\psi^\varepsilon(t)]\,,$$

where $\langle\cdot,\cdot\rangle$ denotes the scalar product in $L^2(R^{n+d})$ and

$$\mathcal{F}[\psi] = \int_{R^{n+d}} F(|\psi|)dx\,dz\,, \quad \text{with } F(s) = \int_0^s f(\sigma)\sigma d\sigma\,.$$

Note that the first two terms in the energy are nonnegative quadratic forms controlling the H^1-norms in the x- and z-directions, respectively.

With Assumption 1, the initial data (11) satisfy $E^\varepsilon[\psi_I] < \infty$ for fixed ε. From [6] (Theorem 9.2.5 and Remark 9.2.7) and Assumption 2 we obtain local-in-time existence for the $(n+d)$-dimensional problem (4) as well as energy and mass conservation:

$$E^\varepsilon[\psi^\varepsilon(t)] = E^\varepsilon[\psi_I]\,, \qquad \|\psi^\varepsilon(t)\|_{2,2} = \|\psi_I\|_{2,2} = \|\varphi_I\|_2\,. \tag{13}$$

Considering the limit of $\varepsilon^2 E^\varepsilon$ when $\varepsilon \to 0$, we immediately obtain uniform bounds for the dominant term. The main difficulty consists in finding uniform bounds on $\left\langle H^\perp \psi^\varepsilon(t), \psi^\varepsilon(t)\right\rangle$. Once we have this, we can derive uniform bounds on the H^1-norm of $\psi^\varepsilon(t)$.

For the notation of norms we use the following conventions.

Definition 2. *Let $0 < T \le \infty$, $1 \le p, q, r \le \infty$, and $u(t,x)$, $v(t,x,z)$ functions of $t \in (0,T)$, $x \in R^n$, and $z \in R^d$. Then we define the norms*

$$\begin{aligned} \|u(t)\|_p &:= \|u(t,\cdot)\|_{L^p(R^n)}\,, \\ \|u\|_{r(p)} &:= \Big\| \|u(\cdot)\|_p \Big\|_{L^r((0,T))}\,, \\ \|v(t)\|_{q,p} &:= \Big\| \|v(t,\cdot)\|_p \Big\|_{L^q(R^d)}\,, \\ \|v\|_{r(q,p)} &:= \Big\| \|v(\cdot)\|_{q,p} \Big\|_{L^r((0,T))}\,, \end{aligned}$$

and the corresponding Banach spaces are denoted by L^p_x, $L^r_t L^p_x$, $L^q_z L^p_x$, and $L^r_t L^q_z L^p_x$.

Taking into account the expansion (5) of the $(n+d)$-dimensional wavefunction ψ^ε with respect to the orthonormal basis $(\omega_k)_{k\ge 0}$ of eigenfunctions gives

$$\|\psi^\varepsilon(t)\|_{2,2}^2 = \sum_{k=0}^{\infty} \|\phi_k^\varepsilon(t)\|_2^2, \tag{14}$$

$$\|\nabla_x \psi^\varepsilon(t)\|_{2,2}^2 = \sum_{k=0}^{\infty} \|\nabla_x \phi_k^\varepsilon(t)\|_2^2. \tag{15}$$

At first sight, the energy equation seems to be of limited use, since it is dominated by the contributions in the z-direction. However, with the mass conservation this part can be written as

$$\begin{aligned} \langle H\psi^\varepsilon(t), \psi^\varepsilon(t)\rangle &= \sum_{k=0}^{\infty} \mu_k \|\phi_k^\varepsilon(t)\|_2^2 \\ &= \sum_{k=1}^{\infty} (\mu_k - \mu_0)\|\phi_k^\varepsilon(t)\|_2^2 + \mu_0 \|\varphi_I\|_2^2\,, \end{aligned} \tag{16}$$

and, on the other hand,

$$\langle H\psi_I, \psi_I\rangle = \mu_0 \|\varphi_I\|_2^2\,. \tag{17}$$

By using (16) and (17) we can rewrite the energy conservation as

$$\begin{aligned} &\left\langle H^\perp \psi^\varepsilon(t), \psi^\varepsilon(t)\right\rangle + \frac{1}{\varepsilon^2}\sum_{k=1}^{\infty}(\mu_k - \mu_0)\|\phi_k^\varepsilon(t)\|_2^2 + 2\mathcal{F}[\psi^\varepsilon(t)] \\ &= \left\langle H^\perp \psi_I, \psi_I\right\rangle + 2\mathcal{F}[\psi_I]\,. \end{aligned} \tag{18}$$

In the case of defocusing nonlinearities all terms in this equation are nonnegative, and we immediately obtain uniform boundedness of $\psi^\varepsilon(t)$ in $H^1(R^{n+d})$, as well as the statement that the mass remains concentrated to the ground state as $\varepsilon \to 0$. The rest of this section is devoted to proving the same results (Lemmas 3 and 4) without sign assumption on the nonlinearity.

By applying Lemma 5 from the appendix with $r = \alpha + 2$, we can control the term coming from the nonlinearity:

$$|\mathcal{F}[\psi^\varepsilon(t)]| \le \|\psi^\varepsilon(t)\|_{2+\alpha,2+\alpha}^{2+\alpha} \le c\|\nabla_x \psi^\varepsilon(t)\|_{2,2}^{n\alpha/2} \|\nabla_z \psi^\varepsilon(t)\|_{2,2}^{d\alpha/2}\,, \tag{19}$$

where here and in the following c denotes possibly different ε-independent, positive constants. Consequently, the energy conservation multiplied by ε^2 yields

$$\varepsilon^2 \|\nabla_x \psi^\varepsilon(t)\|_{2,2}^2 + \|\nabla_z \psi^\varepsilon(t)\|_{2,2}^2 \le c + c\varepsilon^2 \|\nabla_x \psi^\varepsilon(t)\|_{2,2}^{n\alpha/2} \|\nabla_z \psi^\varepsilon(t)\|_{2,2}^{d\alpha/2}\,,$$

and, from the Young inequality,

$$\varepsilon^2 \|\nabla_x \psi^\varepsilon(t)\|_{2,2}^2 + \|\nabla_z \psi^\varepsilon(t)\|_{2,2}^2 \le c + \varepsilon^2 \eta \|\nabla_x \psi^\varepsilon(t)\|_{2,2}^2 + \varepsilon^2 C(\eta) \|\nabla_z \psi^\varepsilon(t)\|_{2,2}^{\frac{2d\alpha}{4-n\alpha}}\,.$$

Remark. The constraint $\alpha < \frac{4}{n}$ in Assumption 2 guarantees that the exponent remains positive.

With the choice $\eta = \frac{1}{2}$ we deduce

$$\|\nabla_z \psi^\varepsilon(t)\|_{2,2}^2 \le c + \varepsilon^2 c \|\nabla_z \psi^\varepsilon(t)\|_{2,2}^{\frac{2d\alpha}{4-n\alpha}}\,.$$

In the case $\theta = \frac{2d\alpha}{4-n\alpha} < 2$ we conclude that

$$\|\nabla_z \psi^\varepsilon\|_{\infty(2,2)} \le c. \tag{20}$$

For $\theta > 2$ we obtain the result by the standard bootstrap argument, since $\|\nabla_z \psi_I\|_{2,2}$ is independent of ε (see [4, Lemma 2.9]). Using (19) with (20) in (18), we get

$$\|\nabla_x \psi^\varepsilon(t)\|_{2,2}^2 + \|x\psi^\varepsilon(t)\|_{2,2}^2 + \frac{1}{\varepsilon^2}\sum_{k=1}^{\infty} (\mu_k - \mu_0)\|\phi_k^\varepsilon(t)\|_2^2 \le c + c\|\nabla_x \psi^\varepsilon(t)\|_{2,2}^{n\alpha/2}\,. \tag{21}$$

Since, by $\alpha < 4/n$, the exponent in the last term is smaller than 2, uniform boundedness of $\|\nabla_x \psi^\varepsilon(t)\|_{2,2}$ follows.

It is now easy to prove the following two results on uniform boundedness and on the uniform smallness of the contributions from excited states.

LEMMA 3. *Let the assumptions of Theorem 1 be satisfied, let ψ^ε be the solution of* (4), (11), *let ϕ^ε be defined by* (7), *and let φ be the solution of* (10), (12). *Then*

$$\psi^\varepsilon \in L^\infty((0,\infty);\, H^1(R^{n+d}))\,, \qquad \phi^\varepsilon, \varphi \in L^\infty((0,\infty);\, H^1(R^n))\,,$$

uniformly in ε.

Proof. From (16) and (21) it is immediately clear that $\langle H^\perp \psi^\varepsilon(t), \psi^\varepsilon(t)\rangle + \langle H\psi^\varepsilon(t), \psi^\varepsilon(t)\rangle$ is uniformly bounded with respect to ε and t. The observation that this term dominates the $H^1(R^{n+d})$-norm completes the proof of the first statement of the lemma.

The representation of ψ^ε in terms of the eigenstates shows

$$\langle H^\perp \psi^\varepsilon(t), \psi^\varepsilon(t)\rangle \ge \frac{1}{2}(\|\nabla_x \phi^\varepsilon\|_2^2 + \|x\phi^\varepsilon\|_2^2)\,,$$

which proves the statement for ϕ^ε. Finally, the statement for the ε-independent φ is a consequence of the existence theory. ∎

LEMMA 4. *With the assumptions of the previous lemma,*

$$\|(I-\Pi)\psi^\varepsilon\|_{\infty(p,2)} \le c\,\varepsilon$$

holds with an ε-independent constant c and with $p \in [2, \frac{2d}{d-2}]$ if $d \ge 3$, $p \in [2,\infty)$ if $d = 2$, and $p \in [2,\infty]$ if $d = 1$.

Proof. Using (21) we obtain

$$\begin{aligned}\|(I-\Pi)\psi^\varepsilon(t)\|_{2,2}^2 &= \sum_{k=1}^\infty \|\phi_k^\varepsilon(t)\|_2^2 \\ &\le \frac{1}{\mu_1-\mu_0}\sum_{k=1}^\infty (\mu_k-\mu_0)\|\phi_k^\varepsilon(t)\|_2^2 \le c\varepsilon^2\,,\end{aligned}$$

i.e., the statement of the lemma with $p = 2$. On the other hand we estimate

$$\begin{aligned}\|\nabla_z(I-\Pi)\psi^\varepsilon(t)\|_{2,2}^2 &\le \langle H(I-\Pi)\psi^\varepsilon(t), (I-\Pi)\psi^\varepsilon(t)\rangle = \sum_{k=1}^\infty \mu_k\|\phi_k^\varepsilon(t)\|_2^2 \\ &= \sum_{k=1}^\infty (\mu_k-\mu_0)\|\phi_k^\varepsilon(t)\|_2^2 + \mu_0\|(I-\Pi)\psi^\varepsilon(t)\|_{2,2}^2 \le c\varepsilon^2\,.\end{aligned}$$

The result is now a consequence of the Sobolev embedding $H^1(R^d) \hookrightarrow L^p(R^d)$ in z-space. ∎

3. Proof of the main result. The approximation error in Theorem 1 can be split into two parts:

$$\begin{aligned}\|\psi^\varepsilon - \varphi\omega_0 e^{-i\mu_0 t/\varepsilon^2}\|_{\infty(2,2)} &\le \|(I-\Pi)\psi^\varepsilon\|_{\infty(2,2)} + \|\omega_0 e^{-i\mu_0 t/\varepsilon^2}(\phi^\varepsilon-\varphi)\|_{\infty(2,2)} \\ &= \|(I-\Pi)\psi^\varepsilon\|_{\infty(2,2)} + \|\phi^\varepsilon-\varphi\|_{\infty(2)}\,.\end{aligned}$$

The first term is taken care of by Lemma 4. The difference $\chi^\varepsilon := \phi^\varepsilon - \varphi$ solves the problem

$$\begin{aligned}&i\chi_t^\varepsilon = H^\perp\chi^\varepsilon + g^\varepsilon + h^\varepsilon, \\ &\chi^\varepsilon(t=0) = 0\,,\end{aligned} \tag{22}$$

where

$$g^\varepsilon = \overline{f}(|\phi^\varepsilon|)\phi^\varepsilon - \overline{f}(|\varphi|)\varphi$$

and h^ε given by (9).

For the nonlinear Schrödinger equation (22) with harmonic potential, a local dispersion result can be established (see [8, 9], [6, Lemma 9.2.4]). This property allows us to use Strichartz estimates (see [6, Theorem 3.4.1], [5, 12]), and we obtain the following for any admissible pair (q^*, q) and a *bounded* time interval $T < \infty$:

$$\|\chi^\varepsilon\|_{\infty(2)} \le c_T(\|g^\varepsilon\|_{q^*(q)} + \|h^\varepsilon\|_{q^*(q)})\,. \tag{23}$$

A pair (q, q^*) is admissible iff

$$\frac{2n}{n+2} \le q \le 2 \text{ for } n \ge 3\,, \quad 1 < q \le 2 \text{ for } n = 2\,, \quad 1 \le q \le 2 \text{ for } n = 1\,, \tag{24}$$

$$q^* = \frac{4}{4 - n(2/q - 1)}\,. \tag{25}$$

Note that the definition of admissible pair is not the usual one.

Remark. We need a bounded time interval because the constant depends on the length of the time interval. For more details, see [6].

Assumption 2 implies the pointwise estimate

$$|g^\varepsilon| \le c(|\phi^\varepsilon|^\alpha + |\varphi|^\alpha)|\chi^\varepsilon|\,.$$

Applying the Hölder inequality, we obtain

$$\|g^\varepsilon(t)\|_q \le c(\|\phi^\varepsilon\|^\alpha_{2\alpha q/(2-q)} + \|\varphi\|^\alpha_{2\alpha q/(2-q)})\|\chi^\varepsilon\|_2\,.$$

The assumption $\alpha \le 2/(n-2)$ for $n \ge 3$ allows us to choose q such that both (24) is satisfied and $H^1(R^n) \hookrightarrow L_x^{2\alpha q/(2-q)}$. Therefore we can use Lemma 3 to obtain

$$\|g^\varepsilon(t)\|_{q^*(q)} \le c\|\chi^\varepsilon\|_{q^*(2)}\,. \tag{26}$$

For h^ε we also employ Assumption 2 to obtain a pointwise estimate:

$$|h^\varepsilon| \le c\int_{R^d} (|\psi^\varepsilon|^\alpha + |\Pi\psi^\varepsilon|^\alpha)|(I-\Pi)\psi^\varepsilon|\omega_0\,dz\,.$$

Computing the $L^q(R^n)$-norm and applying the Hölder inequality twice (to the x- and z-integrals, respectively) lead to

$$\|h^\varepsilon\|_q \le c(\|\psi^\varepsilon\|^\alpha_{\alpha p', 2\alpha q/(2-q)} + \|\phi^\varepsilon\|^\alpha_{2\alpha q/(2-q)})\|(I-\Pi)\psi^\varepsilon\|_{p,2}\,,$$

whereby $p' = \frac{p}{p-1}$.

Let us recall all the conditions on p and q:

(i) the assumptions of Lemma 4 for p and condition (24) for q are satisfied;

(ii) the embeddings $H^1(R^n) \hookrightarrow L_x^{2\alpha q/(2-q)}$ and $H^1(R^{n+d}) \hookrightarrow L_z^{\alpha p'} L_x^{2\alpha q/(2-q)}$ (see Lemma 5 in the appendix) hold.

All this is possible since $\alpha \le 4/(n+d-2)$ and $\alpha \le 2/(n-2)$ for $n \ge 3$. As a consequence of Lemmas 3 and 4 we obtain

$$\|h^\varepsilon\|_{\infty(q)} \le c\varepsilon\,. \tag{27}$$

With (26) and (27), the Strichartz estimate (23) becomes

$$\|\chi^\varepsilon\|_{\infty(2)} \le c_T(\|\chi^\varepsilon\|_{q^*(2)} + \varepsilon)\,.$$

Using this estimate on the time interval $(0, t)$ with $t \le T$ gives

$$\|\chi^\varepsilon(t)\|_2^{q^*} \le \tilde{c}_T \left(\int_0^t \|\chi^\varepsilon(s)\|_2^{q^*} ds + \varepsilon^{q^*} \right)\,.$$

Now, an application of the Gronwall lemma concludes the proof of Theorem 1.

4. Application: The Gross–Pitaevskii equation. The three-dimensional nonlinear Schrödinger equation with cubic nonlinearity and an external potential is called the Gross–Pitaevskii equation. It models the temporal evolution of Bose–Einstein condensates at temperatures much smaller than the critical condensation temperature [7, 14, 16]. In dimensional form, the Gross–Pitaevskii equation reads

$$i\hbar\psi_t = -\frac{\hbar^2}{2m}\Delta\psi + \frac{m}{2}\left(\omega_x^2|x|^2 + \omega_z^2|z|^2\right)\psi + Ng|\psi|^2\psi\,, \tag{28}$$

where m is the atomic mass, $\hbar$ is the Planck constant, N is the number of atoms in the condensate, and ω_x, ω_z are the trap frequencies in x- and z-directions, respectively. The parameter g describes the interaction between the atoms in the condensate and has the form $g = \hbar^2 a/m$, where a is the scattering length, positive for repulsive interactions and negative for attractive interactions. We consider the cases $n = 1$ and $n = 2$ with $d = 3 - n$. Characteristic lengths of the condensate in the x- and z-directions are $a_x = \sqrt{\hbar/(m\omega_x)}$ and $a_z = \sqrt{\hbar/(m\omega_z)}$, respectively.

Let us write (28) in dimensionless form. With the scaling

$$x = a_x\tilde{x}\,, \quad z = a_z\tilde{z}\,, \quad \psi = \frac{\tilde{\psi}}{\sqrt{a_x^n a_z^{3-n}}}\,, \quad t = \frac{\tilde{t}}{\omega_x}\,,$$

and skipping the tildes, we obtain

$$i\psi_t = -\frac{1}{2}\Delta_x\psi + \frac{|x|^2}{2}\psi + \frac{\omega_z}{\omega_x}\left(-\frac{1}{2}\Delta_z\psi + \frac{|z|^2}{2}\psi\right) + N\frac{a}{a_z^{3-n}a_x^{n-2}}|\psi|^2\psi\,.$$

In experiments it is observed that in a strongly anisotropic confinement the motion of particles is quenched in one or two directions. This means that by changing the shape of the confining potential, lower-dimensional Bose–Einstein condensates are obtained. They are called disk-shaped or cigar-shaped condensates, respectively. This is the motivation to consider the Gross–Pitaevskii equation with strongly anisotropic confining harmonic potential; thus

$$\varepsilon^2 := \frac{\omega_x}{\omega_z} \ll 1\,.$$

Furthermore we assume the case of weak coupling, namely,

$$\gamma := \frac{Na}{a_z^{3-n}a_x^{n-2}} = \mathcal{O}(1)\,.$$

We then have the equation

$$i\psi_t = -\frac{1}{2}\Delta_x\psi + \frac{|x|^2}{2}\psi + \frac{1}{\varepsilon^2}\left(-\frac{1}{2}\Delta_z\psi + \frac{|z|^2}{2}\psi\right) + \gamma|\psi|^2\psi\,,$$

where $\gamma|\psi|^2 = f(|\psi|)$ with γ positive, if we consider repulsive interactions, e.g., for ^{23}Na and ^{87}Rb, or negative for attractive interactions, e.g., for ^{7}Li. Obviously, Assumption 2 on f holds with $\alpha = 2$.

For *repulsive interactions* ($\gamma > 0$) we have global existence of the solution of the $(n+d)$-dimensional Schrödinger equation if $\alpha < 4/(n+d-2)$. Since $\alpha = 2$ we obtain the condition $n+d<4$, which includes the physically interesting case $n+d=3$. The limiting lower-dimensional Gross–Pitaevskii equation is

$$i\varphi_t = H^{\perp}\varphi + \gamma_0|\varphi|^2\varphi \quad \text{with } \gamma_0 = \gamma\int_{R^d}\omega_0^4(z)dz\,.$$

On the one hand, if we consider the strong confinement in one direction ($d=1$), we obtain a two-dimensional approximate equation ($n=2$). In this case we speak about a *disk-shaped condensate*. On the other hand, we consider a strong confinement in 2 dimensions ($d=2$). Accordingly, the approximate equation is one-dimensional ($n=1$) and we call the condensate a *cigar-shaped condensate*. Theorem 1 can be applied in both cases.

In the case of *attractive interactions*, thus for $\gamma<0$, we get stronger constraints on the dimensions, namely, $n=1$ and $d<3$. Thus, Theorem 1 can only be applied for the reduction from three dimensions to one (cigar-shaped condensate).

Appendix. Anisotropic Sobolev inequalities. In this section, we state anisotropic Sobolev embeddings and a generalized Gagliardo–Nirenberg inequality. The proof of this lemma, rather straightforward, is skipped. It uses standard Sobolev embeddings and Gagliardo–Nirenberg inequalities, combined with interpolation estimates. We generalize here a result of [2] (see also [11], where a similar Sobolev embedding is obtained). Recall that in this paper the whole dimension is $n+d$ and the space variable is written (x,z), where $x\in R^n$ and $z\in R^d$.

LEMMA 5. *Let $2\le p,q\le\infty$ be such that*

$$\frac{n}{p(n+d)}+\frac{d}{q(n+d)}\ge\frac{1}{2}-\frac{1}{n+d}$$

(with $q<\infty$ if $d=2$; $p<\infty$ if $n=2$; and strict inequality if $n=d=1$). Then

$$H^1(R^{n+d})\hookrightarrow L^q_z(R^d;L^p_x(R^n)).$$

Furthermore, for any $r\in[2,\frac{2(n+d)}{n+d-2}]$ we have

$$\|u\|_{L^r_{x,z}}\le C\|u\|_{L^2_{x,z}}^{1-(n+d)(\frac{1}{2}-\frac{1}{r})}\|\nabla_x u\|_{L^2_{x,z}}^{n(\frac{1}{2}-\frac{1}{r})}\|\nabla_z u\|_{L^2_{x,z}}^{d(\frac{1}{2}-\frac{1}{r})}\,.$$

REFERENCES

[1] W. BAO, D. JAKSCH, AND P. A. MARKOWICH, *Numerical solution of the Gross–Pitaevskii equation for Bose–Einstein condensation*, J. Comput. Phys., 187 (2003), pp. 318–342.

[2] N. BEN ABDALLAH AND F. MÉHATS, *On a Vlasov–Schrödinger–Poisson model*, Comm. Partial Differential Equations, 29 (2004), pp. 173–206.

[3] N. BEN ABDALLAH, F. MÉHATS, AND O. PINAUD, *Adiabatic approximation of the Schrödinger–Poisson system with a partial confinement*, SIAM J. Math. Anal., 36 (2005), pp. 986–1013.

[4] R. CARLES, *Nonlinear Schrödinger equations with repulsive harmonic potential and applications*, SIAM J. Math. Anal., 35 (2003), pp. 823–843.

[5] R. CARLES, *Remarks on nonlinear Schrödinger equations with harmonic potential*, Ann. Henri Poincaré, 3 (2002), pp. 757–772.

[6] T. CAZENAVE, *Semilinear Schrödinger Equations*, Courant Lect. Notes Math. 10, AMS, Providence, RI, 2003.

[7] F. DALFOVO, S. GIORGINI, L. P. PITAEVSKII, AND S. STRINGARI, *Theory of Bose–Einstein condensation in trapped gases*, Rev. Mod. Phys., 71 (1999), pp. 463–512.

[8] D. FUJIWARA, *A construction of the fundamental solution for the Schrödinger equation*, J. Analyse Math., 35 (1979), pp. 41–96.

[9] D. Fujiwara, *Remarks on the convergence of the Feynman path integrals*, Duke Math. J., 47 (1980), pp. 559–600.

[10] J. Ginibre and G. Velo, *The global Cauchy problem for the nonlinear Schrödinger equation revisited*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 2 (1985), pp. 309–327.

[11] D. Iftimie and G. Raugel, *Some results on the Navier-Stokes equations in thin* 3*D domains*, J. Differential Equations, 169 (2001), pp. 281–331.

[12] M. Keel and T. Tao, *Endpoint Strichartz estimates*, Amer. J. Math., 120 (1998), pp. 955–980.

[13] P. Leboeuf and N. Pavloff, *Bose–Einstein beams: Coherent propagation through a guide*, Phys. Rev. A, 64 (2001), article 033602.

[14] E. H. Lieb, R. Seiringer, and J. Yngvason, *One-dimensional behaviour of dilute, trapped Bose gases*, Comm. Math. Phys., 244 (2004), pp. 347–393.

[15] Y. G. Oh, *Cauchy problem and Ehrenfest's law of nonlinear Schrödinger equations with potentials*, J. Differential Equations, 81 (1989), pp. 255–274.

[16] L. Pitaevskii and S. Stringari, *Bose–Einstein Condensation*, Internat. Ser. Monogr. Phys. 116, Oxford Science Publications, Oxford, UK, 2003.

[17] R. S. Strichartz, *Restriction of Fourier transform to quadratic surfaces and decay of solutions of wave equations*, Duke Math. J., 44 (1977), pp. 705–714.

[18] G. Teschl, *Mathematical Methods in Quantum Mechanics, with Applications to Schrödinger operators*, Lecture Notes, University of Vienna, 2005; available online from http://www.mat.univie.ac.at/~gerald/ftp/book-schroe.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 200–217

STABLE DETERMINATION OF AN INCLUSION BY BOUNDARY MEASUREMENTS*

G. ALESSANDRINI† AND M. DI CRISTO‡

Abstract. We deal with the problem of determining an inclusion within an electrical conductor from electrical boundary measurements. Under mild a priori assumptions we establish an optimal stability estimate.

Key words. Dirichlet-to-Neumann map, discontinuous conductivity, optimal stability

AMS subject classifications. 35R30, 35A08

DOI. 10.1137/S003614100444191X

1. Introduction.

In this paper we deal with an inverse boundary value problem which is a special instance of the well-known Calderón's inverse conductivity problem [C]. Given a bounded domain Ω in $\mathbb{R}^n$, $n \geq 2$, with reasonably smooth boundary, an open set D, compactly contained in Ω, and a constant $k > 0$, $k \neq 1$, consider, for any $f \in H^{1/2}(\partial\Omega)$, the weak solution $u \in H^1(\Omega)$ to the Dirichlet problem

$$\operatorname{div}((1 + (k-1)\chi_D)\nabla u) = 0 \qquad \text{in } \Omega, \tag{1.1}$$

$$u = f \qquad \text{on } \partial\Omega, \tag{1.2}$$

where χ_D denotes the characteristic function of the set D. We will denote by $\Lambda_D : H^{1/2}(\partial\Omega) \to H^{-1/2}(\partial\Omega)$ the so-called Dirichlet-to-Neumann map, that is the operator which maps the Dirichlet data onto the corresponding Neumann data $\frac{\partial u}{\partial \nu}_{|\partial\Omega}$. The inverse problem that we examine here is to determine D when Λ_D is given.

In 1988 Isakov [I1] proved the uniqueness. The purpose of the present paper is to prove a result of stability. In fact we prove that, under mild a priori assumptions on the regularity and topology of D, there is a continuous dependence of D (in the Hausdorff metric) from Λ_D with a modulus of continuity of logarithmic type; see Theorem 2.4. Let us stress that, indeed, this rate of continuity is the optimal one, as it was shown by examples in the recent paper [DC-R] by the second author and Luca Rondi.

We wish to mention here a closely related, but different, problem which attracted a lot of attention starting with the papers of Friedman [F] and Friedman and Gustafsson [F-G]. That is the one of determining D when, instead of full knowledge of the Dirichlet-to-Neumann map, only one, or few, pairs of Dirichlet and Neumann data are available; see [A-I], [I2] for extended bibliographical accounts. Unfortunately, for such a problem, the uniqueness question, not to mention stability, remains a largely open issue.

Let us illustrate briefly the main steps of our arguments. We must recall that Isakov's approach to uniqueness is essentially based on two arguments,

*Received by the editors March 10, 2004; accepted for publication (in revised form) December 13, 2004; published electronically September 20, 2005. This research was supported by MIUR, grant 2002013279.

http://www.siam.org/journals/sima/37-1/44191.html

†Dipartimento di Scienze Matematiche, Università degli Studi di Trieste, 34100 Trieste, Italy (alessang@univ.trieste.it).

‡Dipartimento di Matematica, Politecnico di Milano, 20133 Milano, Italy (dicristo@mate.polimi.it).

(a) the Runge approximation theorem,

(b) the use of solutions with Green's function type singularities.

Also, here we shall use singular solutions, and indeed we shall need an accurate study of their asymptotic behavior when the singularity gets close to the set of discontinuity ∂D of the conductivity coefficient $1+(k-1)\chi_D$ in (1.1); see Proposition 3.4. On the other hand, it seems that Runge's theorem, in its whole generality, is typically based on nonconstructive arguments (Lax [L], Kohn and Vogelius [K-V]) and thus, is not suited for stability estimates. We wish to mention here that a constructive version of a Runge type theorem, for special geometrical configurations and for Helmholtz equation, has indeed been used by Potthast [P1, P2] for the purpose of stability estimates in the different, but related, inverse problem of obstacle scattering. Unfortunately, by this approach, very restrictive geometrical assumptions on the unknown obstacle are required. In this paper we introduce a different approach based on quantitative estimates of unique continuation (see Proposition 3.5) which avoids the use of Runge type arguments and allows one to treat a wide generality of inclusions.

In section 2 we formulate our main hypotheses and state the stability result, Theorem 2.4. In section 3 we prove Theorem 2.4 on the basis of some auxiliary propositions, whose proofs are deferred to in section 4.

2. The main result. Let us introduce our regularity and topological assumptions on the conductor Ω and on the unknown inclusion D. To this purpose we shall need the following definitions. In places, we shall denote a point $x\in\mathbb{R}^n$ by $x=(x',x_n)$, where $x'\in\mathbb{R}^{n-1}$, $x_n\in\mathbb{R}$.

Definition 2.1. *Let Ω be a bounded domain in $\mathbb{R}^n$. Given α, $0<\alpha\le 1$, we shall say that a portion S of $\partial\Omega$ is of class $C^{1,\alpha}$ with constants $\overline{r}$, $L>0$ if, for any $P\in S$, there exists a rigid transformation of coordinates under which we have $P=0$ and*

$$\Omega\cap B_{\overline{r}}(0)=\{x\in B_{\overline{r}}\,:\,x_n>\varphi(x')\},$$

where φ is a $C^{1,\alpha}$ function on $B_{\overline{r}}(0)\subset\mathbb{R}^{n-1}$ satisfying $\varphi(0)=|\nabla\varphi(0)|=0$ and $\|\varphi\|_{C^{1,\alpha}(B_{\overline{r}}(0))}\le L\overline{r}$.

Definition 2.2. *We shall say that a portion S of $\partial\Omega$ is of Lipschitz class with constants $\overline{r}$, $L>0$ if for any $P\in S$, there exists a rigid transformation of coordinates under which we have $P=0$ and*

$$\Omega\cap B_{\overline{r}}(0)=\{x\in B_{\overline{r}}\,:\,x_n>\varphi(x')\},$$

where φ is a Lipschitz continuous function on $B_{\overline{r}}(0)\subset\mathbb{R}^{n-1}$ satisfying $\varphi(0)=0$ and $\|\varphi\|_{C^{0,1}(B_{\overline{r}}(0))}\le L\overline{r}$.

Remark 2.3. We have chosen to scale all norms in a such a way that they are dimensionally equivalent to their argument. For instance, for any $\varphi\in C^{1,\alpha}(B_{\overline{r}}(0))$ we set

$$\|\varphi\|_{C^{1,\alpha}(B_{\overline{r}}(0))}=\|\varphi\|_{L^\infty(B_{\overline{r}}(0))}+\overline{r}\|\nabla\varphi\|_{L^\infty(B_{\overline{r}}(0))}+\overline{r}^{1+\alpha}|\nabla\varphi|_{\alpha,B_{\overline{r}}(0)}.$$

For given numbers $\overline{r}$, M, $\widetilde{\delta}$, $L>0$, $0<\alpha<1$, we shall assume

(H1) the domain Ω satisfies the following conditions:

$$|\Omega|\le M\overline{r}^n, \tag{2.1}$$

where $|\cdot|$ denotes the Lebesgue measure of Ω,

$$\partial\Omega \text{ is of class } C^{1,\alpha} \text{ with constants } \overline{r},\ L, \tag{2.2}$$

(H2) the inclusion D satisfies the following conditions:

$$\Omega \setminus \overline{D} \quad \text{is connected,} \tag{2.3}$$

$$\text{dist}(D, \partial\Omega) \geq \widetilde{\delta}, \tag{2.4}$$

$$\partial D \text{ is of class } C^{1,\alpha} \text{ with constants } \overline{r},\ L. \tag{2.5}$$

In what follows, we shall refer to numbers k, n, $\overline{r}$, M, $\widetilde{\delta}$, L, α as to the a priori data. We shall denote by D_1 and D_2 two possible inclusions in Ω, both satisfying the properties mentioned. We shall denote by Λ_{D_i}, $i = 1, 2$, the Dirichlet-to-Neumann map Λ_D when $D = D_i$. We can now state the main theorem.

Theorem 2.4. *Let $\Omega \subset \mathbb{R}^n$, $n \geq 2$, satisfy* (H1). *Let $k > 0$, $k \neq 1$ be given. Let D_1 and D_2 be two inclusions in Ω satisfying* (H2). *If, given $\varepsilon > 0$, we have*

$$\|\Lambda_{D_1} - \Lambda_{D_2}\|_{\mathcal{L}(H^{1/2}, H^{-1/2})} \leq \varepsilon, \tag{2.6}$$

then

$$d_{\mathcal{H}}(\partial D_1, \partial D_2) \leq \omega(\varepsilon),$$

where ω is an increasing function on $[0, +\infty)$, which satisfies

$$\omega(t) \leq C|\log t|^{-\eta}, \qquad \text{for every} \quad 0 < t < 1$$

and C, η, $C > 0$, $0 < \eta \leq 1$, are constants only depending on the a priori data.

Here $d_{\mathcal{H}}$ denotes the Hausdorff distance between bounded closed sets of $\mathbb{R}^n$ and $\|\cdot\|_{\mathcal{L}(H^{1/2}H^{-1/2})}$ denotes the operator norm on the space of bounded linear operators between $H^{1/2}(\partial\Omega)$ and $H^{-1/2}(\partial\Omega)$.

Remark 2.5. It should be emphasized that in this statement the unknown inclusion may be disconnected.

Remark 2.6. Several variations of the above results could be devised with minor adaptations on the arguments. Just to mention one, an analogous result would be obtained if the Neumann-to-Dirichlet maps N_{D_i} are available instead of the Dirichlet-to-Neumann maps Λ_{D_i}.

3. Proof of Theorem 2.4. Before proving Theorem 2.4, we shall state some auxiliary propositions, whose proofs are collected in section 4. Here and in what follows, we shall deal with various sets associated to Ω, D_1, and D_2. We shall use the following notation.

Definition 3.1. *We denote by $\mathcal{G}$ the connected component of $\Omega \setminus (D_1 \cup D_2)$, whose boundary contains $\partial\Omega$, $\Omega_D = \Omega \setminus \overline{\mathcal{G}}$, $\Omega_{\overline{r}} = \{x \in \mathcal{C}\Omega : dist(x, \Omega) \leq \overline{r}\}$ and $\mathcal{S}_{2\overline{r}} = \{x \in \mathbb{R}^n : \overline{r} \leq dist(x, \Omega) \leq 2\overline{r}\}$.*

We introduce a variation of the Hausdorff distance which we call *modified distance*.

Definition 3.2. *We shall call the modified distance between D_1 and D_2 the number*

$$d_\mu(D_1, D_2) = \max\left\{ \sup_{x \in \partial D_1 \cap \partial\Omega_D} \text{dist}(x, D_2), \sup_{x \in \partial D_2 \cap \partial\Omega_D} \text{dist}(x, D_1) \right\}. \tag{3.1}$$

This notion is an adaptation of the one introduced in [A-B-R-V], which was also called *modified distance*. In order to distinguish such notions, we call d_μ the present one, whereas the one in [A-B-R-V] was denoted by d_m. On the other hand, we need to stress the common peculiarities: such modified distances do not satisfy the axioms of a metric and in general do not dominate the Hausdorff distance (see section 3 in [A-B-R-V] for related arguments). The following proposition provides sufficient conditions under which d_μ dominates $d_{\mathcal{H}}$; see [A-B-R-V, Proposition 3.6] for a related statement.

PROPOSITION 3.3. *Let Ω be an open set in $\mathbb{R}^n$ satisfying* (H1). *Let D_1, D_2 be two bounded open inclusions of Ω satisfying* (H2). *Then*

$$d_{\mathcal{H}}(\partial D_1, \partial D_2) \le c d_\mu(D_1, D_2), \tag{3.2}$$

where c depends only on the a priori assumptions.

With no loss of generality, we can assume that there exists a point O of $\partial D_1 \cap \partial\Omega_D$, where the maximum in Definition 3.1 is attained, that is

$$d_\mu = d_\mu(D_1, D_2) = \operatorname{dist}(O, D_2). \tag{3.3}$$

As is well known, the Dirichlet-to-Neumann map Λ_D associated to problem (1.1), (1.2) is defined by

$$< \Lambda_D u, v >= \int_\Omega (1 + (k-1)\chi_D)\nabla u \cdot \nabla v, \tag{3.4}$$

for every $u \in H^1(\Omega)$ solution to (1.1) and for every $v \in H^1(\Omega)$. Here $< \cdot, \cdot >$ denotes the dual pairing between $H^{-1/2}(\partial\Omega)$ and $H^{1/2}(\partial\Omega)$. With a slight abuse of notation we shall write

$$< g, f >= \int_{\partial\Omega} g f \, d\sigma,$$

for any $f \in H^{1/2}(\partial\Omega)$ and $g \in H^{-1/2}(\partial\Omega)$. Let $\Gamma_D(x, y)$ be the fundamental solution for the operator $\operatorname{div}((1 + (k-1)\chi_D)\nabla \cdot)$, thus

$$\operatorname{div}((1 + (k-1)\chi_D)\nabla\Gamma_D(\cdot, y)) = -\delta(\cdot - y), \tag{3.5}$$

where $y, w \in \mathbb{R}^n$, δ denotes the Dirac distribution. We shall denote by Γ_{D_1}, Γ_{D_2} such fundamental solutions when $D = D_1$, D_2, respectively. Recalling the well-known identity

$$\int_\Omega (1 + (k-1)\chi_{D_1})\nabla u_1 \cdot \nabla u_2 - \int_\Omega (1 + (k-1)\chi_{D_2})\nabla u_1 \cdot \nabla u_2 = \int_{\partial\Omega} u_1 [\Lambda_{D_1} - \Lambda_{D_2}] u_2,$$

which holds for every $u_i \in H^1(\Omega)$, $i = 1, 2$, solutions to (1.1) when $D = D_i$, respectively (see [I2] formula (5.0.4), section 5.0), we have

$$\begin{aligned} &\textstyle\int_\Omega (1 + (k-1)\chi_{D_1})\nabla\Gamma_{D_1}(\cdot, y) \cdot \nabla\Gamma_{D_2}(\cdot, w) \\ &\textstyle- \int_\Omega (1 + (k-1)\chi_{D_2})\nabla\Gamma_{D_1}(\cdot, y) \cdot \nabla\Gamma_{D_2}(\cdot, w) \\ = &\textstyle\int_{\partial\Omega} \Gamma_{D_1}(\cdot, y)[\Lambda_{D_1} - \Lambda_{D_2}](\Gamma_{D_2}(\cdot, w)) d\sigma \qquad \forall\, y, w \in \mathcal{C}\overline{\Omega}. \end{aligned} \tag{3.6}$$

Let us define, for $y, w \in \mathcal{G} \cup \mathcal{C}\Omega$

$$S_{D_1}(y, w) = (k-1) \int_{D_1} \nabla\Gamma_{D_1}(\cdot, y) \cdot \nabla\Gamma_{D_2}(\cdot, w), \tag{3.7}$$

$$S_{D_2}(y, w) = (k-1) \int_{D_2} \nabla\Gamma_{D_1}(\cdot, y) \cdot \nabla\Gamma_{D_2}(\cdot, w), \tag{3.8}$$

$$f(y, w) = S_{D_1}(y, w) - S_{D_2}(y, w). \tag{3.9}$$

Thus (3.6) can be rewritten as

$$f(y,w) = \int_{\partial\Omega} \Gamma_{D_1}(\cdot, y)[\Lambda_{D_1} - \Lambda_{D_2}](\Gamma_{D_2}(\cdot, w))d\sigma \qquad \forall\, y, w \in \mathcal{C}\overline{\Omega}. \tag{3.10}$$

From now on we shall consider the dimension $n \geq 3$, since the case $n = 2$ can be treated similarly through minor adaptations regarding the fundamental solutions. Up to a transformation of coordinates, we can assume that O, defined in (3.3), is the origin of the coordinate system. Let $\nu(O)$ be the outer unit normal vector to $\partial\Omega_D$ in the origin O. Such a normal is indeed well defined since we are assuming that O realizes the modified distance between D_1 and D_2; therefore, in a small neighborhood of O, $\partial\Omega_D$ is made of a part of ∂D_1, which is known to be $C^{1,\alpha}$. We will rotate the coordinate system in such a way that $\nu(O) = (0, \dots, 0, -1)$. Taking $y = w = h\nu(O)$, with $h > 0$, we want to evaluate $f(y,y)$ and $S_{D_1}(y,y)$ in term of h, for h small. Then, evaluating S_{D_2} in term of d_μ, we will get the stability estimate for the modified distance and thus, using Proposition 3.3, for the Hausdorff distance. An important ingredient for evaluating f and S_{D_1} is the behavior of the fundamental solution. We state now a proposition that collects all the results on Γ_{D_i}, $i = 1, 2$, that we will need throughout the paper. For $x = (x', x_n)$, where $x' \in \mathbb{R}^{n-1}$ and $x_n \in \mathbb{R}$, we set $x^\star = (x', -x_n)$. We shall denote with χ^+ the characteristic function of the half-space $\{x_n > 0\}$ and with Γ_+ the fundamental solution of the operator $\mathrm{div}((1+(k-1)\chi^+)\nabla\cdot)$. If Γ is the standard fundamental solution of the Laplace operator, we have that (see, for instance, [A-I-P], Theorem 4)

$$\Gamma_+(x,y) = \begin{cases} \frac{1}{k}\Gamma(x,y) + \frac{k-1}{k(k+1)}\Gamma(x,y^\star) & \text{for } x_n > 0, y_n > 0, \\ \frac{2}{k+1}\Gamma(x,y) & \text{for } x_n y_n < 0, \\ \Gamma(x,y) - \frac{k-1}{k+1}\Gamma(x,y^\star) & \text{for } x_n < 0, y_n < 0. \end{cases} \tag{3.11}$$

The following proposition holds.

Proposition 3.4. *Let $D \subset \mathbb{R}^n$ be an open set whose boundary is of class $C^{1,\alpha}$, with constants $\overline{r}$, L.*

(i) *There exists a constant $c_1 > 0$ depending on k, n, α and L only, such that*

$$|\nabla_x \Gamma_D(x,y)| \leq c_1 |x-y|^{1-n}, \tag{3.12}$$

for every $x, y \in \mathbb{R}^n$,

(ii) *There exist constants c_2, $c_3 > 0$ depending on k, n, α and L only, such that*

$$\left|\Gamma_D(x,y) - \Gamma_+(x,y)\right| \leq \frac{c_2}{\overline{r}^\alpha}|x-y|^{2-n+\alpha}, \tag{3.13}$$

$$\left|\nabla_x\Gamma_D(x,y) - \nabla_x\Gamma_+(x,y)\right| \leq \frac{c_3}{\overline{r}^{\alpha^2}}|x-y|^{1-n+\alpha^2}, \tag{3.14}$$

for every $x \in D \cap B_r(O)$, and for every $y = h\nu(O)$, with $0 < r < \overline{r}_0$, $0 < h < \overline{r}_0$, where $\overline{r}_0 = \left(\min\left\{\frac{1}{2}(8L)^{-1/\alpha}, \frac{1}{2}\right\}\right)\frac{\overline{r}}{2}$.

The next two propositions give us quantitative estimates on f and S_{D_1} when we move y towards O, along $\nu(O)$. Proposition 3.5 makes use of quantitative estimates of unique continuation, whereas Proposition 3.6 is mainly based on the asymptotic estimates of fundamental solutions obtained in Proposition 3.4.

Proposition 3.5. *Let Ω be an open set in $\mathbb{R}^n$ satisfying* (H1)*. Let D_1, D_2 be two inclusions in Ω verifying* (H2) *and let $y = h\nu(O)$, with O defined in* (3.3)*. If,* *given $\varepsilon > 0$, we have*

$$\|\Lambda_{D_1} - \Lambda_{D_2}\|_{\mathcal{L}(H^{1/2},H^{-1/2})} \leq \varepsilon,$$

then for every h, $0 < h < \overline{c}\,\overline{r}$, where $0 < \overline{c} < 1$, depends on L,

$$|f(y,y)| \leq C\frac{\varepsilon^{Bh^F}}{h^A}, \tag{3.15}$$

where $0 < A < 1$ and $C, B, F > 0$ are constants that depend only on the a priori data.

Proposition 3.6. *Let Ω be an open set in $\mathbb{R}^n$ satisfying* (H1). *Let D_1, D_2 be two inclusions in Ω verifying* (H2) *and $y = h\nu(O)$. Then for every h, $0 < h < \overline{r}_0/2$,*

$$|S_{D_1}(y,y)| \geq c_1 h^{2-n} - c_2 d_\mu^{2-2n} + c_3, \tag{3.16}$$

where c_1, c_2, and c_3 are positive constants depending only on the a priori data. Here $\overline{r}_0$ is the number introduced in Proposition 3.4.

Now we have all the tools that we need to prove Theorem 2.4.

Proof of Theorem 2.4. Let $O \in \partial D_1$ satisfying (3.3), that is

$$d_\mu(D_1, D_2) = \operatorname{dist}(O, D_2) = d_\mu.$$

Then, for $y = h\nu(O)$, with $0 < h < h_1$, where $h_1 = \min\{d_\mu, \overline{c}\,\overline{r}, \overline{r}_0/2\}$, using (3.12), we have

$$|S_{D_2}(y,y)| \leq c\int_{D_2} \frac{1}{(d_\mu - h)^{n-1}}\frac{1}{(d_\mu - h)^{n-1}}dx = c\frac{1}{(d_\mu - h)^{2n-2}}|D_2|. \tag{3.17}$$

Using Proposition 3.5, we have

$$\begin{aligned}|S_{D_1}(y,y)| - |S_{D_2}(y,y)| &\leq |S_{D_1}(y,y) - S_{D_2}(y,y)| \\ = |f(y,y)| \leq c\frac{\varepsilon^{Bh^F}}{h^A}&.\end{aligned}$$

On the other hand, by Proposition 3.6 and (3.17)

$$|S_{D_1}(y,y)| - |S_{D_2}(y,y)| \geq c_1 h^{2-n} - c_2(d_\mu - h)^{2-2n}.$$

Thus we have

$$c_3 h^{2-n} - c_4(d_\mu - h)^{2-2n} \leq \frac{\varepsilon^{Bh^F}}{h^A}.$$

That is

$$\begin{aligned}c_4(d_\mu - h)^{2-2n} &\geq c_3 h^{2-n} - \frac{\varepsilon^{Bh^F}}{h^A} = h^{2-n}\big(c_3 - \varepsilon^{Bh^F}h^{\widetilde{A}}\big) \\ &\geq c_5 h^{2-n}\big(1 - \varepsilon^{Bh^F}h^{\widetilde{A}}\big),\end{aligned} \tag{3.18}$$

where $\widetilde{A} = n - 2 - A$, $\widetilde{A} > 0$. Let $h = h(\varepsilon)$ where $h(\varepsilon) = \min\{|\ln\varepsilon|^{-\frac{1}{2F}}, d_\mu\}$, for $0 < \varepsilon \leq \varepsilon_1$, with $\varepsilon_1 \in (0,1)$ such that $\exp(-B|\ln\varepsilon_1|^{1/2}) = 1/2$. If $d_\mu \leq |\ln\varepsilon|^{-\frac{1}{2F}}$ the theorem follows using Proposition 3.3. In the other case we have

$$\varepsilon^{Bh(\varepsilon)^F}h(\varepsilon)^{\widetilde{A}} \leq \varepsilon^{B|\ln\varepsilon|^{-1/2}} \leq \exp\big(-B|\ln\varepsilon|^{1/2}\big).$$

Then, for any ε, $0 < \varepsilon < \varepsilon_1$,

$$(d_\mu - h(\varepsilon))^{2-2n} \geq c_6 h(\varepsilon)^{2-n},$$

that is, solving for d_μ, and recalling that, in this case, $h(\varepsilon) = |\ln \varepsilon|^{-\frac{1}{2F}}$,

$$d_\mu \leq c_7 |\ln \varepsilon|^{-\delta \frac{n-2}{2n-2}}, \tag{3.19}$$

where $\delta = 1/(2F)$. When $\varepsilon \geq \varepsilon_1$, then

$$d_\mu \leq \operatorname{diam} \Omega$$

and, in particular, when $\varepsilon_1 \leq \varepsilon < 1$

$$d_\mu \leq \operatorname{diam} \Omega \frac{|\ln \varepsilon|^{-\frac{1}{2F}}}{|\ln \varepsilon_1|^{-\frac{1}{2F}}}.$$

Finally, using Proposition 3.3, the theorem follows. ∎

4. Proofs of the auxiliary propositions. We premise the proof of Proposition 3.3 with one lemma.

LEMMA 4.1. *Let Ω be an open set in $\mathbb{R}^n$ satisfying* (H1). *Let D be a bounded open inclusion of Ω satisfying* (H2). *Then for every $P \in \partial D$, there exists a continuous path γ in $\Omega \setminus \overline{D}$ with one endpoint in P and the other on $\partial\Omega$, such that for every $z \in \gamma$*

$$|z - P| \leq c \operatorname{dist}(z, D), \tag{4.1}$$

where c is a positive constant depending on the a priori data only.

Proof. Using Lemma 5.2 of [A-B-R-V], (which adapted arguments due to Lieberman [Li]), we approximate $\operatorname{dist}(\cdot, \partial D)$ with a regularized distance $\tilde{d}$ such that $\tilde{d} \in C^2(\Omega \setminus D) \cup C^{1,\alpha}(\overline{\Omega \setminus D})$ and the following facts hold:

$$\gamma_0 \leq \frac{\operatorname{dist}(x, \partial D)}{\tilde{d}(x)} \leq \gamma_1,$$
$$|\nabla \tilde{d}(y)| \geq c_1 \quad \text{for every } y \in \Omega \text{ s.t. } \operatorname{dist}(y, \partial D) < b\overline{r},$$
$$\|\tilde{d}\|_{1,\alpha} \leq c_2 \overline{r},$$

where γ_0, γ_1, b, c_1 and c_2 are positive constants only depending on L and α. We define for $0 < h < a\overline{r}$, with a depending on L and α only,

$$E_h = \{x \in \Omega \setminus \overline{D} \, : \, \tilde{d}(x) > h\}.$$

Arguing as in Lemma 5.3 of [A-B-R-V], E_h is connected with boundary of class C^1 and

$$\widetilde{c}_1 h \leq \operatorname{dist}(x, \partial D) \leq \widetilde{c}_2 h \qquad \forall\, x \in \partial E_h \cap \Omega, \tag{4.2}$$

where $\widetilde{c}_1$, $\widetilde{c}_2$ are positive constants depending on L and α only. Let us fix $P \in \partial D$. Let $\nu(P)$ be the outer unit normal to ∂D in P (we recall that ∂D is $C^{1,\alpha}$). Since (4.2), there exists a point $P' \in E_h$ such that $P' = \tilde{h}\nu(P)$, where $\tilde{h}$ is a positive constant $\widetilde{c}_1 h < \tilde{h} < \widetilde{c}_2 h$. We denote by $\overline{PP'}$ the segment whose endpoints are P and P'. Since E_h is connected, there exists a continuous path $\gamma' \subset E_h$ with one endpoint P' and the other on $\partial\Omega$. Since $\gamma' \subset E_h$ we have that for every $x \in \gamma'$, $\operatorname{dist}(x, \partial D) \geq ch$, where c is a positive constant. We then define $\gamma = \gamma' \cup \overline{PP'}$ and the lemma follows. ∎

Proof of Proposition 3.3. Let us fix $P \in \partial D_1$. We distinguish the following two cases:

(i) $P \in \partial D_1 \cap \partial \mathcal{G}$,
(ii) $P \in \partial D_1 \setminus \partial \mathcal{G}$.
If case (i) occurs then,

$$\mathrm{dist}(P, \partial D_2) = \mathrm{dist}(P, \overline{D}_2) \leq d_\mu.$$

Let us consider case (ii). Let γ be the continuous path constructed in Lemma 4.1 from P to $\partial\Omega$. Since $P \notin \partial\mathcal{G}$, there exists $z \in \gamma \cap \partial D_2 \cap \partial\Omega_D$.

$$\mathrm{dist}(z, D_1) \leq \sup_{x \in \partial D_2 \cap \partial \Omega_D} \big\{ \mathrm{dist}(x, D_1) \big\} \leq d_\mu(D_1, D_2).$$

Thus

$$|z - P| \leq c d_\mu(D_1, D_2),$$

where $c > 0$ is the constant appearing in (4.1). On the other hand,

$$\mathrm{dist}(P, \partial D_2) \leq |z - P|.$$

So we obtain that for every $P \in \partial D_1$,

$$\mathrm{dist}(P, \partial D_2) \leq c d_\mu(D_1, D_2).$$

Similarly, one can show that for every $Q \in \partial D_2$

$$\mathrm{dist}(Q, \partial D_1) \leq c d_\mu(D_1, D_2).$$

Then we conclude that

$$d_{\mathcal{H}}(\partial D_1, \partial D_2) \leq c d_\mu(D_1, D_2). \qquad \square$$

Proof of Proposition 3.4. Let us prove (i).
Let us consider the case $x \in D$ and $y \in \partial D$. The cases in which $x, y \in D$ or $x, y \in \mathcal{C}D$ are trivial. Let $h = |x - y|$. Let c be a positive number less than $\frac{1}{1+2\sqrt{n}}$. We distinguish the following two cases:
(a) $\mathrm{dist}(x, \partial D) < ch$,
(b) $\mathrm{dist}(x, \partial D) \geq ch$.
Let us consider the case (a). Let $P \in \partial D$ be such that $|P - x| = \mathrm{dist}(x, \partial D)$. For every $r > 0$, let $Q_r(P)$ be the cube centered at P, with sides of length $2r$ and parallel to the coordinates axes. We have that the ball $B_r(P)$ is inscribed into $Q_r(P)$; in particular, $x \in Q_{ch}(P)$. On the other hand,

$$|P - y| \geq |y - x| - |P - x| \geq h(1 - c).$$

Then, due to our choice of c, $|P - y| > (2ch)\sqrt{n}$, that is $y \notin Q_{2ch}(P)$. Thus

$$\mathrm{div}_z\big((1 + (k-1)\chi_D)\nabla_z \Gamma_D(z, y)\big) = 0 \qquad \text{in } Q_{\frac{3}{2}ch}(P)$$

and for the piecewise $C^{1,\alpha}$ regularity of Γ_D, proved in [DB-E-F], see also [L-V], we have

$$\|\nabla \Gamma_D(\cdot, y)\|_{L^\infty(Q_{ch}(P))} \leq \frac{\overline{c}_1}{h} \|\Gamma_D(\cdot, y)\|_{L^\infty(Q_{\frac{3}{2}ch}(P))}, \tag{4.3}$$

where $\overline{c}_1$ depends on L, k, n, and α only. Using the pointwise bound of Γ_D with Γ (see [L-S-W]), we have

$$\|\Gamma_D(\cdot,y)\|_{L^\infty\left(Q_{\frac{3}{2}ch}(P)\right)} \leq \overline{c}_2\left(\frac{ch}{2}\right)^{2-n}, \tag{4.4}$$

where $\overline{c}_2$ depends on n and k only. Hence, by (4.3) and (4.4) we get

$$|\nabla_x\Gamma_D(x,y)| \leq \|\nabla\Gamma_D(\cdot,y)\|_{L^\infty(Q_{ch}(P))} \leq \overline{c}_3 h^{1-n} = \overline{c}_3|x-y|^{1-n}, \tag{4.5}$$

where $\overline{c}_3$ depends on L, k, n, and α only.

If case (b) occurs, then $Q_{\frac{ch}{\sqrt{n}}}(x) \subset D$. Hence

$$\begin{aligned}
|\nabla_x\Gamma_D(x,y)| &\leq \|\nabla\Gamma_D(\cdot,y)\|_{L^\infty\left(Q_{\frac{ch}{2\sqrt{n}}}(x)\right)} \leq \frac{\overline{c}_4}{h}\|\Gamma_D(\cdot,y)\|_{L^\infty\left(Q_{\frac{c}{\sqrt{n}}}(P)\right)} \\
&\leq \frac{\overline{c}_4}{h}(h(1-c))^{2-n} = \overline{c}_4' h^{1-n} = \overline{c}_4'|x-y|^{1-n},
\end{aligned}$$

where $\overline{c}_4$, $\overline{c}_4'$ depend on L, k, n, and α only.

Let us prove (ii).

Let us fix $r_1 = \min\left\{\frac{1}{2}(8L)^{-1/\alpha}\overline{r}, \frac{\overline{r}}{2}\right\}$. Recalling Definition 2.1, we have that

$$\partial D \cap B_{\overline{r}}(0) = \{x \in B_{\overline{r}}(0) \,:\, x_n = \varphi(x')\},$$

where $\varphi \in C^{1,\alpha}(\mathbb{R}^{n-1})$ satisfying $\varphi(0) = |\nabla\varphi(0)| = 0$. Let $\theta \in C^\infty(\mathbb{R})$ be such that $0 \leq \theta \leq 1$, $\theta(t) = 1$, for $|t| < 1$, $\theta(t) = 0$, for $|t| > 2$ and $|\frac{d\theta}{dt}| \leq 2$. We consider the following change of variables $\xi = \Phi(x)$ defined by

$$\begin{cases} \xi' = x' \\ \xi_n = x_n - \varphi(x')\theta\left(\frac{|x'|}{r_1}\right)\theta\left(\frac{x_n}{r_1}\right). \end{cases}$$

It can be verified that, with the given choice of r_1, the following properties of Φ hold:

$$\Phi(Q_{2r_1}(0)) = Q_{2r_1}(0), \tag{4.6}$$

$$\Phi(Q_{r_1}(0) \cap D) = Q^+_{r_1}(0), \tag{4.7}$$

$$c^{-1}|x_1 - x_2| \leq |\Phi(x_1) - \Phi(x_2)| \leq c|x_1 - x_2| \qquad \forall\, x_1, x_2 \in \mathbb{R}^n, \tag{4.8}$$

$$|\Phi(x) - x| \leq \frac{c}{\overline{r}^\alpha}|x|^{1+\alpha} \qquad \forall\, x \in \mathbb{R}^n, \tag{4.9}$$

$$|D\Phi(x) - I| \leq \frac{c}{\overline{r}^\alpha}|x|^\alpha \qquad \forall\, x \in \mathbb{R}^n, \tag{4.10}$$

where $Q^+_{r_1}(0) = \{x \in Q_{r_1}(0) \,:\, x_n > 0\}$ and $c \geq 1$ depends on L and α only. Φ is a $C^{1,\alpha}$ diffeomorphism from $\mathbb{R}^n$ into itself. Let us define the cylinder C_{r_1} as

$$C_{r_1} = \{x \in \mathbb{R}^n \,:\, |x'| < r_1, |x_n| < r_1\}.$$

For $x, y \in C_{r_1}$, we have that $\widetilde{\Gamma}_D(\xi,\eta) = \Gamma_D(x,y)$, where $\xi = \Phi(x)$, $\eta = \Phi(y)$, is solution of

$$\mathrm{div}_\xi((1 + (k-1)\chi^+)B(\xi)\nabla_\xi\widetilde{\Gamma}_D(\xi,\eta)) = -\delta(\xi - \eta), \tag{4.11}$$

where $B = \frac{JJ^T}{\det J}$, with $J = \frac{\partial\xi}{\partial x}(\Phi^{-1}(\xi))$. We observe that B is of class C^α and $B(0) = I$. Let us consider

$$\widetilde{R}(x,y) = \widetilde{\Gamma}_D(x,y) - \Gamma_+(x,y),$$

where we keep the notation x, y to indicate ξ, η. By the properties of Γ_+ and by (4.11), $\widetilde{R}$ satisfies

$$\operatorname{div}_x((1+(k-1)\chi^+)\nabla_x\widetilde{R}(x,y)) = \operatorname{div}_x((1+(k-1)\chi^+)(I-B)\nabla_x\widetilde{\Gamma}_D(x,y)).$$

Let $\widetilde{L} > 0$, depending on the a priori data only, be such that $\overline{\Omega} \subset B_{\widetilde{L}}(0)$. Thus, using the fundamental solution Γ_+ we obtain

$$\begin{aligned}
-\widetilde{R}(x,y) &= \int_{B_{\widetilde{L}}(0)} (1+(k-1)\chi^+)(B-I)\nabla_z\Gamma_+(z,y)\cdot\nabla_z\widetilde{\Gamma}_D(z,x)dz \\
&+ \int_{\partial B_{\widetilde{L}}(0)} (1+(k-1)\chi^+)\left[\widetilde{R}(x,z)\frac{\partial\Gamma_+}{\partial\nu}(z,y) - \Gamma_+(z,y)\frac{\partial\widetilde{R}}{\partial\nu}(x,z)\right]d\sigma(z) \\
&= \int_{B_{\widetilde{L}}(0)\cap C_{r_1}} (1+(k-1)\chi^+)(B-I)\nabla_z\Gamma_+(z,y)\cdot\nabla_z\widetilde{\Gamma}_D(z,x)dz \\
&+ \int_{B_{\widetilde{L}}(0)\setminus C_{r_1}} (1+(k-1)\chi^+)(B-I)\nabla_z\Gamma_+(z,y)\cdot\nabla_z\widetilde{\Gamma}_D(z,x)dz \\
&+ \int_{\partial B_{\widetilde{L}}(0)} \left[\widetilde{R}(x,z)\frac{\partial\Gamma_+}{\partial\nu}(z,y) - \Gamma_+(z,y)\frac{\partial\widetilde{R}}{\partial\nu}(x,z)\right]d\sigma(z).
\end{aligned}$$

For $|x|, |y| < r_1/2$, the last two integrals are bounded. Using (3.12) we obtain

$$\begin{aligned}
|\widetilde{R}(x,y)| &\le c\left(1 + \int_{C_{r_1}} |z|^\alpha |x-z|^{1-n}|y-z|^{1-n}dz\right) \\
&= c\left(1 + I_1 + I_2\right),
\end{aligned}$$

where c depends on L, α, k, and n and

$$\begin{aligned}
I_1 &= \int_{\{|z|<4h\}\cap C_{r_1}} |z|^\alpha|x-z|^{1-n}|y-z|^{1-n}dz, \\
I_2 &= \int_{\{|z|>4h\}\cap C_{r_1}} |z|^\alpha|x-z|^{1-n}|y-z|^{1-n}dz.
\end{aligned}$$

Now

$$\begin{aligned}
I_1 &\le \int_{|w|<4} h^\alpha|w|^\alpha h^{1-n}\Big|\frac{x}{h}-w\Big|^{1-n} h^{1-n}\Big|\frac{y}{h}-w\Big|^{1-n}h^n dw \\
&= h^{\alpha+2-n}\int_{|w|<4}|w|^\alpha\Big|\frac{x}{h}-w\Big|^{1-n}\Big|\frac{y}{h}-w\Big|^{1-n}dw \\
&\le h^{\alpha+2-n}F(\xi,\eta),
\end{aligned}$$

where $h = |x-y|$ and

$$F(\xi,\eta) = 4^{\alpha}\int_{|w|<4} |\xi - w|^{1-n}|\eta - w|^{1-n}dw$$

and $\xi = x/h$ and $\eta = y/h$. From standard bounds (see, for instance, [M, Chapter 2, section 11]) it is not difficult to see that

$$F(\xi,\eta) \leq \text{const.} < \infty,$$

for all $\xi,\eta \in \mathbb{R}^n$, $|\xi-\eta| = 1$. Thus

$$I_1 \leq c|x-y|^{\alpha+2-n}.$$

Let us now consider I_2. Since $|y| = -y_n \leq |x-y| = h$, we can deduce and $|z| \leq \frac{4}{3}|y-z|$ and $|z| \leq 2|x-z|$ and thus obtain that

$$I_2 \leq c\int_{|z|>4h} |z|^{\alpha+1-n+1-n}dz \leq ch^{\alpha+2-n}.$$

Then we conclude that

$$|\widetilde{R}(x,y)| \leq c|x-y|^{\alpha+2-n}, \tag{4.12}$$

for every $|x|, |y| < r_1/2$, where c depends on L, α, k, and n only. Let us go back to the original coordinates system. We observe that if $x \in \Phi^{-1}(B^+_{r_1/2}(0))$ and $y = e_n y_n$, with $y_n \in (-r_1/2, 0)$, then $|\Phi(x) - x|$ is bounded by $c|x-y|^{1+\alpha}$. Namely, since $\Phi(x)\cdot y \leq 0$ and $\Phi(y) = y$, by (4.8) we have

$$c^{-1}|x| \leq |\Phi(x)| \leq |\Phi(x) - y| \leq c|x-y|. \tag{4.13}$$

On the other hand, by (4.9) and (4.13)

$$|\Phi(x) - x| \leq \frac{c}{\overline{r}^{\alpha}}|x|^{1+\alpha} \leq \frac{c'}{\overline{r}^{\alpha}}|x-y|^{1+\alpha}. \tag{4.14}$$

We have

$$\begin{aligned}
R(x,y) &= \Gamma_D(x,y) - \Gamma_+(x,y)\\
&= \Gamma_D(x,y) - \Gamma_+(x,y) + \Gamma_+(\Phi(x),\Phi(y)) - \Gamma_+(\Phi(x),\Phi(y))\\
&= \widetilde{R}(\Phi(x),\Phi(y)) + \Gamma_+(\Phi(x),y) - \Gamma_+(x,y).
\end{aligned}$$

Using (4.8), (4.9), (4.12), and (4.14) we obtain

$$\begin{aligned}
&|\Gamma_D(x,y) - \Gamma_+(x,y)|\\
&\leq \frac{c}{\overline{r}^{\alpha}}|x-y|^{\alpha+2-n} + \frac{c}{\overline{r}^{\alpha}}\|\nabla\Gamma_+(\cdot,y)\|_{L^\infty(Q_{r_1})}|x-\Phi(x)|\\
&\leq \frac{c}{\overline{r}^{\alpha}}|x-y|^{\alpha+2-n} + \frac{c'}{\overline{r}^{\alpha}}|x-y|^{1+\alpha}h^{1-n}\\
&\leq \frac{c''}{\overline{r}^{\alpha}}|x-y|^{\alpha+2-n},
\end{aligned}$$

where c'' depends on k, n, α, and L only. We estimate now the first derivative of R. To estimate the first derivative of $\widetilde{R}$ let us consider a cube $Q \subset B^+_{r_1/4}(x)$ of side $cr_1/4$, with $0 < c < 1$, such that $x \in \partial Q$. The following interpolation inequality holds:

$$\|\nabla \widetilde{R}(\cdot, y)\|_{L^\infty(Q)} \leq c \|\widetilde{R}(\cdot, y)\|^{1-\delta}_{L^\infty(Q)} |\nabla \widetilde{R}(\cdot, y)|^{\delta}_{\alpha, Q},$$

where $\delta = \frac{1}{1+\alpha}$, c depends on L only, and

$$|\nabla \widetilde{R}|_{\alpha, Q} = \sup_{x, x' \in Q, x \neq x'} \frac{|\nabla \widetilde{R}(x, y) - \nabla \widetilde{R}(x', y)|}{|x - x'|^\alpha}.$$

Since, from the piecewise Hölder continuity of $\nabla \Gamma_D$ (see (4.3)), and also of $\nabla \Gamma_+$, (see (3.11)), we have that

$$|\nabla \widetilde{R}(\cdot, y)|_{\alpha, Q} \leq |\nabla \widetilde{\Gamma}_D(\cdot, y)|_{\alpha, Q} + |\nabla \Gamma_+(\cdot, y)|_{\alpha, Q} \leq c h^{-\alpha + 1 - n},$$

where c depends on L only, thus we conclude that

$$|\nabla_x \widetilde{R}(x, y)| \leq \frac{c}{\overline{r}^\eta} h^{(\alpha + 2 - n)(1 - \delta)} h^{(-\alpha + 1 - n)\delta} = \frac{c}{\overline{r}^\eta} h^{1 - n + \eta},$$

where $\eta = \frac{\alpha^2}{1+\alpha}$. Thus

$$|\nabla_x \widetilde{R}(x, y)| \leq \frac{c}{\overline{r}^\eta} |x - y|^{\eta + 1 - n}, \tag{4.15}$$

where $\eta = \frac{\alpha^2}{1+\alpha}$ and c depends on L only. Concerning Γ_+ we have

$$\begin{aligned}
&|\nabla_x \Gamma_+(\Phi(x), y) - \nabla_x \Gamma_+(x, y)| \\
&= |D\Phi(x)^T \nabla \Gamma_+(\cdot, y)_{|\Phi(x)} - \nabla_x \Gamma_+(x, y)| \\
&\leq |(D\Phi(x)^T - I) \nabla \Gamma_+(\cdot, y)_{|\Phi(x)}| \\
&\quad + |\nabla \Gamma_+(\cdot, y)_{|\Phi(x)} - \nabla_x \Gamma_+(x, y)| \\
&\leq \frac{c}{\overline{r}^\alpha} \|\nabla \Gamma_+(\cdot, y)\|_{L^\infty(Q_{r_1})} |x - \Phi(x)| + |\nabla \Gamma_+(\cdot, y)|_{\alpha, Q} |\Phi(x) - x|^\alpha \\
&\leq \frac{c'}{\overline{r}^\alpha} h^{1+\alpha} h^{1-n} + \frac{c}{\overline{r}^{\alpha^2}} h^{-\alpha + 1 - n} h^{(1+\alpha)\alpha} \\
&\leq \frac{c}{\overline{r}^{\alpha^2}} h^{1 - n + \alpha^2},
\end{aligned}$$

where c depends on k, n, α, and L only. $\square$

Proof of Proposition 3.5. Let us fix $\overline{y} \in \mathcal{S}_{2\overline{r}}$, where $\mathcal{S}_{2\overline{r}}$ is the set introduced in Definition 3.1, and let us consider $f(\overline{y}, \cdot)$. We have that

$$\Delta_w f(\overline{y}, w) = 0 \qquad \text{in } \mathcal{C}\overline{\Omega}_D. \tag{4.16}$$

For $w \in \mathcal{S}_{2\overline{r}}$, by (2.6), (3.10), and (3.12) we have

$$|f(\overline{y}, w)| \leq C(\overline{r}, L, M) \|\Lambda_{D_1} - \Lambda_{D_2}\| = \widetilde{\varepsilon}. \tag{4.17}$$

Let us now estimate $f(\overline{y}, w)$ when $w \in \mathcal{G}$. Again recalling Definition 3.1, we define $\mathcal{G}^h = \{x \in \mathcal{G} : \mathrm{dist}(x, \Omega_D) \geq h\}$. For every $w \in \mathcal{G}^h$, we have that

$$|S_{D_1}(\overline{y}, w)| \leq |k-1| \int_{D_1} |\nabla_x \Gamma_{D_1}(x, \overline{y})|\, |\nabla_x \Gamma_{D_2}(x, w)| dx$$

$$\leq c \int_{D_1} |x - w|^{1-n} dx \leq c h^{1-n}. \tag{4.18}$$

Similarly, $|S_{D_2}(\overline{y}, w)| \leq c h^{1-n}$. Then we conclude that

$$|f(\overline{y}, w)| \leq c h^{1-n} \qquad \text{in } \mathcal{G}^h.$$

At this stage we shall make use of the three spheres inequality for supremum norms of harmonic functions v; see, for instance, [K-M], [K]. For every l_1, l_2, $1 < l_1 < l_2$ and for every $x \in \mathcal{G} \cup \mathcal{S}_{2\overline{r}} \cup \Omega_{\overline{r}}$ there exists $\tau \in (0, 1]$, depending only on l_1, l_2 and n such that

$$\|v\|_{L^\infty(B_{l_1 r}(x))} \leq \|v\|^{\tau}_{L^\infty(B_r(x))} \|v\|^{1-\tau}_{L^\infty(B_{l_2 r}(x))}.$$

We apply it for $v(\cdot) = f(\overline{y}, \cdot)$ in the ball $B_{\overline{r}}(\overline{x})$, where $\overline{x} \in \mathcal{S}_{2\overline{r}}$ be such that $\mathrm{dist}(\overline{x}, \Gamma) = \overline{r}/2$, where $\Gamma = \{x \in \mathbb{R}^n : \mathrm{dist}(x, \Omega) = \overline{r}\} \subset \partial \mathcal{S}_{2\overline{r}}$, $l_1 = 3r = 3\overline{r}/2$, and $l_2 = 4r = 2\overline{r}$, then we obtain

$$\|f(\overline{y}, \cdot)\|_{L^\infty(B_{3\overline{r}/2}(\overline{x}))} \leq \|f(\overline{y}, \cdot)\|^{\tau}_{L^\infty(B_{\overline{r}/2}(\overline{x}))} \|f(\overline{y}, \cdot)\|^{1-\tau}_{L^\infty(B_{2\overline{r}}(\overline{x}))}. \tag{4.19}$$

For every $\overline{w} \in \mathcal{G}^h$, we denote with γ a simple arc in $\overline{\mathcal{G}} \cup \overline{\Omega}_{\overline{r}} \cup \overline{\mathcal{S}}_{2\overline{r}}$ joining $\overline{x}$ to $\overline{w}$. Let us define $\{x_i\}$, $i = 1, \ldots, s$ as follows $x_1 = \overline{x}$, $x_{i+1} = \gamma(t_i)$, where $t_i = \max\{t : |\gamma(t) - x_i| = \overline{r}\}$ if $|x_i - \overline{w}| > \overline{r}$, otherwise let $i = s$ and stop the process. By construction, the balls $B_{\overline{r}/2}(x_i)$ are pairwise disjoint, $|x_{i+1} - x_i| = \overline{r}$ for $i = 1, \ldots, s-1$, $|x_s - \overline{w}| \leq \overline{r}$. For (2.1), there exists β such that $s \leq \beta$. An iterated application of the three spheres inequality (4.19) for $f(\overline{y}, \cdot)$ (see, for instance, [A-B-R-V, pg. 780], [A-DB, Appendix E]) gives that for any r, $0 < r < \overline{r}$,

$$\|f(\overline{y}, \cdot)\|_{L^\infty(B_{r/2}(\overline{w}))} \leq \|f(\overline{y}, \cdot)\|^{\tau^s}_{L^\infty(B_{r/2}(\overline{x}))} \|f(\overline{y}, \cdot)\|^{1-\tau^s}_{L^\infty(\mathcal{G})}. \tag{4.20}$$

We can now estimate the right-hand side of (4.20) by (4.17) and (4.18) and obtain, for any r, $0 < r < \overline{r}$

$$\|f(\overline{y}, \cdot)\|_{L^\infty(B_{r/2}(\overline{w}))} \leq c (h^{1-n})^{1-\tau^s} \varepsilon^{\tau^s} \leq c (h^{1-n})^{A} \varepsilon^{\widetilde{\beta}}, \tag{4.21}$$

where $\widetilde{\beta} = \tau^\beta$ and $A = 1 - \widetilde{\beta}$. Let $O \in \partial D_1$, as defined in (3.3), that is

$$d(O, D_2) = d_\mu(D_1, D_2).$$

There exists a $C^{1,\alpha}$ neighborhood U of O in $\partial \Omega_D$ with constants $\overline{r}$ and L. Thus there exists a nontangential vector field $\widetilde{\nu}$, defined on U such that the truncated cone

$$C(O, \widetilde{\nu}(O), \theta, \overline{r}) = \left\{ x \in \mathbb{R}^n : \frac{(x - O) \cdot \widetilde{\nu}(O)}{|x - O|} > \cos\theta,\ |x - O| < \overline{r} \right\} \tag{4.22}$$

satisfies

$$C(O, \widetilde{\nu}(O), \theta, \overline{r}) \subset \mathcal{G},$$

where $\theta = \arctan(1/\overline{L})$. Let us define

$$\lambda_1 = \min\left\{\frac{\overline{r}}{1+\sin\theta}, \frac{\overline{r}}{3\sin\theta}\right\},$$

$$\theta_1 = \arcsin\left(\frac{\sin\theta}{4}\right),$$

$$w_1 = O + \lambda_1\nu,$$

$$\rho_1 = \lambda_1 \sin\theta_1.$$

We have that $B_{\rho_1}(w_1) \subset C(O, \widetilde{\nu}(O), \theta_1, \overline{r})$, $B_{4\rho_1}(w_1) \subset C(O, \widetilde{\nu}(O), \theta, \overline{r})$. Let $\overline{w} = w_1$, since $\rho_1 \leq \overline{r}/2$, we can use (4.21) in the ball $B_{\rho_1}(\overline{w})$ and we can approach $O \in \partial D_1$ by constructing a sequence of balls contained in the cone $C(O, \widetilde{\nu}(O), \theta_1, \overline{r})$. We define, for $k \geq 2$

$$w_k = O + \lambda_k\nu, \qquad \lambda_k = \chi\lambda_{k-1}, \qquad \rho_k = \chi\rho_{k-1}, \qquad \text{with } \chi = \frac{1-\sin\theta_1}{1+\sin\theta_1}.$$

Hence $\rho_k = \chi^{k-1}\rho_1$, $\lambda_k = \chi^{k-1}\lambda_1$, and

$$B_{\rho_{k+1}}(w_{k+1}) \subset B_{\rho_{3k}}(w_k) \subset B_{\rho_{4k}}(w_k) \subset C(O, \nu, \theta, \overline{r}).$$

Denoting $d(k) = |w_k - O| - \rho_k = \lambda_k - \rho_k$, we have $d(k) = \chi^{k-1}d(1)$, with $d(1) = \lambda_1(1 - \sin\theta)$. For any r, $0 < r \leq d(1)$, let $k(r)$ be the smallest integer such that $d(k) \leq r$, that is

$$\frac{\left|\log\frac{r}{d(1)}\right|}{\left|\log\chi\right|} \leq k(r) - 1 \leq \frac{\left|\log\frac{r}{d(1)}\right|}{\left|\log\chi\right|} + 1.$$

By an iterated application of the three spheres inequality over the chain of balls $B_{\rho_1}(w_1), \dots, B_{\rho_{k(r)}}(w_{k(r)})$, we have

$$\begin{aligned} \|f(\overline{y}, \cdot)\|_{L^\infty(B_{\rho_{k(r)}}(w_{k(r)}))} &\leq c(h^{1-n})^{A(1-\tau^{k(r)-1})}\varepsilon^{\beta\tau^{k(r)-1}} \\ &\leq c(h^{1-n})^{A}\varepsilon^{\beta\tau^{k(r)-1}}, \end{aligned} \tag{4.23}$$

for $0 < r < c\overline{r}$, where $0 < c < 1$ depends on L only.

Let us now consider $f(y, w)$ as a function of y. First observe that

$$\Delta_y f(y, w) = 0 \qquad \text{in } \mathcal{C}\Omega_D, \qquad \forall\, w \in \mathcal{C}\Omega_D.$$

For $y, w \in \mathcal{G}^h$, $y \neq w$, using (3.12), we have

$$|S_{D_1}(y, w)| \leq c\int_{D_1} |x - y|^{1-n}|x - w|^{1-n}dx \leq ch^{2-n};$$

similarly, for S_{D_2}. Therefore

$$|f(y, w)| \leq ch^{2-2n} \qquad \text{with } y, w \in \mathcal{G}^h.$$

Finally, for $y \in \mathcal{S}_{2\overline{r}}$ and $w \in \mathcal{G}^h$, using (4.23), we have

$$|f(y, w)| \leq c(h^{1-n})^A\varepsilon^{\beta\tau^{k(h)-1}}.$$

Proceeding as before, let us fix $w \in \mathcal{G}$ such that $\mathrm{dist}(w, \partial\Omega_D) = h$ and $\widetilde{y} \in \mathcal{S}_{2\overline{r}}$ such that $\mathrm{dist}(\widetilde{y}, \Gamma) = \overline{r}/2$. Taking $r = \overline{r}/2$, $l_1 = 3r$, $l_2 = 4r$, $y_1 = O + \lambda_1 \nu$, and using iteratively the three spheres inequality, we have

$$\|f(y,w)\|_{L^\infty(B_{\overline{r}/2}(y_1))} \leq \|f(y,w)\|^{\tau^s}_{L^\infty(B_{\overline{r}/2}(\widetilde{y}))} \|f(y,w)\|^{1-\tau^s}_{L^\infty(\mathcal{G})},$$

where τ and s are the same number established previously. Therefore

$$\begin{aligned}
\|f(y,w)\|^{\tau^s}_{L^\infty(B_{\overline{r}/2}(y_1))} &\leq c(h^{2-2n})^{1-\tau^s}(h^{1-n})^{A\tau^s}(\varepsilon^{\beta\tau^{k(h)-1}})^{\tau^s} \\
&\leq c(h^{2-2n})^{1-\gamma}(h^{1-n})^{A\tau^s}(\varepsilon^{\beta\tau^{k(h)-1}})^{\gamma} \\
&\leq c(h^{2-2n})^{A'}(\varepsilon^{\beta\tau^{k(h)-1}})^{\gamma},
\end{aligned}$$

where $\gamma = \tau^\beta$, with β as before, so $0 < \gamma < 1$, and $A' = A\tau^s + 1 - \gamma$. Once more, let us apply iteratively the three spheres inequality over a chain of balls contained in a cone with vertex in O and we obtain

$$\|f(y,w)\|_{L^\infty(B_{\rho_k}(y_{k(h)}))} \leq c(h^{2-2n})^{A'(1-\tau^{k(h)-1})}(\varepsilon^{\beta\tau^{k(h)-1}})^{\gamma\tau^{k(h)-1}}. \tag{4.24}$$

Now, from (4.24), choosing $y = w = h\nu(O)$, where $\nu(O)$ is the exterior unit normal to $\partial\Omega_D$ in O, we obtain

$$|f(y,y)| \leq ch^{A''}(\varepsilon^{\beta\tau^{k(h)-1}})^{\gamma\tau^{k(h)-1}}, \tag{4.25}$$

where $A'' = -(2-2n)\beta A' > 0$. We observe that, for $0 < h < c\overline{r}$, where $0 < c < 1$ depends on L, $k(h) \leq c|\log h| = -c\log h$, so we can write

$$\tau^{k(h)} = \mathrm{e}^{-c\log h \log(\tau)} = h^{-c\log\tau} = h^{c|\log\tau|} = h^F,$$

with $F = c|\log\tau|$. Therefore

$$\begin{aligned}
|f(y,y)| &\leq h^{-A''}\varepsilon^{B\tau^{k(h)}} \\
&= \mathrm{e}^{-A''\log h}\mathrm{e}^{B\tau^{k(h)}\log\varepsilon} \\
&= \mathrm{e}^{-A''\log h + B'h^F\log\varepsilon}.
\end{aligned}$$

Then in (4.25) we obtain that

$$|f(y,y)| \leq \mathrm{e}^{-A'\log h + B'h^F\log\varepsilon} = \frac{\varepsilon^{B'h^F}}{h^{A'}}. \qquad \square$$

Proof of Proposition 3.6. Let us consider $y = h\nu(O)$, where $\nu(O)$ is the exterior outer normal to $\partial\Omega_D$ in O with O defined as in (3.3), $0 < h < \overline{r}_0$, where $\overline{r}_0$ is the number introduced in Proposition 3.4 and $x \in D_1$ such that $|x-y| < r$, with $0 < r < \overline{r}_0$. Let us first observe that since $O \in \partial D_1$ and $x \in D_1$, for Γ_{D_1} we have the asymptotic formula (3.14), which says that

$$\left|\nabla_x\Gamma_{D_1}(x,y) - \nabla_x\Gamma_+(x,y)\right| \leq c_1|x-y|^{1-n+\delta}.$$

Furthermore, since we are in the situation in which $x \in D_1$ and $y \notin D_1$, for (3.11), $\Gamma_+(x,y) = 2/(k+1)\Gamma(x,y)$, where $\Gamma(x,y)$ denotes the standard fundamental solution of the Laplace operator. Let us now consider $\Gamma_{D_2}(x,y)$. With our choice of O, x and y, we know that $y \notin D_2$ but we do not have any information on x; that is, we do not know in which side of the interface ∂D_2 it is. Thus we have to distinguish different situations.

If $x \in B_r(O) \cap D_1 \cap D_2$, then we have the asymptotic formula (3.11) for Γ_{D_2} and from Lemma 3.1 of [A] the following formula holds:

$$\nabla_x \Gamma_{D_1}(x,y) \cdot \nabla_x \Gamma_{D_2}(x,y) \geq c|x-y|^{2-2n}. \tag{4.26}$$

Now consider the case $x \in (D_1 \setminus D_2) \cap B_r(O)$. In this region let us consider a smaller ball $B_\rho(O)$ centered in O with radius ρ where $0 < \rho < d_\mu$. Since the definition of d_μ, we have $B_\rho \cap D_2 = \emptyset$. If x and y are in $B_\rho(O)$, we have

$$\begin{cases} \Delta_x\big(\Gamma_{D_2}(x,y) - \Gamma(x,y)\big) = 0 & \text{in } B_\rho(O), \\ \big[\Gamma_{D_2}(x,y) - \Gamma(x,y)\big]_{|\partial B_\rho(O)} \leq c\rho^{2-n}. \end{cases} \tag{4.27}$$

Thus by the maximum principle

$$\big|\Gamma_{D_2}(x,y) - \Gamma(x,y)\big| \leq c_1\rho^{2-n} \qquad \forall\, x, y \in B_\rho(O), \tag{4.28}$$

and by interior gradient bound

$$\big|\nabla_x \Gamma_{D_2}(x,y) - \nabla_x \Gamma(x,y)\big| \leq c_2\rho^{1-n} \qquad \forall\, x \in B_{\rho/2}(O), \forall\, y \in B_\rho(O). \tag{4.29}$$

Thus, using Lemma 3.1 of [A], in $B_{\rho/2}(O)$ we obtain the formula

$$\nabla_x \Gamma_{D_1}(x,y) \cdot \nabla_x \Gamma_{D_2}(x,y) \geq c_3|x-y|^{2-2n} - c_4\rho^{2-2n}. \tag{4.30}$$

Let us consider $h \leq \overline{r}_0/2$ and $B_r(O) = \{x \in \mathbb{R}^n : |x - O| < r\}$, with $0 < r < \overline{r}_0$. Then we have

$$\begin{aligned} &|S_{D_1}(y,y)| \\ &= |k-1| \left| \int\limits_{D_1 \cap B_r(O)} \nabla\Gamma_{D_1} \cdot \nabla\Gamma_{D_2} dx + \int\limits_{D_1 \setminus B_r(O)} \nabla\Gamma_{D_1} \cdot \nabla\Gamma_{D_2} dx \right| \\ &\geq |k-1| \left| \int\limits_{D_1 \cap B_r(O)} \nabla\Gamma_{D_1} \cdot \nabla\Gamma_{D_2} dx \right| - |k-1| \left| \int\limits_{D_1 \setminus B_r(O)} \nabla\Gamma_{D_1} \cdot \nabla\Gamma_{D_2} dx \right| \end{aligned}$$

The first term can be estimated as follows:

$$\begin{aligned} &\left| \int_{D_1 \cap B_r(O)} \nabla\Gamma_{D_1} \cdot \nabla\Gamma_{D_2} dx \right| \\ &= \left| \int_{(D_1 \cap D_2) \cap B_r(O)} \nabla\Gamma_{D_1} \cdot \nabla\Gamma_{D_2} dx + \int_{(D_1 \setminus D_2) \cap B_r(O)} \nabla\Gamma_{D_1} \cdot \nabla\Gamma_{D_2} dx \right| \\ &\geq \left| \int_{(D_1 \cap D_2) \cap B_r(O)} \nabla\Gamma_{D_1} \cdot \nabla\Gamma_{D_2} dx + \int_{(D_1 \setminus D_2) \cap B_\rho(O)} \nabla\Gamma_{D_1} \cdot \nabla\Gamma_{D_2} dx \right| \\ &\quad - \left| \int_{[(D_1 \setminus D_2) \cap B_r(O)] \setminus B_\rho(O)} \nabla\Gamma_{D_1} \cdot \nabla\Gamma_{D_2} dx \right|. \end{aligned}$$

In conclusion, choosing $\rho = d_\mu/2$ and using (4.26) and (3.12) we obtain

$$
\begin{aligned}
|S_D(y,y)| \geq c_1 & \int\limits_{[(D_1\cap D_2)\cap B_r(O)]\cup[(D_1\setminus D_2)\cap B_{d_\mu/2}(O)]} |x-y|^{2-2n}dx \\
& - c_2 \int\limits_{[(D_1\setminus D_2)\cap B_r(O)]\setminus B_{d_\mu/2}(O)} |x-y|^{1-n}|x-y|^{1-n}dx \\
& - c_3 \int\limits_{D_1\setminus B_r(O)} |x-y|^{1-n}|x-y|^{1-n}dx \\
& \geq c_4 h^{2-n} - c_5 d_\mu^{2-2n} - c_7. \qquad \square
\end{aligned}
$$

Acknowledgments. The authors wish to express their gratitude to Professors Edi Rosset and Sergio Vessella for fruitful discussions on the topics of this paper. The authors also thank the referees for having pointed out the related papers [P1, P2].

REFERENCES

[A] G. Alessandrini, *Singular solutions of elliptic equations and the determination of conductivity by boundary measurements*, J. Differential Equations, 84 (1990), pp. 252–272.

[A-B-R-V] G. Alessandrini, E. Beretta, E. Rosset, and S. Vessella, *Optimal stability for inverse elliptic boundary value problems with unknown boundaries*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 29 (2000), pp. 755–806.

[A-DB] G. Alessandrini and E. Di Benedetto, *Determining* 2*-dimensional cracks in* 3*-dimensional bodies: Uniqueness and stability*, Indiana Univ. Math. J., 46 (1997), pp. 1–82.

[A-I] G. Alessandrini and V. Isakov, *Analyticity and uniqueness for the inverse conductivity problem*, Rend. Istit. Mat. Univ. Trieste, 28 (1997), pp. 351–369.

[A-I-P] G. Alessandrini, V. Isakov, and J. Powell, *Local uniqueness in the inverse conductivity problem with one measurement*, Trans. Amer. Math. Soc., 347 (1995), pp. 3031–3041.

[C] A. P. Calderón, *On an inverse boundary value problem*, Seminar on Numerical Analysis and its Applications to Continuum Physics, Soc. Brasil. Mat., Rio de Janeiro, 1980, pp. 65–73.

[DB-E-F] E. Di Benedetto, C. Elliott, and A. Friedman, *The free boundary of a flow in a porous body heated from its boundary*, Nonlinear Anal., 10 (1986), pp. 879–900.

[DC-R] M. Di Cristo and L. Rondi, *Examples of exponential instability for inverse inclusion and scattering problems*, Inverse Problems, 19 (2003), pp. 685–701.

[F] A. Friedman, *Detection of mines by electric measurements*, SIAM J. Appl. Math., 47 (1987), pp. 201–212.

[F-G] A. Friedman and B. Gustafsson, *Identification of the conductivity coefficient in an elliptic equation*, SIAM J. Math. Anal., 18 (1987), pp. 777–787.

[I1] V. Isakov, *On uniqueness of recovery of a discontinuous conductivity coefficient*, Comm. Pure Appl. Math., 41 (1988), pp. 865–877.

[I2] V. Isakov, *Inverse problems for partial differential equations*, Springer-Verlag, New York, 1998.

[K-V] R. Kohn and M. Vogelius, *Determining conductivity by boundary measurements,* II*, Interior results*, Comm. Pure Appl. Math., 38 (1985), pp. 643–667.

[K-M] J. Korevaar and J. Meyers, *Logarithmic convexity for supremum norms of harmonic functions*, Bull. London Math. Soc., 26 (1994), pp. 353–362.

[K] I. Kukavica, *Quantitative uniqueness for second-order elliptic operators*, Duke Math. J., 91 (1998), pp. 225–240.

[L] P. Lax, *A stability theorem for solutions of abstract differential equations, and its applications to the study of the local behavior of solutions of elliptic equations*, Comm. Pure Appl. Math., 9 (1956), pp. 747–766.

[L-V] Y. Y. Li and M. Vogelius, *Gradient estimates for solutions to divergence form elliptic equations with discontinuous coefficients*, Arch. Ration. Mech. Anal., 153 (2000), pp. 91–151.

[Li] G. M. Lieberman, *Regularized distance and its applications*, Pacific J. Math., 117 (1985), pp. 329–352.

[L-S-W] W. Littman, G. Stampacchia, and H. Weinberger, *Regular points for elliptic equations with discontinuous coefficients*, Ann. Scuola Norm. Sup. Pisa 3, 17 (1963), pp. 43–77.

[M] C. Miranda, *Partial differential equations of elliptic type*, Springer-Verlag, New York-Berlin, 1970.

[P1] R. Potthast, *Stability estimates and reconstructions in inverse acoustic scattering using singular sources*, J. Comput. Appl. Math., 114 (2000), pp. 247–274.

[P2] R. Potthast, *Point sources and multipoles in inverse scattering theory*, Chapman & Hall, Boca Raton, FL, 2001.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 218–237

AXISYMMETRIC TE-MODES IN A SELF-FOCUSING DIELECTRIC*

CHARLES A. STUART† AND HUAN-SONG ZHOU‡

Abstract. We extend and improve earlier work on the existence of positive solutions $u \in H_0^1(0,\infty)$ of the nonlinear eigenvalue problem

$$u''(r) - \frac{3}{4r^2}u(r) + g\left(r, \frac{u(r)}{\sqrt{r}}\right)u(r) - k^2u(r) = 0 \quad \text{for } r > 0,$$

where $g \in C([0,\infty)^2)$ is such that

(i) $0 < A \leq g(r,s) \leq B < \infty$ for all $r, s \geq 0$, and

(ii) $g(r,s)$ is a nondecreasing function of s for each $r \geq 0$.

This problem is central to the study of guided TE-modes propagating in an axisymmetric, self-focusing dielectric such as an optical fiber. The function g is a nonlinear refractive index and we are now able to dispense with unnecessary restrictions about its behavior.

Key words. Nonlinear dielectric, guided waves, mountain pass theorem

AMS subject classifications. 35Q60, 78A60, 34B16

DOI. 10.1137/S0036141004441751

1. Introduction.

This paper is concerned with the existence of positive solutions $u \in H_0^1(0,\infty)$ of the nonlinear eigenvalue problem

$$u''(r) - \frac{3}{4r^2}u(r) + g\left(r, \frac{u(r)}{\sqrt{r}}\right)u(r) - k^2u(r) = 0 \quad \text{for } r > 0, \tag{1.1}$$

where $g \in C([0,\infty)^2)$ is such that

(i) $0 < A \leq g(r,s) \leq B < \infty$ for all $r, s \geq 0$, and

(ii) $g(r,s)$ is a nondecreasing function of s for each $r \geq 0$.

As was shown in [9], this problem is central to the study of guided transverse electric field modes (TE-modes) propagating in an axisymmetric, self-focusing dielectric such as an optical fiber. These modes are typical examples of what are more generally known as spatial solitons [8]. The derivation of (1.1) from Maxwell's equation as a model for guided waves in a self-focusing dielectric medium is reviewed in section 2.

The present contribution extends and improves our earlier work [12] on (1.1), which dealt only with the case of a homogeneous medium, $g(r,s) = g(s)$. In this situation, our main result, Theorem 5.2 in section 5, shows that (1.1) has a solution, provided that $g(0) < k^2 < \lim_{s\to\infty} g(s)$. In [12], we obtained the same conclusion, but the function g was required to satisfy several additional technical conditions, namely, (H1)–(H3) in Theorem 2.1 of [12], which we are now able to avoid. This improvement has been made possible by recent progress on several related problems [1, 4, 5, 13, 15], which will be discussed presently.

*Received by the editors March 4, 2004; accepted for publication (in revised form) November 30, 2004; published electronically September 20, 2005. This research was supported by NSFC and EPFL-IMA.
http://www.siam.org/journals/sima/37-1/44175.html

†IACS-FSB, Section de Mathématiques, Ecole Polytechnique Fédérale Lausanne, CH-1015 Lausanne, Switzerland (charles.stuart@epfl.ch).

‡Wuhan Institute of Physics and Mathematics, Chinese Academy of Sciences, P.O. Box 71010, Wuhan 430071, People's Republic of China (hszhou@wipm.ac.cn).

Our approach to (1.1) is based on the characterization of solutions of (1.1) as critical points of an energy functional $\Phi_k : H_0^1(0,\infty) \to \mathbb{R}$. For the values of k for which there are positive solutions, this functional is indefinite in the sense that $\inf \Phi_k = -\infty$ and $\sup \Phi_k = \infty$, and the critical points which we discuss, are saddle points of mountain pass type. Furthermore, by (i), all the terms in $\Phi_k(u)$ have quadratic growth as $u \to \infty$ in $H_0^1(0,\infty)$. As we pointed out in [12], this means that the main difficulty, which has to be overcome in using a mountain pass method, is to establish the boundedness in $H_0^1(0,\infty)$ of the Palais–Smale sequences, which are found due to the mountain pass geometry of Φ_k. It was for this step in [12] that we required the undesirable additional hypotheses about g mentioned above. Later, using an argument from [4], we were able to establish the boundedness of Palais–Smale sequences for a similar energy functional in [13] without such assumptions about g. We now return to the physical problem (1.1) and, by following the same approach, we are able to avoid the unnecessary restrictions on g and also to extend the discussion to inhomogeneous media. Compared to our work in [13], there is a further improvement here in that we are now able to avoid any assumptions about the behavior of $g(r,s)$ as $r \to \infty$. In [13] we assumed that $g(r,0)$ is independent of r and that $\lim_{r\to\infty} g_\infty(r)$ exists, where $g_\infty(r) = \lim_{s\to\infty} g(r,s)$. Again, this kind of situation has been treated recently in a similar type of problem [5].

In (1.1) the lack of compactness in the embeddings of $H_0^1(0,\infty)$ into spaces such as $L^p(0,\infty)$ is compensated for by the radial symmetry of the underlying problem, which means that g depends on $r^{-1/2}u(r)$ rather than simply on $u(r)$. The related work in [1, 4, 5] deals with elliptic equations without such radial symmetry and the lack of compactness is handled by arguments of the concentration-compactness type.

Finally, let us mention that (1.1) was first treated by a constrained variational principle in which k^2 occurs as a Lagrange multiplier [9]. In that method the constraint is the power of the guided mode and so it has the advantage of establishing the existence of such modes with prescribed power. On the other hand, it does not yield the existence of guided modes for all wave numbers in the admissible range. We have restricted our attention to positive solutions of (1.1) since they lead to the fundamental TE-mode but, as Ruppen has shown [7], an infinity of higher modes exist under suitable assumptions about g. Guided TE-modes can also propagate in a defocusing axisymmetric medium under appropriate conditions [6].

The rest of this paper is organized as follows. In the next section we show how (1.1) is related to TE-modes as special solutions of Maxwell's equations in a nonlinear dielectric. The variational formulation on which our work is based is set out in section 3 and the mountain pass structure of the associated energy functional is established in section 4. The main result is Theorem 5.2, which asserts the existence of solutions of (1.1) provided that $-\Lambda_0 < k^2 < -\Lambda_\infty$, where Λ_0 and Λ_∞ are numbers determined by the functions $g_0(r) = g(r,0)$ and $g_\infty(r) = \lim_{s\to\infty} g(r,s)$, respectively, in (4.2), (4.3). This result is established by showing that any Palais–Smale sequence (4.4), which arises from the mountain pass geometry of Φ_k, has a strongly convergent subsequence. Finally, we show that, if $k^2 \notin [-\Lambda_0, -\Lambda_\infty]$, then (1.1) has no positive solutions in $H_0^1(0,\infty)$.

2. Guided TE-modes. In a medium without free charges, Maxwell's equations can be written as

$$
\begin{array}{llll}
(1) & \partial_t B = -c\nabla \wedge E, & (2) & \partial_t D = c\nabla \wedge H, \\
(3) & \nabla \cdot B = 0, & (4) & \nabla \cdot D = 0,
\end{array} \tag{2.1}
$$

where $c > 0$ is the speed of light in a vacuum.

A field $F : \mathbb{R}^4 \to \mathbb{R}^3$ is monochromatic if it has the form

$$F(x,t) = F_1(x)\cos\omega t + F_2(x)\sin\omega t \quad \text{for } x \in \mathbb{R}^3 \text{ and } t \in \mathbb{R} \tag{2.2}$$

for some frequency $\omega > 0$ and functions $F_1, F_2 : \mathbb{R}^3 \to \mathbb{R}^3$. We are concerned with solutions of (2.1) in which all the fields are monochromatic.

In this case, the constitutive assumptions used in the study of wave propagation in optical fibers can be expressed in the form

$$H = B \text{ and } D(x,t) = \epsilon\left(\omega, x, \frac{1}{2}\left[|E_1(x)|^2 + |E_2(x)|^2\right]\right) E(x,t) \tag{2.3}$$

when $E(x,t) = E_1(x)\cos\omega t + E_2(x)\sin\omega t$, where the function $\epsilon : J \times \mathbb{R}^3 \times [0,\infty) \to (0,\infty)$ is the dielectric response for an isotropic dielectric medium without absorption for frequencies ω in the interval J.

When the medium is axisymmetric (as in an optical fiber), it is convenient to use cylindrical polar coordinates (r, θ, z), where the z-axis is chosen to be the axis of symmetry of the medium and we use the standard notation

$$i_r = (\cos\theta, \sin\theta, 0), \quad i_\theta = (-\sin\theta, \cos\theta, 0), \quad i_z = (0,0,1).$$

Then the response has the following properties:

$$\text{(A)} \begin{cases} \text{For } \omega \in J, \text{ there is a function } \varepsilon(\omega,\cdot,\cdot) : [0,\infty) \times [0,\infty) \to (0,\infty) \\ \text{such that } \epsilon(\omega, x, s) = \varepsilon(\omega, r, s), \text{ where } \varepsilon(\omega,\cdot) \in C([0,\infty)^2) \text{ and} \\ \text{there are constants } a(\omega) \text{ and } b(\omega) \text{ such that} \\ \qquad 0 < a(\omega) \le \varepsilon(\omega, r, s) \le b(\omega) < \infty \text{ for all } r, s \ge 0. \\ \text{Furthermore, } \varepsilon(\omega, r, s) \to \varepsilon(\omega, r, 0) \text{ as } s \to 0, \text{ uniformly for } r \ge 0. \end{cases}$$

A self-focusing medium is characterized by the following further property of the dielectric response:

(AS) For $\omega \in J$ and $r \ge 0$ fixed, $\varepsilon(\omega, r, s)$ is a nondecreasing function of s.

A field $F : \mathbb{R}^4 \to \mathbb{R}^3$ is a traveling wave if it has the form

$$F(x,t) = w(x - t\xi) \text{ for } x \in \mathbb{R}^3 \text{ and } t \in \mathbb{R} \tag{2.4}$$

for some nonzero vector $\xi \in \mathbb{R}^3$ and function $w : \mathbb{R}^3 \to \mathbb{R}^3$. Here $|\xi|$ is the wave speed and $\xi/|\xi|$ is the direction of propagation. For waves propagating in the direction of the axis of symmetry, we can suppose that $\xi = (0,0,s)$ for some wave speed $s > 0$.

A field $F : \mathbb{R}^4 \to \mathbb{R}^3$ is axisymmetric (with respect to the x_3-axis) if

$$F(\Gamma_\theta x, t) = \Gamma_\theta F(x,t) \text{ for } x \in \mathbb{R}^3 \text{ and } t \in \mathbb{R}$$

for all rotations Γ_θ around the x_3-axis.

A TE-mode is a solution of Maxwell's equations in which the electric field is an axisymmetric, monochromatic traveling wave which is everywhere transverse to the direction of propagation. In such modes, the electric field can be expressed as

$$E(x,t) = v(r) i_\theta \cos(kz - \omega t), \tag{2.5}$$

where $v : [0,\infty) \to \mathbb{R}$ is the amplitude and $k > 0$ is the wave number, related to the wave speed by $s = \omega/k$. To ensure that (2.5) defines a field which is continuously differentiable on $\mathbb{R}^4$, we must suppose that v has the properties

$$v \in C^1([0,\infty)) \quad \text{and} \quad v(0) = 0. \tag{2.6}$$

To establish the existence of TE-modes, one must show that there exist functions $v \not\equiv 0$ satisfying (2.6) and constants $k > 0$ and $\omega \in J$ such that (2.1), (2.3) are satisfied by electromagnetic fields, where E has the form (2.5). As is shown in [9], this reduces to satisfying the second order differential equation

$$\begin{gathered} v''(r) + \frac{1}{r}v'(r) - \frac{1}{r^2}v(r) + \left(\frac{\omega}{c}\right)^2 \varepsilon\left(\omega, r, \frac{1}{2}v(r)^2\right) v(r) - k^2 v(r) = 0 \quad \text{for } r > 0, \\ \text{where} \quad v \in C^2((0,\infty)) \text{ and } v(0) = 0. \end{gathered} \tag{2.7}$$

In fact, if v satisfies (2.7) and E is defined by (2.5), a complete solution of (2.1) and (2.3) is obtained by setting

$$\begin{aligned} D(x,t) &= \varepsilon\left(\omega, r, \frac{1}{2}v(r)^2\right) v(r) i_\theta \cos(kz - \omega t) \text{ and} \\ B(x,t) &= H(x,t) \\ &= \frac{c}{\omega}\left\{\frac{1}{r}[rv(r)]' i_z \sin(kz - \omega t) - kv(r) i_r \cos(kz - \omega t)\right\}. \end{aligned} \tag{2.8}$$

A guided TE-mode must also satisfy the following guidance conditions, which ensure that the electromagnetic energy is finite and that all the fields decay to zero as the distance from the axis of symmetry becomes infinite. As is shown in [9], these physical boundary conditions reduce to

$$\int_0^\infty \{v(r)^2 + v'(r)^2\}\, r dr < \infty, \tag{2.9}$$

$$v(r) \to 0 \text{ and } v'(r) \to 0 \text{ as } r \to \infty \tag{2.10}$$

for TE-modes in the form (2.5).

From now on we fix a value of ω in J and try to find pairs (k, v) such that (2.7)–(2.10) are satisfied. To simplify the notation we set

$$g(r,s) = \left(\frac{\omega}{c}\right)^2 \varepsilon\left(\omega, r, \frac{1}{2}s^2\right) \text{ for } r \geq 0 \text{ and } s \in \mathbb{R} \tag{2.11}$$

and

$$u(r) = r^{1/2} v(r). \tag{2.12}$$

Equation (2.7) becomes

$$u''(r) - \frac{3}{4r^2}u(r) + g\left(r, \frac{u(r)}{\sqrt{r}}\right) u(r) - k^2 u(r) = 0 \quad \text{for } r > 0 \tag{2.13}$$

and, as we verify in Theorem 5.2, the other conditions in (2.7)–(2.10) are all satisfied provided that u belongs to the Sobolev space

$$H = H_0^1(0,\infty) = \{u \in H^1(0,\infty) : u(0) = 0\}. \tag{2.14}$$

More precisely, if $(k,u)\in(0,\infty)\times H$ and (2.13) is satisfied in the weak sense, namely,

(2.15)
$$\int_0^\infty u'(r)\varphi'(r)dr=\int_0^\infty\left\{g\left(r,\frac{u(r)}{\sqrt{r}}\right)-k^2-\frac{3}{4r^2}\right\}u(r)\varphi(r)dr \text{ for all } \varphi\in C_0^\infty((0,\infty)),$$

then the function v defined by (2.12) satisfies (2.7)–(2.10) and, consequently, it generates a guided TE-mode through the formulae (2.5) and (2.8).

The remainder of the paper deals with problem (2.13) so, for future reference, we recall some basic properties of elements of H.

We reserve $\langle\cdot,\cdot\rangle$ and $|\cdot|_2$ as notation for the usual scalar product and norm on $L^2(0,\infty)$ and, more generally, $|\cdot|_p$ denotes the usual norm on $L^p(0,\infty)$ for $1\le p\le\infty$. Then H is a Hilbert space with scalar product

$$\langle u,v\rangle_1=\langle u',v'\rangle+\langle u,v\rangle=\int_0^\infty[u'v'+uv]dr \tag{2.16}$$

and norm

$$\|u\|_1=\{|u'|_2^2+|u|_2^2\}^{1/2}. \tag{2.17}$$

An element $z\in H$ is, after modification on a set of measure zero, continuous on $[0,\infty)$ and satisfies the inequalities

$$|z|_\infty^2\le|z|_2|z'|_2, \tag{2.18}$$

$$r^{-1/2}|z(r)|\le|z'|_2 \text{ for all } r>0, \tag{2.19}$$

$$\lim_{r\to0}r^{-1/2}z(r)=0, \tag{2.20}$$

$$|z/r|_2\le2|z'|_2 \quad \text{(Hardy's inequality)}, \tag{2.21}$$

and

$$\int_0^\infty\left[z'+\frac{z}{2r}\right]^2dr=\int_0^\infty\left\{(z')^2+\frac{3z^2}{4r^2}\right\}dr\le4|z'|_2^2. \tag{2.22}$$

3. Variational formulation. Positive solutions of (2.15) are the critical points of a functional $\Phi_k:H\to\mathbb{R}$, which we now introduce. With g defined by (2.11) for a response satisfying conditions (A) and (AS), we set

$$g_0(r)=g(r,0) \text{ and } g_\infty(r)=\lim_{s\to\infty}g(r,s), \tag{3.1}$$

$$\widetilde{g}(r,s)=g(r,s)-g_0(r), \tag{3.2}$$

and then

$$\overline{G}(r,s)=\int_0^s g(r,\tau^+)\tau d\tau=\frac{1}{2}g_0(r)s^2+\widetilde{G}(r,s) \text{ with } \widetilde{G}(r,s)=\int_0^s\widetilde{g}(r,\tau^+)\tau d\tau$$

$$\text{and}\quad G(r,s)=\int_0^s g(r,r^{-1/2}\tau^+)\tau d\tau=r\overline{G}(r,r^{-1/2}s) \tag{3.3}$$

$$=\frac{1}{2}g_0(r)s^2+r\widetilde{G}(r,r^{-1/2}s). \tag{3.4}$$

Note that, for all $r > 0$, $G(r, \cdot) \in C^1(\mathbb{R})$ with

$$\partial_s G(r,s) = g(r, r^{-1/2}s^+)s$$

and

$$\frac{1}{2}A(\omega)s^2 \leq \frac{1}{2}g_0(r)s^2 \leq G(r,s) \leq \frac{1}{2}g_\infty(r)s^2 \leq \frac{1}{2}B(\omega)s^2 \tag{3.5}$$

for all $r > 0$ and $s \in \mathbb{R}$, where $A(\omega) = \left(\frac{\omega}{c}\right)^2 a(\omega)$ and $B(\omega) = \left(\frac{\omega}{c}\right)^2 b(\omega)$.

Recalling (2.21), we can now define a functional Φ_k on H by setting

$$\Phi_k(u) = \frac{1}{2}\int_0^\infty \left\{ |u'(r)|^2 + \frac{3}{4r^2}u(r)^2 + k^2 u(r)^2 \right\} dr - \int_0^\infty G(r, u(r))dr$$

for $u \in H$.

Lemma 3.1. *Under assumptions* (A) *and* (AS), *we have the following:*

(a) $\Phi_k \in C^1(H, \mathbb{R})$ *and*

$$\Phi_k'(u)v = \int_0^\infty u'(r)v'(r)dr - \int_0^\infty \left\{ g\left(r, \frac{u(r)^+}{\sqrt{r}}\right) - k^2 - \frac{3}{4r^2} \right\} u(r)v(r)dr \tag{3.6}$$

for all $u, v \in H$.

(b) *Furthermore, for* $t > 0$ *and* $u \in H$,

$$\Phi_k(tu) \leq \frac{1}{2}(t^2 - 1)\Phi_k'(u)u + \Phi_k(u).$$

Remark 3.1. It follows from (3.6) that if $\Phi_k'(u) = 0$ and $u \geq 0$, then (k, u) satisfies (2.15).

Proof. (a) First, we note that $G : (0, \infty) \times \mathbb{R} \to \mathbb{R}$ is a function of Carathéodory type having the following properties:

(i) $G(r, \cdot) \in C^1(\mathbb{R})$ for all $r > 0$.

(ii) $0 \leq G(r,s) \leq \frac{1}{2}B(\omega)s^2$ for all $s \in \mathbb{R}$.

(iii) $|\partial_s G(r,s)| \leq B(\omega)|s|$ for all $s \in \mathbb{R}$.

Using standard results about Nemytskii operators [14], it follows easily that

$$u \longmapsto \int_0^\infty G(r, u(r))dr$$

is a continuously Fréchet differentiable mapping of $L^2(0, \infty)$ into $\mathbb{R}$. (See Theorem 2.8 of [2], for example.) Since H is continuously embedded in $L^2(0, \infty)$, the same is true when it is considered as mapping from H into $\mathbb{R}$.

Using (2.21), it follows immediately that

$$\int_0^\infty \left\{ u'v' + \frac{3}{4r^2}uv + k^2 uv \right\} dr$$

is a bounded, symmetric bilinear form on H. Thus

$$u \longmapsto \int_0^\infty \left\{ u'(r)^2 + \frac{3}{4r^2}u(r)^2 + k^2 u(r)^2 \right\} dr$$

is a continuously Fréchet differentiable mapping of H into $\mathbb{R}$.

Combining these observations we see that $\Phi_k \in C^1(H, \mathbb{R})$.

(b) For $t > 0$ and $u \in H$,

$$\begin{aligned}
\Phi_k(tu) &= \frac{t^2}{2} \int_0^\infty \left\{ u'(r)^2 + \frac{3}{4r^2} u(r)^2 + k^2 u(r)^2 \right\} dr - \int_0^\infty G(r, tu(r)) dr \\
&= \frac{t^2}{2} \left[\Phi_k'(u)u + \int_0^\infty g\left(r, \frac{u(r)^+}{\sqrt{r}}\right) u(r)^2 dr \right] - \int_0^\infty G(r, tu(r)) dr \\
&= \frac{t^2}{2} \Phi_k'(u)u + \int_0^\infty \left\{ \frac{1}{2} g\left(r, \frac{u(r)^+}{\sqrt{r}}\right) t^2 u(r)^2 - G(r, tu(r)) \right\} dr.
\end{aligned}$$

For $r > 0$ fixed, set

$$h(t) = \frac{1}{2} g\left(r, \frac{u(r)^+}{\sqrt{r}}\right) t^2 u(r)^2 - G(r, tu(r)).$$

Then

$$\begin{aligned}
h'(t) &= g\left(r, \frac{u(r)^+}{\sqrt{r}}\right) tu(r)^2 - [\partial_s G(r, tu(r))] u(r) \\
&= g\left(r, \frac{u(r)^+}{\sqrt{r}}\right) tu(r)^2 - g(r, r^{-1/2} tu(r)^+) tu(r)^2 \\
&= tu(r)^2 \left\{ g\left(r, \frac{u(r)^+}{\sqrt{r}}\right) - g\left(r, \frac{tu(r)^+}{\sqrt{r}}\right) \right\}.
\end{aligned}$$

The monotonicity of $g(r, \cdot)$ shows that $h'(t) \geq 0$ for $0 < t \leq 1$ and $h'(t) \leq 0$ for $t > 1$. Consequently, $h(t) \leq h(1)$ for all $t > 0$ and so

$$\frac{1}{2} g\left(r, \frac{u(r)^+}{\sqrt{r}}\right) t^2 u(r)^2 - G(r, tu(r)) \leq \frac{1}{2} g\left(r, \frac{u(r)^+}{\sqrt{r}}\right) u(r)^2 - G(r, u(r)).$$

Thus

$$\begin{aligned}
\Phi_k(tu) &\leq \frac{t^2}{2} \Phi_k'(u)u + \int_0^\infty \left\{ \frac{1}{2} g\left(r, \frac{u(r)^+}{\sqrt{r}}\right) u(r)^2 - G(r, u(r)) \right\} dr \\
&= \frac{t^2}{2} \Phi_k'(u)u + \Phi_k(u) - \frac{1}{2} \Phi_k'(u)u. \qquad \square
\end{aligned}$$

4. Mountain pass geometry. Recalling (3.5), we set

(4.1)
$$\Lambda_0 = \inf \left\{ \int_0^\infty \left(|u'(r)|^2 + \left[\frac{3}{4r^2} - g_0(r) \right] u(r)^2 \right) dr : u \in H \text{ and } \int_0^\infty u(r)^2 dr = 1 \right\}$$

(4.2)
$$\Lambda_\infty = \inf \left\{ \int_0^\infty \left(|u'(r)|^2 + \left[\frac{3}{4r^2} - g_\infty(r) \right] u(r)^2 \right) dr : u \in H \text{ and } \int_0^\infty u(r)^2 dr = 1 \right\}$$

and observe that

$$(k^2 + \Lambda_\infty) \int_0^\infty u(r)^2 dr \leq \Phi_k(u) \leq (k^2 + \Lambda_0) \int_0^\infty u(r)^2 dr$$

for all $u \in H$. Furthermore, since

$$\inf\left\{\int_0^\infty \left(|u'(r)|^2 + \frac{3}{4r^2}u(r)^2\right) dr : u \in H \text{ and } \int_0^\infty u(r)^2 dr = 1\right\} = 0,$$

as is easily seen using (2.21), it follows from (3.5) that

$$-\frac{1}{2}B(\omega) \le \Lambda_\infty \le \Lambda_0 \le -\frac{1}{2}A(\omega) < 0. \tag{4.3}$$

LEMMA 4.1. *Consider* $k^2 > -\Lambda_0$. *Then*

(a) $\left(\int_0^\infty \left\{|u'(r)|^2 + \left[\frac{3}{4r^2} + k^2 - g_0(r)\right] u(r)^2\right\} dr\right)^{1/2}$

defines a norm, denoted by $\|u\|_0$, *which is equivalent to the usual norm* $\|u\|_1$ *on* H *defined by* (2.17).

(b) *There exist constants* $\alpha, \rho > 0$ *such that*

$$\Phi_k(u) \ge \alpha > 0 \textit{ for all } u \in S_\rho = \{u \in H : \|u\|_0 = \rho\}.$$

Remark 4.1. The norm $\|u\|_0$ is derived from a scalar product on H which we denote by $\langle\cdot,\cdot\rangle_0$.

Proof. (a) First we note that

$$\begin{aligned}\int_0^\infty &\left\{|u'(r)|^2 + \left[\frac{3}{4r^2} + k^2 - g_0(r)\right] u(r)^2\right\} dr \\ &\le \int_0^\infty \{|u'(r)|^2 + 3|u'(r)|^2 + k^2 u(r)^2\} dr \\ &\le \max\{4, k^2\} \|u\|_1^2 \text{ for all } u \in H\end{aligned}$$

by (2.21) and $g_0 \ge 0$.

On the other hand, for $t \in (0,1)$,

$$\begin{aligned}&\int_0^\infty \left\{|u'(r)|^2 + \left[\frac{3}{4r^2} + k^2 - g_0(r)\right] u(r)^2\right\} dr \\ &\quad = k^2 \int_0^\infty u(r)^2 dr + t\int_0^\infty \left\{|u'(r)|^2 + \left[\frac{3}{4r^2} - g_0(r)\right] u(r)^2\right\} dr \\ &\qquad\qquad + (1-t)\int_0^\infty \left\{|u'(r)|^2 + \left[\frac{3}{4r^2} - g_0(r)\right] u(r)^2\right\} dr \\ &\quad \ge (k^2 + t\Lambda_0)\int_0^\infty u(r)^2 dr + (1-t)\int_0^\infty \left\{|u'(r)|^2 + \left[\frac{3}{4r^2} - g_0(r)\right] u(r)^2\right\} dr \\ &\quad \ge \{k^2 + t\Lambda_0 - (1-t)|g_0|_\infty\}\int_0^\infty u(r)^2 dr + (1-t)\int_0^\infty |u'(r)|^2 dr\end{aligned}$$

for all $u \in H$. Since $k^2 + \Lambda_0 > 0$, we can choose $t_k \in (0,1)$ such that $k^2 + t_k\Lambda_0 - (1 - t_k)|g_0|_\infty > 0$, and then

$$\begin{aligned}\int_0^\infty &\left\{|u'(r)|^2 + \left[\frac{3}{4r^2} + k^2 - g_0(r)\right] u(r)^2\right\} dr \\ &\ge \min\{k^2 + t\Lambda_0 - (1-t)|g_0|_\infty, 1 - t_k\} \|u\|_1^2 \text{ for all } u \in H.\end{aligned}$$

(b) By (2.19) and part (a), there is a constant C such that

$$r^{-1/2}\left|u(r)\right| \leq \left\|u\right\|_1 \leq C\left\|u\right\|_0 \text{ for all } r>0 \text{ and } u\in H.$$

It follows from (A) that, given any $\eta>0$, there exists $\gamma>0$ such that

$$g\left(r,\frac{u(r)}{\sqrt{r}}\right) \leq g(r,0)+\eta = g_0(r)+\eta$$

for all $r>0$ and all $u\in H$ with $\left\|u\right\|_0<\gamma$. Hence

$$G(r,u(r)) = r\int_0^{u(r)/\sqrt{r}} g(r,\tau^+)\tau d\tau \leq \frac{1}{2}\{g_0(r)+\eta\}u(r)^2$$

and so, for $\left\|u\right\|_0<\gamma$,

$$\begin{aligned}
\Phi_k(u) &= \frac{1}{2}\int_0^\infty \left\{|u'(r)|^2+\left[\frac{3}{4r^2}+k^2\right]u(r)^2\right\}dr - \int_0^\infty G(r,u(r))dr \\
&\geq \frac{1}{2}\int_0^\infty \left\{|u'(r)|^2+\frac{3}{4r^2}u(r)^2+[k^2-g_0(r)-\eta]u(r)^2\right\}dr \\
&= \frac{1}{2}\left\|u\right\|_0^2-\frac{\eta}{2}\left|u\right|_2^2 \\
&\geq \frac{1}{2}\left\|u\right\|_0^2-\frac{\eta C^2}{2}\left\|u\right\|_0^2 = \frac{1}{2}(1-\eta C^2)\left\|u\right\|_0^2.
\end{aligned}$$

Choosing $\eta=1/2C^2$, it suffices to set $\rho=\gamma/2$ and $\alpha=\rho^2/4$. ∎

LEMMA 4.2. *Suppose that* $-\Lambda_0<k^2<-\Lambda_\infty$. *There exists* $e\in H$ *such that*

$$\left\|e\right\|_0>\rho \text{ and } \Phi_k(e)<0,$$

where ρ *is given by Lemma* 4.1(b).

Proof. Choose $\varepsilon>0$ such that $k^2+\Lambda_\infty+\varepsilon<0$. By the definition of Λ_∞, there exists an element $w\in H,\ w\not\equiv 0$ such that

$$\int_0^\infty \left\{|w'(r)|^2+\left[\frac{3}{4r^2}-g_\infty(r)\right]w(r)^2\right\}dr < (\Lambda_\infty+\varepsilon)\int_0^\infty w(r)^2dr$$

and, replacing w by $|w|$, we may assume that $w\geq 0$. Now, for $t>0$,

$$\frac{\Phi_k(tw)}{t^2} = \frac{1}{2}\int_0^\infty \left\{|w'(r)|^2+\left[\frac{3}{4r^2}+k^2\right]w(r)^2\right\}dr - t^{-2}\int_0^\infty G(r,tw(r))dr,$$

where

$$0\leq t^{-2}G(r,tw(r)) \leq \frac{1}{2}B(\omega)w(r)^2$$

by (3.5). The dominated convergence theorem implies that

$$\lim_{t\to\infty}\frac{\Phi_k(tw)}{t^2} = \frac{1}{2}\int_0^\infty \left\{|w'(r)|^2+\left[\frac{3}{4r^2}+k^2\right]w(r)^2\right\}dr - \int_0^\infty \lim_{t\to\infty} t^{-2}G(r,tw(r))dr.$$

For $r,s>0$,

$$s^{-2}G(r,s) = rs^{-2}\int_0^{s/\sqrt{r}} g(r,\tau)\tau d\tau = \int_0^1 g\left(r,\frac{s\sigma}{\sqrt{r}}\right)\sigma d\sigma \to g_\infty(r)\int_0^1 \sigma d\sigma = \frac{1}{2}g_\infty(r)$$

as $s \to \infty$ by dominated convergence. Hence

$$\lim_{t\to\infty} t^{-2}G(r,tw(r)) = \lim_{t\to\infty}[tw(r)]^{-2}G(r,tw(r))w(r)^2 = \frac{1}{2}g_\infty(r)w(r)^2$$

if $w(r) > 0$, whereas $\lim_{t\to\infty} t^{-2}G(r,tw(r)) = 0$ if $w(r) = 0$. Thus, since $w \geq 0$, we have that

$$\begin{aligned}\lim_{t\to\infty}\frac{\Phi_k(tw)}{t^2} &= \frac{1}{2}\int_0^\infty \left\{|w'(r)|^2 + \left[\frac{3}{4r^2}+k^2\right]w(r)^2\right\}dr - \int_0^\infty \left[\frac{1}{2}g_\infty(r)w(r)^2\right]dr \\ &= \frac{1}{2}\int_0^\infty \left\{|w'(r)|^2 + \left[\frac{3}{4r^2}-g_\infty(r)\right]w(r)^2\right\}dr + \frac{1}{2}k^2\int_0^\infty w(r)^2dr \\ &< \frac{1}{2}(\Lambda_\infty+\varepsilon+k^2)\int_0^\infty w(r)^2dr < 0\end{aligned}$$

by the choice of w and ε. It suffices to set $e = tw$ for t large enough. □

Lemmas 4.1 and 4.2 show that for $-\Lambda_0 < k^2 < -\Lambda_\infty$, the functional Φ_k has what is called the mountain pass geometry. Indeed, setting

$$\begin{aligned}c &= \inf_{\gamma\in\Gamma}\max_{t\in[0,1]}\Phi_k(\gamma(t)), \text{ where}\\ \Gamma &= \{\gamma \in C([0,1],H) : \gamma(0) = 0 \text{ and } \gamma(1) = e\},\end{aligned}$$

the mountain pass geometry described in Lemmas 4.1 and 4.2 implies that

$$\text{there exists a sequence } \{u_n\} \subset H \text{ such that} \tag{4.4}$$

$$\Phi_k(u_n) \xrightarrow{n} c > 0, \tag{4.5}$$

$$\Phi_k'(u_n)u_n \xrightarrow{n} 0, \tag{4.6}$$

$$\Phi_k'(u_n)\varphi \xrightarrow{n} 0, \text{ for all } \varphi \in H. \tag{4.7}$$

The properties (4.5) and (4.7) follow from the existence of what is often called a Palais–Smale sequence at level c, and the fact that the additional property (4.6) is also available was first shown by Cerami; see [3, Chapter IV], for example. If $\{u_n\}$ has a subsequence converging strongly u in H,

$$\Phi_k'(u_n)\varphi \xrightarrow{n} 0 \text{ for all } \varphi \in H$$

and we have a critical point of $\Phi_k$with $\Phi_k(u) = c$, so $u \not\equiv 0$. In the next section we show that $\{u_n\}$ has such a subsequence.

5. Existence of a ground state. We begin by establishing some further properties of the sequence (4.4)–(4.7). The main problems are to show that

(i) $\{u_n\}$ is bounded in H_0^1, and then

(ii) $u_n \xrightarrow{n} u$ weakly in $H \Longrightarrow u_n \xrightarrow{n} u$ strongly in H.

LEMMA 5.1. *The sequence* (4.4)–(4.7) *is bounded in H.*

Proof. If $\{u_n\}$ is not bounded in H, there is a subsequence, which we can still denote by $\{u_n\}$, such that $\|u_n\|_0 \xrightarrow{n} \infty$.

Next we set

$$w_n = 2\sqrt{c}\frac{u_n}{\|u_n\|_0} \text{ and } t_n = \frac{2\sqrt{c}}{\|u_n\|_0}. \tag{5.1}$$

By passing to a further subsequence, we can assume that $w_n \overset{n}{\rightharpoonup} w$ weakly in H. The rest of the proof is split into the following steps. First we show that $w \not\equiv 0$, and then we deduce that $w > 0$. Finally we obtain a contradiction, showing that $\{u_n\}$ must be bounded in H.

Step 1 ($w \not\equiv 0$). By Lemma 3.1(b),

$$\Phi_k(w_n) = \Phi_k(t_n u_n) \leq \frac{1}{2}(t_n^2 - 1)\Phi_k'(u_n)u_n + \Phi_k(u_n)$$

and so

$$\limsup_{n\to\infty} \Phi_k(w_n) \leq c \tag{5.2}$$

by (4.5) and (4.6) since $t_n \to 0$.

On the other hand, in the notation (3.4),

$$\begin{aligned}\Phi_k(u) &= \frac{1}{2}\int_0^\infty \left\{|u'(r)|^2 + \left[\frac{3}{4r^2} + k^2 - g_0(r)\right]u(r)^2\right\}dr - \int_0^\infty r\widetilde{G}(r, r^{-1/2}u(r))dr \\ &= \frac{1}{2}\|u\|_0^2 - \int_0^\infty r\widetilde{G}(r, r^{-1/2}u(r))dr,\end{aligned}$$

where

$$\begin{aligned}0 \leq r\widetilde{G}(r, r^{-1/2}u(r)) &= r\int_0^{u(r)/\sqrt{r}} \{g(r,\tau^+) - g_0(r)\}\tau d\tau \\ &\leq r\{g_\infty(r) - g_0(r)\}\int_0^{u(r)/\sqrt{r}} \tau d\tau \leq \frac{1}{2}B(\omega)u(r)^2.\end{aligned}$$

Hence for any $R > 0$,

$$\begin{aligned}0 &\leq \int_0^\infty r\widetilde{G}(r, r^{-1/2}u(r))dr \\ &\leq \int_0^R \frac{1}{2}B(\omega)u(r)^2 dr + \int_R^\infty r\widetilde{G}(r, r^{-1/2}u(r))dr,\end{aligned}$$

and for $r \geq R$,

$$\begin{aligned}0 \leq r\widetilde{G}(r, r^{-1/2}u(r)) &= r\int_0^{u(r)/\sqrt{r}} \widetilde{g}(r,\tau^+)\tau dt \\ &\leq r\widetilde{g}(r, r^{-1/2}u(r)^+)\int_0^{u(r)/\sqrt{r}} \tau dt \leq \frac{1}{2}\widetilde{g}(r, r^{-1/2}|u|_\infty)u(r)^2\end{aligned}$$

by the monotonicity of the function $\widetilde{g}(r,s) = g(r,s) - g_0(r)$ in s.

Hence

$$\begin{aligned}0 &\leq \int_0^\infty r\widetilde{G}(r, r^{-1/2}u(r))dr \\ &\leq \int_0^R \frac{1}{2}B(\omega)u(r)^2 dr + \int_R^\infty \frac{1}{2}\widetilde{g}(r, r^{-1/2}|u|_\infty)u(r)^2 dr \\ &\leq \frac{1}{2}B(\omega)\int_0^R u(r)^2 dr + \frac{1}{2}\sup_{r\geq R}\widetilde{g}(r, r^{-1/2}|u|_\infty)\int_R^\infty u(r)^2 dr. \qquad (5.3)\end{aligned}$$

Since $\|w_n\|_0 = 2\sqrt{c}$ for all n, there is a constant $K > 0$ such that $|w_n|_2 \leq K$ and $|w_n|_\infty \leq K$. Thus putting $u = w_n$ in (5.3),

$$\begin{aligned} 0 &\leq \int_0^\infty r\widetilde{G}(r, r^{-1/2} w_n(r))dr \\ &\leq \frac{1}{2} B(\omega) \int_0^R w_n(r)^2 dr + \frac{1}{2} \sup_{r \geq R} \widetilde{g}(r, r^{-1/2} K) K^2 \end{aligned}$$

for all n. If $w_n \overset{n}{\rightharpoonup} 0$ weakly in H, $\int_0^R w_n(r)^2 dr \xrightarrow{n} 0$ for any $R > 0$, and so

$$\limsup_{n\to\infty} \int_0^\infty r\widetilde{G}(r, r^{-1/2} w_n(r))dr \leq \frac{1}{2} \sup_{r \geq R} \widetilde{g}(r, r^{-1/2} K) K^2$$

for all $R > 0$. But it follows from (A) that

$$\lim_{R\to\infty} \sup_{r \geq R} \widetilde{g}(r, r^{-1/2} K) = 0,$$

so in fact,

$$\limsup_{n\to\infty} \int_0^\infty r\widetilde{G}(r, r^{-1/2} w_n(r))dr = 0.$$

Finally,

$$\begin{aligned} \liminf_{n\to\infty} \Phi_k(w_n) &= \liminf_{n\to\infty} \left\{ \frac{1}{2} \|w_n\|_0^2 - \int_0^\infty r\widetilde{G}(r, r^{-1/2} w_n(r))dr \right\} \\ &= 2c - \limsup_{n\to\infty} \int_0^\infty r\widetilde{G}(r, r^{-1/2} w_n(r))dr \\ &\geq 2c, \end{aligned}$$

contradicting (5.2). Hence $w \not\equiv 0$.

Step 2 ($w > 0$). For all $\varphi \in C_0^\infty((0,\infty))$,

$$\begin{aligned} \langle w_n, \varphi \rangle_0 &= \int_0^\infty \left\{ w_n' \varphi' + \left[k^2 + \frac{3}{4r^2} - g_0(r) \right] w_n \varphi \right\} dr \\ &= t_n \int_0^\infty \left\{ u_n' \varphi' + \left[k^2 + \frac{3}{4r^2} - g_0(r) \right] u_n \varphi \right\} dr \\ &= t_n \left\{ \Phi_k'(u_n)\varphi + \int_0^\infty \widetilde{g}(r, r^{-1/2} u_n(r)^+) u_n \varphi dr \right\} \\ &= t_n \Phi_k'(u_n)\varphi + \int_0^\infty \widetilde{g}(r, r^{-1/2} u_n(r)^+) w_n \varphi dr. \end{aligned}$$

By (4.7), $t_n \Phi_k'(u_n)\varphi \to 0$ since $t_n \to 0$. Furthermore,

$$\begin{aligned} \widetilde{g}(r, r^{-1/2} u_n(r)^+) w_n(r) &= \widetilde{g}\left(r, \frac{w_n(r)^+}{t_n \sqrt{r}}\right) w_n(r) \\ &\to \begin{cases} [g_\infty(r) - g_0(r)] w(r) & \text{if } w(r) > 0, \\ 0 & \text{if } w(r) \leq 0 \end{cases} \end{aligned}$$

as $n \to \infty$ and

$$\left| \widetilde{g}\left(r, \frac{w_n(r)^+}{t_n\sqrt{r}}\right) w_n(r) \right| \le B(\omega)K \text{ for all } r > 0.$$

Hence, by dominated convergence we have that

$$\int_0^\infty \widetilde{g}(r, r^{-1/2}u_n(r)^+)w_n\varphi dr \xrightarrow{n} \int_0^\infty [g_\infty(r) - g_0(r)]w(r)^+\varphi(r)dr$$

and, consequently,

$$\langle w, \varphi\rangle_0 = \int_0^\infty [g_\infty(r) - g_0(r)]w(r)^+\varphi(r)dr \tag{5.4}$$

for all $\varphi \in C_0^\infty((0,\infty))$. But $C_0^\infty((0,\infty))$ is dense in H and it follows easily that (5.4) holds for all $\varphi \in H$. Putting $\varphi = w^-$ in (5.4), we obtain

$$\left\|w^-\right\|_0 = \left\langle w, w^-\right\rangle_0 = 0,$$

showing that $w^- \equiv 0$. Hence $w \equiv w^+$ and (5.4) can be written as

$$\int_0^\infty \left\{ w'\varphi' + \left[k^2 + \frac{3}{4r^2} - g_\infty(r)\right] w\varphi \right\} dr = 0 \text{ for all } \varphi \in C_0^\infty((0,\infty)). \tag{5.5}$$

It follows that w' has a generalized derivative on $(0,\infty)$ and that

$$w''(r) = \left\{ k^2 + \frac{3}{4r^2} - g_\infty(r) \right\} w(r) \text{ on } (0,\infty). \tag{5.6}$$

Since $w \ge 0$ on $(0,\infty)$, but $w \not\equiv 0$, the maximum principle implies that $w > 0$ on $(0,\infty)$.

Step 3 ($\{u_n\}$ is bounded). We show that, for $k^2 > -\Lambda_\infty$, (5.5) cannot have a positive solution.

For $R > 0$, set

$$m(R) = \inf \left\{ \int_0^R \left(|u'(r)|^2 + \left[\frac{3}{4r^2} - g_\infty(r)\right] u(r)^2 \right) dr : \ u \in H_0^1(0,R) \text{ and } \int_0^R u(r)^2 dr = 1 \right\}.$$

Recalling that $g_\infty \in L^\infty$ and that $H_0^1(0,R)$ is compactly embedded in $L^2(0,R)$, it follows easily that there exists an element $u_R \in H_0^1(0,R)\backslash\{0\}$ such that

$$\int_0^R \left\{ |u_R'(r)|^2 + \left[\frac{3}{4r^2} - g_\infty(r)\right] u_R(r)^2 \right\} dr = m(R) \int_0^R u_R(r)^2 dr$$

and, replacing u_R by $|u_R|$, we can suppose that $u_R \ge 0$ on $[0,R]$. But there is a Lagrange multiplier λ such that

$$\int_0^R \left\{ u_R'\varphi' + \left[\frac{3}{4r^2} - g_\infty(r)\right] u_R\varphi \right\} dr = \lambda \int_0^R u_R\varphi dr \text{ for all } \varphi \in H_0^1(0,R)$$

and, putting $\varphi = u_R$, we have $\lambda = m(R)$. Thus, u_R has a generalized derivative on $(0, R)$ and

$$-u_R''(r) + \left[\frac{3}{4r^2} - g_\infty(r)\right] u_R(r) = m(R) u_R(r) \text{ on } (0, R), \tag{5.7}$$

$$u_R(0) = u_R(R) = 0. \tag{5.8}$$

It follows that $u_R \in H^2(\delta, R)$ for all $\delta \in (0, R)$. We do not claim that $u_R \in H^2(0, R)$ but we can assume that $u_R \in C^1((0, R])$. Since $k^2 < -\Lambda_\infty$, we can choose $\varepsilon > 0$ such that $k^2 + 2\varepsilon < -\Lambda_\infty$. By the definition of Λ_∞, there exists $z \in H\backslash\{0\}$ such that

$$\int_0^\infty \left\{|z'(r)|^2 + \left[\frac{3}{4r^2} - g_\infty(r)\right] z(r)^2\right\} dr \le (\Lambda_\infty + \varepsilon) \int_0^\infty z(r)^2 dr$$

and then, by (2.21) and the density of $C_0^\infty((0, \infty))$ there exists an element $v \in C_0^\infty((0, \infty))$ such that

$$\int_0^\infty \left\{|v'(r)|^2 + \left[\frac{3}{4r^2} - g_\infty(r)\right] v(r)^2\right\} dr \le (\Lambda_\infty + 2\varepsilon) \int_0^\infty v(r)^2 dr.$$

Now choose R such that $\operatorname{supp} v \subset (0, R)$. It follows that $m(R) \le \Lambda_\infty + 2\varepsilon$. From (5.7) we obtain

$$-u_R'(R)w(R) + u_R'(\delta)w(\delta) + \int_\delta^R u_R' w' + \left\{\frac{3}{4r^2} - g_\infty(r)\right\} u_R w dr = m(R) \int_\delta^R u_R w dr$$

for any $\delta \in (0, R)$. Hence

$$\lim_{\delta \to 0} u_R'(\delta)w(\delta) = u_R'(R)w(R) - \int_0^R u_R' w' + \left\{\frac{3}{4r^2} - g_\infty(r) - m(R)\right\} u_R w dr$$

since $r^{-2} u_R(r) w(r)$ is integrable on $(0, R)$ by (2.21).

By (2.21) and the density of $C_0^\infty((0, \infty))$ in H, it follows that (5.5) holds for all $\varphi \in H$. We denote by $\widetilde{u_R}$ the function obtained by extending u_R by zero for all $r \ge R$. Setting $\varphi = \widetilde{u_R}$ in (5.5), we have that

$$\int_0^R u_R' w' dr + \int_0^R \left\{\frac{3}{4r^2} - g_\infty(r)\right\} u_R w dr = -k^2 \int_0^R u_R w dr.$$

Hence,

$$\begin{aligned}
\lim_{\delta \to 0} u_R'(\delta)w(\delta) &= u_R'(R)w(R) + \{m(R) + k^2\} \int_0^R u_R w dr \\
&\le u_R'(R)w(R) + \{\Lambda_\infty + 2\varepsilon + k^2\} \int_0^R u_R w dr \\
&< u_R'(R)w(R)
\end{aligned}$$

by the positivity of w and u_R and the choice of ε and R. Since $u_R \ge 0$ on $[0, R]$ and $u_R(R) = 0$, we have $u_R'(R)w(R) \le 0$ and, consequently, there exists $\delta_0 > 0$ such that $u_R'(r)w(r) < 0$ for all $r \in (0, \delta_0)$. Recalling that $u_R \in H_0^1(0, R)$ and that $w > 0$ on $(0, R)$, this yields

$$u_R(\delta_0) = \int_0^{\delta_0} u_R'(r) dr < 0,$$

contradicting the positivity of u_R on $(0, R)$.

Thus the initial assumption that the sequence $\{u_n\}$ is unbounded implies the existence of an element w whose properties lead to a contradiction. This establishes the boundedness of $\{u_n\}$. $\square$

We are now ready to prove the main result of this paper.

THEOREM 5.2. *Let the dielectric response satisfy the assumptions* (A) *and* (AS). *Let* $\omega \in J$ *be such that* $\Lambda_\infty < \Lambda_0$. *Then for every wave number* $k > 0$ *such that* $\Lambda_\infty < -k^2 < \Lambda_0$ *there is a positive solution* $u \in H\backslash\{0\}$ *of* (2.15). *Furthermore, the function* v *defined by* (2.12) *has the following properties:*

(i) $v \in C^2([0,\infty))$ *with* $v(0) = v''(0) = 0$ *and* $v > 0$ *on* $(0, \infty)$.

(ii) v *satisfies* (2.9) *and* (2.10).

Remark 5.1. Observe that the conditions (A) and (AS) are satisfied by a linear dielectric response,

$$\varepsilon(\omega, r, s) = \varepsilon_0(\omega, r) \text{ for all } r, s \geq 0,$$

provided that $\varepsilon_0(\omega, \cdot) \in C([0,\infty))$ and $0 < a(\omega) \leq \varepsilon_0(\omega, r) \leq b(\omega) < \infty$ for all $r, s \geq 0$. In this case, $\Lambda_\infty = \Lambda_0$. Thus the requirement that $\Lambda_\infty < \Lambda_0$ in Theorem 5.2 ensures that the dielectric response is in fact nonlinear.

Remark 5.2. In a homogeneous medium, $\varepsilon(\omega, r, s) = \varepsilon(\omega, 0, s)$ for all $r, s \geq 0$, and so

$$\Lambda_0 = \left(\frac{\omega}{c}\right)^2 \varepsilon(\omega, 0, 0) \text{ and } \Lambda_\infty = \lim_{s\to\infty} \left(\frac{\omega}{c}\right)^2 \varepsilon(\omega, 0, s).$$

Thus for a homogeneous self-focusing material, $\Lambda_\infty = \Lambda_0$ if and only if the dielectric response is linear.

Proof. We fix $k > 0$ such that $\Lambda_\infty < -k^2 < \Lambda_0$ and consider the sequence (4.4)–(4.7). By Lemma 5.1 this sequence is bounded in H and so we can suppose that $u_n \xrightarrow{n} u$ weakly in H.

Putting $\varphi = u$ in (3.6) we get

$$\begin{aligned}
\Phi_k'(u_n)u &= \int_0^\infty \left\{ u_n'(r) + \left[\frac{3}{4r^2} + k^2\right] u_n(r) \right\} u(r)dr - \int_0^\infty g(r, r^{-1/2}u_n(r)^+)u_n(r)u(r)dr \\
&= \int_0^\infty \left\{ u_n'(r) + \left[\frac{3}{4r^2} + k^2 - g_0(r)\right] u_n(r) \right\} u(r)dr - \int_0^\infty \widetilde{g}(r, r^{-1/2}u_n(r)^+)u_n(r)u(r)dr \\
&= \langle u_n, u\rangle_0 - \int_0^\infty \widetilde{g}(r, r^{-1/2}u_n(r)^+)u_n(r)u(r)dr.
\end{aligned}$$

Similarly,

$$\Phi_k'(u_n)u_n = \|u_n\|_0^2 - \int_0^\infty \widetilde{g}(r, r^{-1/2}u_n(r)^+)u_n(r)u_n(r)dr$$

and hence,

$$\begin{aligned}
\|u_n - u\|_0^2 &= \|u_n\|_0^2 - \langle u_n, u\rangle_0 - \langle u, u_n - u\rangle_0 \\
&= \Phi_k'(u_n)u_n + \int_0^\infty \widetilde{g}(r, r^{-1/2}u_n(r)^+)u_n(r)u_n(r)dr \\
&\qquad - \Phi_k'(u_n)u - \int_0^\infty \widetilde{g}(r, r^{-1/2}u_n(r)^+)u_n(r)u(r)dr - \langle u, u_n - u\rangle_0 \\
&= \Phi_k'(u_n)u_n - \Phi_k'(u_n)u - \langle u, u_n - u\rangle_0 + \int_0^\infty \widetilde{g}(r, r^{-1/2}u_n(r)^+)u_n(r)[u_n(r) - u(r)]dr.
\end{aligned}$$

From the properties of (4.4)–(4.7) and the weak convergence of $\{u_n\}$, we see that $\|u_n - u\|_0 \overset{n}{\to} 0$ provided that

$$\int_0^\infty \widetilde{g}(r, r^{-1/2}u_n(r)^+)u_n(r)[u_n(r) - u(r)]dr \xrightarrow{n} 0.$$

For any $R > 0$,

$$\begin{aligned}
&\left|\int_0^\infty \widetilde{g}(r, r^{-1/2}u_n(r)^+)u_n(r)[u_n(r) - u(r)]dr\right| \\
&\leq \int_0^R \left|\widetilde{g}(r, r^{-1/2}u_n(r)^+)u_n(r)[u_n(r) - u(r)]\right| dr \\
&\qquad\qquad + \int_R^\infty \left|\widetilde{g}(r, r^{-1/2}u_n(r)^+)u_n(r)[u_n(r) - u(r)]\right| dr \\
&\leq \int_0^R B(\omega)\left|u_n(r)[u_n(r) - u(r)]\right| dr + \int_R^\infty \widetilde{g}(r, r^{-1/2}\left|u_n\right|_\infty)\left|u_n(r)[u_n(r) - u(r)]\right| dr
\end{aligned}$$

by the monotonicity of $\widetilde{g}(r, \cdot)$. The boundedness of $\{u_n\}$ in H ensures that there is a constant $K > 0$ such that $|u_n|_2 \leq K$ and $|u_n|_\infty \leq K$ for all n. Hence

$$\begin{aligned}
&\int_R^\infty \widetilde{g}(r, r^{-1/2}\left|u_n\right|_\infty)\left|u_n(r)[u_n(r) - u(r)]\right| dr \\
&\qquad \leq \sup_{r \geq R} \widetilde{g}(r, r^{-1/2}K) \int_R^\infty \left|u_n(r)[u_n(r) - u(r)]\right| dr \\
&\qquad \leq \sup_{r \geq R} \widetilde{g}(r, r^{-1/2}K) 2K^2 \text{ for all } n.
\end{aligned}$$

Since $u_n \xrightarrow{n} u$ uniformly on $[0, R]$, we have that

$$\limsup_{n\to\infty} \left|\int_0^\infty \widetilde{g}(r, r^{-1/2}u_n(r)^+)u_n(r)[u_n(r) - u(r)]dr\right| \leq \sup_{r \geq R} \widetilde{g}(r, r^{-1/2}K)2K^2$$

for all R. But it follows from (A) that

$$\lim_{R\to\infty} \sup_{r \geq R} \widetilde{g}(r, r^{-1/2}K) = 0,$$

and consequently,

$$\lim_{n\to\infty} \int_0^\infty \widetilde{g}(r, r^{-1/2}u_n(r)^+)u_n(r)[u_n(r) - u(r)]dr = 0.$$

Hence $\|u_n - u\|_0 \xrightarrow{n} 0$ and Lemma 3.1, together with properties (4.5) and (4.7), imply that $\Phi_k(u) = c > 0$ and $\Phi_k'(u)\varphi = 0$ for all $\varphi \in H$. Thus $u \not\equiv 0$ and, choosing $\varphi = u^-$, we have that

$$0 = \Phi_k'(u)u^- = \left\|u^-\right\|_0^2.$$

Thus $u \geq 0$, and u satisfies (2.15). This implies that u' has a generalized derivative on $(0, \infty)$ and that (2.13) is satisfied. Thus $u \in C^2((0, \infty))$. Setting

$$h(r) = \frac{3}{4r^2} + k^2 - g\left(r, \frac{u(r)}{\sqrt{r}}\right) \text{ for } r > 0,$$

we have that $h \in C((0,\infty))$ and $u''(r) = h(r)u(r)$ for all $r > 0$. Since $u \geq 0$, but $u \not\equiv 0$ on $(0,\infty)$, we must have that $u(r) > 0$ for all $r > 0$.

Setting $v(r) = r^{-1/2}u(r)$ for $r > 0$, we have that $v \in C^2((0,\infty))$ and, by (2.20), $\lim_{r\to 0} v(r) = 0$. Setting $v(0) = 0$, this means that $v \in C([0,\infty))$. Furthermore, v satisfies (2.7), which can be written as

$$\left[v' + \frac{v}{r}\right]' = \{k^2 - g(r, v(r))\}v(r), \tag{5.9}$$

showing that

$$\lim_{r\to 0}\left[v' + \frac{v}{r}\right]' = 0$$

since $A(\omega) \leq g(r,s) \leq B(\omega)$ for all $r, s \geq 0$ and $v(0) = 0$. This implies that there exists $L \in \mathbb{R}$ such that $\lim_{r\to 0}[v' + \frac{v}{r}] = L$. But $v' + \frac{v}{r} = \frac{1}{r}(rv)'$ and, using L'Hospital's rule,

$$\lim_{r\to 0}\frac{v(r)}{r} = \lim_{r\to 0}\frac{rv(r)}{r^2} = \lim_{r\to 0}\frac{(rv)'}{2r} = \frac{1}{2}\lim_{r\to 0}\left[v' + \frac{v}{r}\right] = \frac{L}{2}.$$

This proves that v is differentiable at 0 with $v'(0) = \frac{L}{2}$. Also,

$$\begin{aligned}\lim_{r\to 0} v'(r) &= \lim_{r\to 0}\left\{v'(r) + \frac{v(r)}{r} - \frac{v(r)}{r}\right\}\\ &= \lim_{r\to 0}\left\{v' + \frac{v}{r}\right\} - \lim_{r\to 0}\frac{v(r)}{r} = L - \frac{L}{2}\\ &= \frac{L}{2}\end{aligned}$$

and we have shown that $v \in C^1([0,\infty))$.

Setting $w(r) = v(r)/r$, we find that (5.9) can be written as

$$[r^3w']' = r^3h(r)w, \text{ where } h(r) = \{k^2 - g(r, v(r))\} \text{ for } r > 0, \tag{5.10}$$

and hence,

$$r^3w'(r) - s^3w'(s) = \int_s^r t^3h(t)w(t)dt \text{ for } 0 < s < r.$$

Since $r^3w'(r) = r^2v'(r) - rv(r) \to 0$ as $r \to 0$, we have that

$$w'(r) = r^{-3}\int_0^r t^3h(t)w(t)dt \text{ for } r > 0,$$

where $h(t)w(t)$ remains bounded as $t \to 0$. Thus $\lim_{r\to 0} w'(r) = 0$. But then,

$$\begin{aligned}\lim_{r\to 0} v''(r) &= \lim_{r\to 0}\left\{\left[v' + \frac{v}{r}\right]' - w'\right\}\\ &= \lim_{r\to 0}\left[v' + \frac{v}{r}\right]' - \lim_{r\to 0} w' = 0,\end{aligned}$$

from which it follows that $v \in C^2([0,\infty))$ with $v''(0) = 0$.

Since $u \in H$ and $A(\omega) \le g(r,s) \le B(\omega)$ for all $r, s \ge 0$, it follows from (2.13) that $u \in H^2(1,\infty)$. This implies that $\lim_{r\to\infty} u(r) = \lim_{r\to\infty} u'(r) = 0$, and hence that

$$\lim_{r\to\infty} v(r) = \lim_{r\to\infty} v'(r) = 0.$$

Straightforward substitution shows that

$$\int_0^\infty v(r)^2\, r dr = \int_0^\infty u(r)^2\, dr \text{ and } \int_0^\infty v'(r)^2\, r dr = \int_0^\infty \left[u' - \frac{u}{2r}\right]^2 dr$$

and it follows from (2.21) that $u' - \frac{u}{2r} \in L^2(0,\infty)$ since $u \in H$. Thus v satisfies (2.9). This completes the proof. □

Finally, we show that the restrictions imposed on k in Theorem 5.2 are quite sharp.

THEOREM 5.3. *Let the dielectric response satisfy assumptions* (A) *and* (AS). *If the wave number k is such that $-k^2 \notin [\Lambda_\infty, \Lambda_0]$, then there is no nonnegative solution u of* (2.15) *in $H\backslash\{0\}$.*

Proof. Suppose that $u \in H\backslash\{0\}$ satisfies (2.15) and that $u \ge 0$ on $(0,\infty)$. As in the proof of Theorem 5.2, it follows that $u \in C^2((0,\infty))$ and $u(r) > 0$ for all $r > 0$.

Putting $\varphi = u$ in (2.15), we find that

$$\int_0^\infty |u'|^2 + \left[\frac{3}{4r^2} + k^2\right] u^2 dr = \int_0^\infty g\left(r, \frac{u(r)}{\sqrt{r}}\right) u^2 dr \le \int_0^\infty g_\infty(r) u^2 dr$$

since $g(r,\cdot)$ is nondecreasing for all $r > 0$. Hence

$$\int_0^\infty |u'|^2 + \left[\frac{3}{4r^2} - g_\infty(r)\right] u^2 dr = -k^2 \int_0^\infty u^2 dr.$$

But the definition of Λ_∞ implies that

$$\int_0^\infty |u'|^2 + \left[\frac{3}{4r^2} - g_\infty(r)\right] u^2 dr \ge \Lambda_\infty \int_0^\infty u^2 dr$$

so we must have $-k^2 \ge \Lambda_\infty$.

Suppose now that $-k^2 > \Lambda_0$. We can choose $\varepsilon > 0$ such that $-k^2 > \Lambda_0 + 2\varepsilon$ and then, by the definition of Λ_0, there exists $v \in H\backslash\{0\}$ such that

$$\int_0^\infty |v'|^2 + \left[\frac{3}{4r^2} - g_0(r)\right] v^2 dr \le [\Lambda_0 + \varepsilon] \int_0^\infty v^2 dr.$$

Using (2.21), it follows from the density of $C_0^\infty((0,\infty))$ in H that there exists $w \in C_0^\infty((0,\infty))\backslash\{0\}$ such that

$$\int_0^\infty |w'|^2 + \left[\frac{3}{4r^2} - g_0(r)\right] w^2 dr \le [\Lambda_0 + 2\varepsilon] \int_0^\infty w^2 dr.$$

For $R > 0$, we now set

$$m(R) = \inf\left\{\int_0^R |z'|^2 + \left[\frac{3}{4r^2} - g_0(r)\right] z^2 dr : z \in H_0^1((0,R)) \text{ and } \int_0^R z^2 dr = 1\right\}$$

and observe that, for any R such that $\operatorname{supp} w \subset (0,R)$, we must have $m(R) \le \Lambda_0 + 2\varepsilon$. For any $R \in (0,\infty)$, it follows from the compactness of the embedding of $H_0^1((0,R))$

in $L^2((0,R))$ and the weak sequential lower semicontinuity of $\int_0^R |z'|^2 + \frac{3}{4r^2}z^2 dr$ on $H_0^1((0,R))$ that there exists $u_R \in H_0^1((0,R))$ such that $\int_0^R u_R^2 dr = 1$ and

$$\int_0^R |u_R'|^2 + \left[\frac{3}{4r^2} - g_0(r)\right] u_R^2 dr = m(R).$$

We may assume that $u_R \geq 0$ on $(0,R)$ and that

$$\int_0^R u_R' z' + \left[\frac{3}{4r^2} - g_0(r)\right] u_R z dr = m(R) \int_0^R u_R z dr \text{ for all } z \in H_0^1((0,R)).$$

This implies that $u_R \in H^2((\delta, R))$ for any $\delta \in (0,R)$ with

$$-u_R'' + \left[\frac{3}{4r^2} - g_0(r)\right] u_R = m(R) u_R \text{ on } (0,R),$$

so we can suppose that $u_R \in C^1((0,R])$. Multiplying by u and integrating we get

$$-u_R'(R)u(R) + u_R'(\delta)u(\delta) + \int_\delta^R u_R' u' + \left[\frac{3}{4r^2} - g_0(r)\right] u_R u dr = m(R) \int_\delta^R u_R u dr$$

for any $\delta \in (0,R)$. Hence

$$\lim_{\delta \to 0} u_R'(\delta)u(\delta) = u_R'(R)u(R) - \int_0^R u_R' u' + \left[\frac{3}{4r^2} - g_0(r) - m(R)\right] u_R u dr$$

since $r^{-2}u_R(r)u(r)$ is integrable on $(0,R)$ by (2.21). On the other hand, if $\widetilde{u_R}$ is the function obtained by extending u_R by zero for $r \geq R$, we can set $\varphi = \widetilde{u_R}$ in (2.15), and this yields

$$\int_0^R u' u_R' + \left[\frac{3}{4r^2} + k^2 - g\left(r, \frac{u(r)}{\sqrt{r}}\right)\right] u_R u dr = 0.$$

Thus we have that

$$\begin{aligned}
\lim_{\delta \to 0} u_R'(\delta)u(\delta) &= u_R'(R)u(R) - \int_0^R \left[g\left(r, \frac{u(r)}{\sqrt{r}}\right) - k^2 - g_0(r) - m(R)\right] u_R u dr \\
&\leq u_R'(R)u(R) + [k^2 + m(R)] \int_0^R u_R u dr \\
&\leq u_R'(R)u(R) + [k^2 + \Lambda_0 + 2\varepsilon] \int_0^R u_R u dr
\end{aligned}$$

for $R > 0$ such that supp $w \subset (0,R)$. But $u > 0$ and $u_R \geq 0$ on $(0,R)$ with $u_R(R) = 0$, so $u_R'(R)u(R) \leq 0$. Also $\int_0^R u_R u dr > 0$ since $u_R \not\equiv 0$ on $(0,R)$. Hence

$$\lim_{\delta \to 0} u_R'(\delta)u(\delta) \leq [k^2 + \Lambda_0 + 2\varepsilon] \int_0^R u_R u dr < 0,$$

since $k^2 + \Lambda_0 + 2\varepsilon < 0$ by the choice of ε. But this means that there exists $\delta_0 > 0$ such that $u_R'(r)u(r) < 0$ for all $r \in (0, \delta_0)$. Recalling that $u_R \in H_0^1((0,R))$, we now have that

$$u_R(\delta_0) = \int_0^{\delta_0} u_R'(r) dr < 0,$$

contradicting the fact that $u_R \geq 0$ on $(0, R)$. Hence we must have $-k^2 \leq \Lambda_0$, completing the proof. $\square$

Remark 5.3. The proceeding result concerns positive solutions, but the first part of the proof shows that, in fact, (2.15) has no solutions at all in $H\backslash\{0\}$ for $-k^2 < \Lambda_\infty$. By strengthening slightly the assumptions on the dielectric response, arguments similar to those used for planar waveguides in Theorem 3.1 of [10] show that (2.15) may have no solutions in $H\backslash\{0\}$ if $-k^2 > \Lambda_0$. See also Theorems 4.2 and 4.4 in [11] for other nonexistence results of this kind used in a related problem concerning cylindrical waveguides.

REFERENCES

[1] D. G. Costa and H. Tehrani, *On a class of asymptotically linear elliptic problems in* $\mathbb{R}^N$, J. Differential Equations, 173 (2001), pp. 470–494.

[2] D. G. De Figueiredo, *The Ekeland Variational Principle with Applications and Detours*, Tata Institute Lecture Notes, Bombay, 1989.

[3] I. Ekeland, *Convexity Methods in Hamiltonian Mechanics*, Springer-Verlag, Berlin, 1990.

[4] L. Jeanjean, *On the existence of bounded Palais–Smale sequences and application to a Landesman–Lazer-type problem on* $\mathbb{R}^N$, Proc. Royal Soc. Edinburgh Sect. A, 129 (1999), pp. 787–809.

[5] L. Jeanjean and K. Tanaka, *A positive solution for an asymptotically linear elliptic problem on* $\mathbb{R}^N$ *autonomous at infinity*, ESAIM Control Optim. Calc. Var., 7 (2002), pp. 597–614.

[6] O. John and C. A. Stuart, *Guidance properties of a cylindrical defocusing waveguide*, Comment Math. Univ. Carolin., 35 (1994), pp. 653–673.

[7] H.-J. Ruppen, *Multiple TE-modes in a cylindric, self-focusing wave guide*, Nonlinear World, 2 (1995), pp. 387–418.

[8] G. I. Stegeman and M. Segev, *Optical spatial solitons and their interactions: Universality and diversity*, Science, 286 (1999), pp. 1518–1523.

[9] C. A. Stuart, *Self-trapping of an electromagnetic field and bifurcation from the essential spectrum*, Arch. Rational Mech. Anal., 113 (1991), pp. 65–96.

[10] C. A. Stuart, *Guidance properties of nonlinear planar waveguides*, Arch. Rational Mech. Anal., 125 (1993), pp. 145–200.

[11] C. A. Stuart, *Cylindrical TM-modes in a homogeneous self-focusing dielectric*, Math. Models Methods Appl. Sci., 6 (1996), pp. 977–1008.

[12] C. A. Stuart and H. S. Zhou, *A variational problem related to self-trapping of an electromagnetic field*, Math. Methods Applied Sci., 19 (1996), 1397–1407.

[13] C. A. Stuart and H. S. Zhou, *Applying the mountain pass theorem to an asymptotically linear elliptic equation on* $\mathbb{R}^N$, Comm. Partial Differential Equations, 24 (1999), pp. 1731–1758.

[14] M. M. Vainberg, *Variational Methods for the Study of Nonlinear Operators*, Holden-Day, 1964.

[15] H.-S. Zhou, *An application of a mountain pass theorem*, Acta Math. Sin. (English Ser.), 18 (2002), pp. 27–36.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 238–250

RECONSTRUCTION OF A POLYNOMIAL FROM ITS RADON PROJECTIONS*

BORISLAV BOJANOV† AND YUAN XU‡

Abstract. A polynomial of degree n in two variables is shown to be uniquely determined by its Radon projections taken over $[n/2]+1$ parallel lines in each of the $(2[(n+1)/2]+1)$ equidistant directions along the unit circle.

Key words. Radon projection, polynomials of two variables, interpolation

AMS subject classifications. 42A38, 42B08, 42B15

DOI. 10.1137/040616516

1. Introduction.

Let f be a function defined on the unit disk B^2 on the plane. A Radon projection of f is the integral of f over a line segment inside B^2. More precisely, for any given pair (θ, t) of a real number $t \in [-1,1]$ and any angle θ, let $I(\theta, t)$ denote the line segment inside the unit disk B^2, where the line passes through the point $(t\cos\theta, t\sin\theta)$ and is perpendicular to the vector $(\cos\theta, \sin\theta)$. Then

$$\mathcal{R}_\theta(f;t) := \int_{I(\theta,t)} f(x,y)dxdy \tag{1.1}$$

defines the Radon projection of f on the line segment $I(\theta, t)$.

The Radon transform $f \mapsto \{\mathcal{R}_\theta(f;t)\}$ associates with f a family of univariate functions of t parameterized by θ. The problem of reconstructing f from full or partial knowledge of $\{\mathcal{R}_\theta(f;t)\}$ has been studied by many authors. It plays an essential role in computer tomography. It is well known that the set of Radon projections $\{\mathcal{R}_\theta(f;t) : 0 \le t \le 1,\ 0 \le \theta \le 2\pi\}$ determines f completely. Furthermore, it is known that if f has compact support in B^2, then f is uniquely determined by any infinite set of Radon projections [16]. In practice, however, the data set is usually finite. Thus, the main problem is to obtain a good approximation to the function from a large collection of its Radon projections. See [13] for background and detailed discussions. Using a finite data set, one may determine a polynomial whose Radon projections agree with the given data. Such an approach has been considered in [10, 11, 12] in connection with computer tomography.

The present paper is concerned with the problem of whether a polynomial of degree n in two variables can be uniquely determined from a set of $(n+1)(n+2)/2$ distinct Radon projections. The solution depends on the arrangement of the lines on which the projections are taken. One early solution was given in [12], in which the Radon projections are taken over all possible line segments $[\mathbf{x}_i, \mathbf{x}_j]$ joining $\mathbf{x}_i$ and $\mathbf{x}_j$, where $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{n+1}$ are $n+2$ equally spaced points on the boundary of the unit disk

*Received by the editors October 6, 2004; accepted for publication (in revised form) January 18, 2005; published electronically September 20, 2005.
http://www.siam.org/journals/sima/37-1/61651.html

†Department of Mathematics, University of Sofia, Blvd. James Boucher 5, 1164 Sofia, Bulgaria (boris@fmi.uni-sofia.bg). This author was supported by the Bulgarian Ministry of Science and Education under project MM–1402/2004.

‡Department of Mathematics, University of Oregon, Eugene, OR 97403-1222 (yuan@math.uoregon.edu). This author was supported in part by the National Science Foundation under grant DMS-0201669.

B^2. In [7] (see also [4, 5]), a more general result was proved, in which $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{n+1}$ are $n+2$ distinct points on the boundary of a convex set, and the result was extended in [8] to higher dimensions. Recently the problem was considered in [1], in which the lines are taken in $(n+1)$ directions, with one line in the first direction, two lines in the second direction, etc. This set has the drawback of lacking symmetry.

The main result of the present paper shows that a set of Radon projections taken over $2\lfloor n/2 \rfloor + 1$ parallel lines on each of the $2\lfloor (n+1)/2 \rfloor + 1$ equidistant directions can determine a polynomial of degree n uniquely. The set of the corresponding line segments possesses a rotation symmetry. If $n = 2m$ is a fixed positive integer, then our method requires the Radon data in $2m+1$ directions, which are taken to be in equally spaced angles along the unit circle, and in each direction we take $m+1$ parallel lines associated with $t_0, t_1, \ldots, t_m$. We prove that for almost all choices of $\{t_k\}$ a polynomial P_n of degree $n = 2m$ in two variables is uniquely determined by the set of its Radon projections over these line segments; more precisely, for any given f, P_n is uniquely determined such that

$$\mathcal{R}_{\phi_j}(P_n; t_k) = \mathcal{R}_{\phi_j}(f; t_k), \quad 0 \le j \le 2m, \quad 0 \le k \le m. \tag{1.2}$$

A similar construction works for $n = 2m-1$. Several examples of the sets of points $\{t_k\}$ that define regular interpolation are given. Furthermore, the polynomial P_n can be easily computed. Thus, the result appears to offer a simple way to recover a polynomial from its Radon data.

The number of conditions (1.2) is equal to $(n+1)(n+2)/2$, which is the dimension of the space of bivariate polynomials of total degree at most n. Hence, (1.2) can be interpreted as an interpolation problem by polynomials based on line integrals. As in the case of interpolation based on points, this bivariate problem is not always regular. The only general configurations in the literature that hold for every integer n are that of Hakopian [8], mentioned above, and the nonsymmetric configuration given recently in [1]. The result of the present paper provides another family of examples. The proof is based on an observation that recovering polynomials from their Radon projections can be reduced, using a formula in [12], to a family of univariate interpolation problems that uses certain special classes of algebraic polynomials. In the case of [1], the strategy of solving the interpolation problem resembles the approach that uses the Bezout theorem for pointwise interpolation. The approach in the present paper follows the strategy of [2], which uses equally spaced points on circles for pointwise interpolation and allows one to step beyond the limitation of the Bezout theorem.

The paper is organized as follows. In the following section we prove the existence and uniqueness of the reconstructing problem. In section 3 we show how to construct the polynomial and we give an outline of the algorithm.

2. Existence and uniqueness of the solution.

2.1. Preliminary. Let $B^2 = \{(x,y) : x^2 + y^2 \le 1\}$ denote the unit disk on the plane. Let θ be the angle in the polar coordinates

$$x = r\cos\theta, \quad y = r\sin\theta, \qquad r \ge 0, \quad 0 \le \theta \le 2\pi.$$

A line ℓ whose slope is $-\cot\theta$ is defined by the equation

$$\ell(x,y) := x\cos\theta + y\sin\theta - t = 0,$$

where t is a real number. Clearly the line ℓ passes through the point $(t\cos\theta, t\sin\theta)$ and is perpendicular to the vector $\xi = (\cos\theta, \sin\theta)$. Since, for a fixed t, the line corresponding to (θ, t) coincides with the line corresponding to $(\pi + \theta, -t)$, we could assume $\theta \in [0, \pi)$. Alternatively, we could assume that $\theta \in [0, 2\pi)$ and $t \geq 0$.

We will also use ℓ to denote the set of points on the line ℓ and introduce the notation

$$I(\theta, t) = \ell \cap B^2, \qquad 0 \leq \theta < \pi, \quad -1 \leq t \leq 1$$

to denote the line segment of ℓ inside B^2. The points on $I(\theta, t)$ can be represented as

$$x = t\cos\theta - s\sin\theta, \quad y = t\sin\theta + s\cos\theta$$

for $s \in [-\sqrt{1-t^2}, \sqrt{1-t^2}]$.

The Radon projection of a function f in the direction θ with a parameter $t \in [-1, 1]$ is denoted by $\mathcal{R}_\theta(f; t)$,

$$\begin{aligned} \mathcal{R}_\theta(f; t) &:= \int_{I(\theta,t)} f(x, y)\,dxdy \\ &= \int_{-\sqrt{1-t^2}}^{\sqrt{1-t^2}} f(t\cos\theta - s\sin\theta, t\sin\theta + s\cos\theta)\,ds. \end{aligned}$$

In the literature it is also called an X-ray. It is clear that $\mathcal{R}_\theta(f; t) = \mathcal{R}_{\pi+\theta}(f; -t)$, so we can assume, for the sake of definiteness, either $0 \leq \theta < \pi$ or $0 \leq t \leq 1$.

2.2. Polynomial bases. Let Π_n^2 denote the space of polynomials of total degree n in two variables, which has dimension $\dim \Pi_n^2 = (n+1)(n+2)/2$. If $P \in \Pi_n^2$, then

$$P(\mathbf{x}) = \sum_{k=0}^{n} \sum_{j=0}^{k} c_{k,j} x^j y^{k-j}, \qquad \mathbf{x} = (x, y).$$

Let U_k denote the Chebyshev polynomial of the second kind [15],

$$U_k(x) = \frac{\sin(k+1)\theta}{\sin\theta}, \qquad x = \cos\theta.$$

For $\xi = (\cos\theta, \sin\theta)$ and $\mathbf{x} = (x, y)$, the ridge polynomial $U_k(\theta; \cdot)$ is defined by

$$U_k(\theta; \mathbf{x}) := U_k(\langle \mathbf{x}, \xi \rangle) = U_k(x\cos\theta + y\sin\theta).$$

Clearly, U_k is an element of Π_k^2 and it is constant on every line that is perpendicular to ξ. The Radon projection of this function in any direction can be easily computed. This is a result due to Marr [12] and it plays a central role in our discussion below.

LEMMA 2.1. *For each $t \in (-1, 1)$, $0 \leq \theta, \phi \leq 2\pi$,*

$$\mathcal{R}_\phi(U_k(\theta; \cdot); t) = \frac{2}{k+1}\sqrt{1-t^2}\,U_k(t)U_k(\cos(\phi - \theta)).$$

The following useful relation follows easily from Marr's formula [11]:

$$\frac{1}{\pi}\int_{B^2} U_k(\theta; \mathbf{x})U_k(\phi; \mathbf{x})\,d\mathbf{x} = \frac{1}{k+1}U_k(\cos(\phi - \theta)). \tag{2.1}$$

Let $\mathcal{V}_n$ denote the space of orthogonal polynomials of degree n on B^2 with respect to the unit weight function; that is, $P \in \mathcal{V}_n$ if P is of degree n and

$$\int_{B^2} P(\mathbf{x})Q(\mathbf{x})\,d\mathbf{x} = 0 \qquad \text{for all } Q \in \Pi_{n-1}^2.$$

For special choices of θ and ϕ, (2.1) becomes an orthogonal relation. Indeed, for $k \in \mathbb{N}$, let

$$\xi_{j,k} = (\cos\theta_{j,k}, \sin\theta_{j,k}), \qquad \theta_{j,k} := \frac{j\pi}{k+1}, \quad 0 \le j \le k.$$

For fixed k, the points $\cos\theta_{j,k}$, $1 \le j \le k$, are zeros of U_k. The ridge polynomials $U_k(\theta; \mathbf{x})$ have the remarkable orthogonal property that $U_k(\theta; \cdot) \in \mathcal{V}_k$. Since the dimension of $\mathcal{V}_k$ is $k+1$, it follows from (2.1) that the set

$$\mathbb{P}_k := \{U_k(\theta_{j,k}; \mathbf{x}) : 0 \le j \le k\}$$

is a basis for $\mathcal{V}_k$. In particular, this shows that the set $\{\mathbb{P}_k : 0 \le k \le n\}$ is a basis for Π_n^2. Together with Lemma 2.1, this proves the following result.

Lemma 2.2. *Every polynomial $P_n \in \Pi_n^2$ can be written uniquely as*

$$P_n(\mathbf{x}) = \sum_{k=0}^{n} \sum_{j=0}^{k} c_{j,k} U_k(\theta_{j,k}; \mathbf{x}). \tag{2.2}$$

Furthermore, for each ϕ and t,

$$\mathcal{R}_\phi(P_n; t) = \sqrt{1-t^2} \sum_{k=0}^{n} \frac{2}{k+1} U_k(t) \sum_{j=0}^{k} c_{j,k} U_k(\cos(\phi - \theta_{j,k})). \tag{2.3}$$

There are several other explicit orthogonal bases for $\mathcal{V}_n$; see, for example, [6]. If $\{Q_j^k : 0 \le k \le n\}$ is an orthogonal basis of $\mathcal{V}_n$, then, as was shown in [17],

$$U_k(\theta_{j,k}; \mathbf{x}) = \frac{1}{k+1} \sum_{l=0}^{k} Q_l^k(x,y) Q_l^k(\cos\theta_{j,k}, \sin\theta_{j,k}). \tag{2.4}$$

One explicit basis of $\mathcal{V}_n$, denoted by $Q_{j,i}(\mathbf{x})$, is given in terms of polar coordinates as (cf. [6])

$$\begin{aligned} Q_{j,1}(x,y) &= h_{n-2j,1} P_j^{(0,n-2j)}(2r^2-1) r^{n-2j} \cos(n-2j)\theta, \quad 0 \le 2j \le n, \\ Q_{j,2}(x,y) &= h_{n-2j,2} P_j^{(0,n-2j)}(2r^2-1) r^{n-2j} \sin(n-2j)\theta, \quad 0 \le 2j \le n-1, \end{aligned}$$

where $P_j^{(\alpha,\beta)}$ denotes the usual Jacobi polynomial, $h_{0,1} = 1/(n+1)$, and $h_{j,1} = h_{j,2} = 1/(2n+2)$ for $1 \le 2j \le n-1$. Using this basis and restricting $\mathbf{x}$ to the boundary of B^2, (2.4) becomes

$$\begin{aligned} U_{2k}(\cos(\phi - \theta_{j,2k})) &= 1 + 2\sum_{l=1}^{k} \left(\cos 2l\theta_{j,2k} \cos 2l\phi + \sin 2l\theta_{j,2k} \sin 2l\phi\right), \\ U_{2k-1}(\cos(\phi - \theta_{j,2k-1})) &= 2\sum_{l=1}^{k} \left(\cos(2l-1)\theta_{j,2k-1} \cos(2l-1)\phi \right. \\ &\qquad\qquad \left. + \sin(2l-1)\theta_{j,2k-1} \sin(2l-1)\phi\right). \end{aligned} \tag{2.5}$$

The above equations also follow from the elementary trigonometric identities

$$1 + 2\sum_{j=1}^{k} \cos 2j\theta = \frac{\sin(2k+1)\theta}{\sin\theta} \quad \text{and} \quad 2\sum_{j=1}^{k} \cos(2j-1)\theta = \frac{\sin 2k\theta}{\sin\theta}$$

with θ replaced by $\phi - \theta_{j,2k}$ and $\phi - \theta_{j,2k-1}$, respectively. These relations allow us to rewrite (2.3) in a form that is easier to work with. In the following we use $\lfloor x \rfloor$ to denote the integer part of x.

LEMMA 2.3. *Let P_n be given as in* (2.2). *Then*

$$\frac{\mathcal{R}_\phi(P_n;t)}{\sqrt{1-t^2}} = \sum_{l=0}^{\lfloor n/2 \rfloor} U_{2l}(t)\left[a_{0,2l} + 2\sum_{j=1}^{l}(a_{j,2l}\cos 2j\phi + b_{j,2l}\sin 2j\phi)\right] + \sum_{l=1}^{\lfloor (n+1)/2 \rfloor} U_{2l-1}(t)\left[2\sum_{j=1}^{l}(a_{j,2l-1}\cos(2j-1)\phi + b_{j,2l-1}\sin(2j-1)\phi)\right], \tag{2.6}$$

in which the coefficients $c_{j,k}$ in (2.2) *and $a_{j,k}$, $b_{j,k}$ are related by*

$$c_{j,2l} = \frac{1}{2}a_{0,2l} + \sum_{k=1}^{l}(a_{k,2l}\cos 2k\theta_{j,2l} + b_{k,2l}\sin 2k\theta_{j,2l}) \tag{2.7}$$

for $0 \le j \le 2l$, $0 \le l \le \lfloor n/2 \rfloor$, and

$$c_{j,2l-1} = \sum_{k=1}^{l}(a_{k,2l-1}\cos(2k-1)\theta_{j,2l-1} + b_{k,2l-1}\sin(2k-1)\theta_{j,2l-1}) \tag{2.8}$$

for $0 \le j \le 2l-1$, $1 \le l \le \lfloor (n+1)/2 \rfloor$.

Proof. We prove only the case of $n = 2m$. Comparing (2.3) and (2.6) shows that

$$\frac{1}{2l+1}\sum_{j=0}^{2l} c_{j,2l}U_{2l}(\cos(\phi - \theta_{j,2l})) = \frac{1}{2}a_{0,2l} + \sum_{j=1}^{l}(a_{j,2l}\cos 2j\phi + b_{j,2l}\sin 2j\phi),$$

$$\frac{1}{2l}\sum_{j=0}^{2l-1} c_{j,2l-1}U_{2l-1}(\cos(\phi - \theta_{j,2l-1})) = \sum_{j=1}^{l}(a_{j,2l-1}\cos(2j-1)\phi + b_{j,2l-1}\sin(2j-1)\phi).$$

Using (2.5) we can rewrite the left-hand side in the form of the right-hand side, which leads to

$$a_{j,2l} = \frac{2}{2l+1}\sum_{k=0}^{2l} c_{k,2l}\cos 2j\theta_{k,2l}, \qquad b_{j,2l} = \frac{2}{2l+1}\sum_{k=0}^{2l} c_{k,2l}\sin 2j\theta_{k,2l}$$

for $1 \le j \le l$, and $2a_{0,2l}$ satisfies the above formula with $j = 0$, and

$$a_{j,2l-1} = \frac{1}{l}\sum_{k=0}^{2l-1} c_{k,2l-1}\cos(2j-1)\theta_{k,2l-1},$$
$$b_{j,2l-1} = \frac{1}{l}\sum_{k=0}^{2l-1} c_{k,2l-1}\sin(2j-1)\theta_{k,2l-1}$$

for $1 \le j \le l$. For fixed l, the above linear relations can be written in the matrix form. After a proper normalization, the coefficient matrix turns out to be an orthogonal matrix. For example, for the even indices, we have

$$\frac{1}{\sqrt{2l+1}}\begin{bmatrix} 1 & 1 & \cdots & 1 \\ \sqrt{2} & \sqrt{2}\cos\frac{2\pi}{2l+1} & \cdots & \sqrt{2}\cos\frac{2l\pi}{2l+1} \\ 0 & \sqrt{2}\sin\frac{2\pi}{2l+1} & \cdots & \sqrt{2}\sin\frac{2l\pi}{2l+1} \\ \vdots & \vdots & \cdots & \vdots \\ \sqrt{2} & \sqrt{2}\cos 2l\frac{2\pi}{2l+1} & \cdots & \sqrt{2}\cos 2l\frac{2l\pi}{2l+1} \\ 0 & \sqrt{2}\sin 2l\frac{2\pi}{2l+1} & \cdots & \sqrt{2}\sin 2l\frac{2l\pi}{2l+1} \end{bmatrix} \cdot \begin{bmatrix} c_{0,2l} \\ c_{1,2l} \\ \vdots \\ c_{2l,2l} \end{bmatrix} = \frac{1}{\sqrt{2l+1}}\begin{bmatrix} a_{0,2l} \\ \frac{a_{1,2l}}{\sqrt{2}} \\ \frac{b_{1,2l}}{\sqrt{2}} \\ \vdots \\ \frac{a_{l,2l}}{\sqrt{2}} \\ \frac{b_{l,2l}}{\sqrt{2}} \end{bmatrix}.$$

The $(l+1)\times(l+1)$ matrix in the left-hand side, including the factor $1/\sqrt{2l+1}$, is orthogonal. Hence, the linear system of equations can be easily reversed, which leads to the stated relations, (2.7) and (2.8). The formulas can also be verified directly by inserting (2.7) and (2.8) into the equations of $a_{j,k}$ and $b_{j,k}$ and using the well-known trigonometric identities

$$\frac{1}{m+1}\sum_{k=0}^{m}\cos 2j\theta_{k,m} + i\frac{1}{m+1}\sum_{k=0}^{m}\sin 2j\theta_{k,m} = \frac{1}{m+1}\sum_{k=0}^{m} e^{2ij\theta_{k,m}} = \delta_{0,j}$$

for $0 \le j \le m$, where $\delta_{0,j} = 1$ if $j \ne 0$ and $\delta_{0,j} = 0$ otherwise, and using the elementary trigonometric identities $2\cos\theta\cos\phi = \cos(\theta+\phi)+\cos(\theta-\phi)$. ∎

To reconstruct the polynomial, we will determine the coefficients $a_{j,k}$ and $b_{j,k}$ and use them in (2.7) and (2.8) to determine $c_{j,k}$ in (2.2). The following representation is useful.

LEMMA 2.4. *Let $n = 2m$ or $n = 2m-1$. Let P_n be given as in* (2.2). *Then*

$$\frac{\mathcal{R}_\phi(P_n;t)}{\sqrt{1-t^2}} = A_0(t) + \sum_{j=1}^{n}\left[A_j(t)\cos j\phi + B_j(t)\sin j\phi\right], \tag{2.9}$$

where

$$A_0(t) = \sum_{l=0}^{\lfloor n/2\rfloor} a_{0,2l}U_{2l}(t),$$

and for $1 \le j \le m$,

$$A_{2j}(t) = 2\sum_{l=j}^{\lfloor n/2\rfloor} a_{j,2l}U_{2l}(t), \qquad B_{2j}(t) = 2\sum_{l=j}^{\lfloor n/2\rfloor} b_{j,2l}U_{2l}(t), \tag{2.10}$$

$$A_{2j-1}(t) = 2\sum_{l=j}^{\lfloor n/2\rfloor} a_{j,2l-1}U_{2l-1}(t), \qquad B_{2j-1}(t) = 2\sum_{l=j}^{\lfloor n/2\rfloor} b_{j,2l-1}U_{2l-1}(t). \tag{2.11}$$

Proof. The proof is obtained from changing the order of summations in (2.6). ∎

2.3. Existence and uniqueness of the solution. We now state the Radon projections from which the polynomial P_n can be uniquely recovered. These are given by

$$\mathcal{R}_{\phi_{j,m}}(P_n;t_k), \qquad 0 \le k \le \lfloor n/2\rfloor, \quad 0 \le j \le 2m, \quad m = \lfloor (n+1)/2\rfloor, \tag{2.12}$$

which has a total of $(\lfloor n/2 \rfloor + 1)(2\lfloor (n+1)/2 \rfloor + 1) = (n+1)(n+2)/2$ projections, which is the same as the dimension of Π_n^2. The angles are chosen to be equidistant,

$$\Theta_m := \left\{ \phi_{j,m} = \frac{2j\pi}{2m+1} : \quad 0 \le j \le 2m \right\}. \tag{2.13}$$

The reason that we choose the equidistant angles lies in the lemma below, which plays a key role in our study. Such a lemma appeared first in [2] and has been used in [3] and [18].

LEMMA 2.5. *For $\phi \in \Theta_m$ and $P_n \in \Pi_n^2$ with $n = 2m$ or $n = 2m-1$,*

$$\frac{\mathcal{R}_\phi(P_n;t)}{\sqrt{1-t^2}} = A_0(t)$$
$$+ \sum_{j=1}^{m} \left[\big(A_j(t) + A_{2m-j+1}(t)\big) \cos j\phi + \big(B_j(t) - B_{2m-j+1}(t)\big) \sin j\phi \right],$$

where we assume that $A_{2m} = B_{2m} = 0$ if $n = 2m-1$.

Proof. Using the expression in the previous lemma, the proof follows from the fact that

$$\cos(2m-j+1)\phi = \cos j\phi \quad \text{and} \quad \sin(2m-j+1)\phi = -\sin j\phi$$

for $\phi \in \Theta_m$. $\square$

For the expression in this lemma, the variable t is in $(-1,1)$. We can also state the result for $\phi \in \widetilde{\Theta}_m = \{(2j+1)\pi/(2m+1) : 0 \le j \le 2m\}$, for which the sign before A_{2m-j+1} and B_{2m-j+1} will be reversed. It is easy to see that $\widetilde{\Theta}_m = \Theta_m + \pi$ modulus 2π. The two expressions are consistent with the fact that $\mathcal{R}_\theta(f;t) = \mathcal{R}_{\pi+\theta}(f;-t)$.

We are now ready to prove our main result, which shows that the polynomial P_n can be uniquely determined by the data (2.12). We need the following function classes: For $n = 2m$ define

$$X_j(t) = \{U_{2m}(t), U_{2m-2}(t), \ldots, U_{2j}(t), U_{2m-1}(t), U_{2m-3}(t), \ldots, U_{2m-2j+1}(t)\}$$

for $1 \le j \le m-1$. We also consider X_j as a column vector in $\mathbb{R}^{m+1}$ and regard

$$\Xi_j(\mathbf{t}) = [X_j(t_0), X_j(t_1), \ldots, X_j(t_m)]$$

as an $(m+1) \times (m+1)$ matrix.

A set of points $\mathbf{t} = \{t_0, t_1, \ldots, t_m\}$ in $(-1,1)$ is said to be asymmetric if t and $-t$ do not both belong to $\mathbf{t}$.

THEOREM 2.1. *Let $\mathbf{t} := \{t_0, t_1, \ldots, t_m\}$ be a set of asymmetric distinct points in $(-1,1)$ such that the matrices $\Xi_1(\mathbf{t}), \Xi_2(\mathbf{t}), \ldots, \Xi_{m-1}(\mathbf{t})$ are all nonsingular; then the polynomial $P \in \Pi_{2m}^2$ is uniquely determined by the set (2.12) of its Radon projections.*

Proof. In order to prove that there is a unique solution, it is sufficient to show that if $\mathcal{R}_{\phi_{j,m}}(P;t_k) = 0$ for $0 \le j \le 2m$ and $0 \le k \le m$, then $P(\mathbf{x}) \equiv 0$.

Setting its left-hand side to zero, the expression in Lemma 2.5 shows that

$$A_0(t_k) + \sum_{j=1}^{m} \left[\big(A_j(t_k) + A_{2m-j+1}(t_k)\big) \cos j\phi + \big(B_j(t_k) - B_{2m-j+1}(t_k)\big) \sin j\phi \right] = 0$$

for $0 \le k \le m$ and $\phi \in \Theta_m$. The left-hand side is a trigonometric polynomial of degree m and it vanishes on $2m+1$ points. Hence, it follows from the uniqueness of the trigonometric interpolation that the coefficients have to be zero, that is,

$$
\begin{aligned}
(2.14)\qquad A_0(t_k) = 0, \qquad & A_j(t_k) + A_{2m-j+1}(t_k) = 0, \\
& B_j(t_k) - B_{2m-j+1}(t_k) = 0, \qquad 1 \le j \le m
\end{aligned}
$$

for $0 \le k \le m$. Since A_j and A_{2m-j+1} have different parities, the coefficients in $A_j(t) + A_{2m-j+1}(t)$ will not combine, and there are exactly $m+1$ terms in this polynomial, which are written as a linear combination of the Chebyshev polynomials $\{U_k\}$. Let $Y_0(t) = \{U_{2m}(t), U_{2m-2}(t), \dots, U_2(t), U_0(t)\}$ and let

$$
\begin{aligned}
Y_{2j}(t) &= \{U_{2m}(t), U_{2m-2}(t), \dots, U_{2j}(t), U_{2m-1}(t), U_{2m-3}(t), \dots, U_{2m-2j+1}(t)\}, \\
Y_{2j-1}(t) &= \{U_{2m-1}(t), U_{2m-3}(t), \dots, U_{2j+1}(t), U_{2m}(t), U_{2m-2}(t), \dots, U_{2m-2j}(t)\}.
\end{aligned}
$$

Furthermore, let Ξ_{Y_j} denote the matrix $\Xi_{Y_j} = [Y_j(t_0), Y_j(t_1), \dots, Y_j(t_m)]$. The coefficients of the linear systems of equations in (2.14) are Ξ_{Y_0}, $\Xi_{Y_{2j}}$ for $1 \le j \le m/2$, and $\Xi_{Y_{2j-1}}$ for $1 \le j \le (m+1)/2$. Hence, in order to prove that (2.14) implies $A_0(t) \equiv 0$, $A_j(t) = B_j(t) \equiv 0$ for $1 \le j \le m$, we need to show that these matrices are all invertible. However, it is easy to see that we have the relation $\Xi_{Y_{2(m-j)-1}} = \Xi_{Y_{2j}}$, which shows, in particular, that $\Xi_{Y_m} = \Xi_{Y_{m-1}}$. Thus, we can deal with $\Xi_{Y_{2j}}$ for $0 \le j \le m-1$. Furthermore, Y_0 contains only even polynomials U_{2l}, $0 \le l \le m$. Using the notation $s = t^2$, it follows that $U_{2l}(\sqrt{s})$ is a polynomial of degree l so that the matrix Ξ_{Y_0} is always invertible as $\mathbf{t}$ is a set of asymmetric distinct points. Thus, since $\Xi_{Y_{2j}} = \Xi_j(\mathbf{t})$, our assumption on $\mathbf{t}$ implies that all matrices are invertible. $\square$

A similar theorem holds in the case of $n = 2m-1$. First, we need to define for $n = 2m-1$,

$$
X_j(t) = \{U_{2m-2}(t), U_{2m-4}(t), \dots, U_{2j}(t), U_{2m-1}(t), U_{2m-3}(t), \dots, U_{2m-2j+1}(t)\}
$$

for $1 \le j \le m-1$ and

$$
X_m(t) = \{U_{2m-1}(t), U_{2m-3}(t), \dots, U_3(t), U_1(t)\}.
$$

We will keep the notion $\Xi_j(\mathbf{t})$ for the matrix built upon from the columns of X_j, which is an $m \times m$ matrix.

THEOREM 2.6. *Let $\mathbf{t} = \{t_0, t_1, \dots, t_{m-1}\}$ be a set of asymmetric distinct points in $(-1,0) \cup (0,1)$ such that the matrices $\Xi_1(\mathbf{t}), \Xi_2(\mathbf{t}), \dots, \Xi_m(\mathbf{t})$ are all nonsingular. Then the polynomial $P \in \Pi^2_{2m-1}$ is uniquely determined by the set* (2.12) *of its Radon projections.*

That the points $t_0, t_1, \dots, t_m$ are all in $(-1,0) \cup (0,1)$ means that none of the points can be zero. In fact, since U_{2k-1} is an odd polynomial, $\Xi_m(\mathbf{t})$ contains only odd polynomials, which vanish at the origin. Otherwise, the proof of this theorem is similar to that of the previous theorem. In the case of $n = 2m$, the set $X_0 = \{U_{2m}, U_{2m-2}, \dots, U_0\}$ does not appear in the conditions of the theorem since the interpolation at distinct points in $[0,1)$ by X_0 is always regular. Such a reduction of conditions does not appears to happen for $n = 2m-1$.

For each fixed j, the determinant of $\Xi_j(\mathbf{t})$ can be considered as a polynomial function in $\mathbf{t} = (t_0, t_1, \dots, t_m)$ so that $\det \Xi_j(\mathbf{t}) = 0$ defines a hypersurface of dimension m. Hence, the determinant is not zero for almost all choices of $\mathbf{t} \in \mathbb{R}^{m+1}$. Evidently, the same holds true for $m+1$ matrices. Hence, we have the following corollary.

COROLLARY 2.7. *For $n = 2m$ or $2m-1$, and for almost all choices of distinct points $t_0, t_1, \dots, t_m$ in $(-1,1)$ for $n = 2m$ or in $(-1,0) \cup (0,1)$ for $n = 2m-1$, the polynomial $P_n \in \Pi^2_n$ is uniquely determined by the set* (2.12) *of its Radon projections.*

2.4. Choices of . We have shown that almost all choices of $\mathbf{t} = \{t_0, t_1, \ldots, t_m\}$ will lead to the unique solution of recovering the polynomial. One may ask if *any* choice of $\mathbf{t} \subset (-1,1)^{m+1}$ will work. The answer, however, is negative, as the following example shows.

Example: $m = 2$. By Theorem 2.1 we have only to choose t_0, t_1, t_2 in $(0,1)$ such that the set $X_2 = \{U_2, U_3, U_4\}$ has nonsingular determinant $\Xi_2(\mathbf{t})$. Recall that

$$U_2(t) = 4t^2 - 1, \quad U_3(t) = 8t^3 - 4t, \quad U_4(t) = 16t^4 - 12t^2 + 1.$$

We can compute the determinant $\Xi_2(\mathbf{t})$ explicitly, and the result is

$$\det \Xi_2(\mathbf{t}) = 32 \prod_{1 \le i < j \le 3} (t_i - t_j)$$
$$\times \left[8t_1^2t_2^2t_3^2 + 4t_1t_2t_3(t_1+t_2+t_3) + (2t_1^2-1)(2t_2^2-1)(2t_3^2-1)\right].$$

Since $t_i \in (0,1)$, the first two terms in the square brackets are positive, while the third term could be negative. This shows, in particular, that $\det \Xi_2(\mathbf{t}) \ne 0$ if one of the t_i's is $\sqrt{2}/2$, or if two of t_1, t_2, t_3 are less than $\sqrt{2}/2$ and one is greater than $\sqrt{2}/2$. However, $\det \Xi_2(\mathbf{t})$ can be zero for some choices of t_0, t_1, t_2.

On the positive side, we give two results that provide sets of points that will ensure the uniqueness of the reconstruction. The first result uses the following theorem due to Obrechkoff [14].

Lemma 2.8. *Let $d\mu$ be a nonnegative weight function on an interval and let $P_0, P_1, P_2, \ldots$ be the orthogonal polynomials with respect to $d\mu$. Denote by α_n the largest zero of P_n. Then the number of zeros of the polynomial $a_0P_0 + a_1P_2 + \cdots + a_nP_n$, where $a_0, a_1, \ldots, a_n$ are real numbers, in the interval $(\alpha_n, +\infty)$ is at most the number of sign changes in the sequence of the coefficients $a_0, a_1, \ldots, a_n$.*

The Chebyshev polynomials $U_0, U_1, U_2, \ldots$ are orthogonal with respect to the weight function $\sqrt{1-x^2}$ on $[-1,1]$ so that the above lemma implies the following.

Proposition 2.9. *Let $\mathbf{t} = \{t_0, t_1, \ldots, t_m\}$ be a set of numbers that satisfies*

$$\cos\frac{\pi}{2m+1} < t_0 < t_1 < \cdots < t_m < 1. \tag{2.15}$$

Then the matrices $\Xi_1(\mathbf{t}), \ldots, \Xi_{m-1}(\mathbf{t})$ are all nonsingular.

Proof. If one of the matrices, say $\Xi_j(\mathbf{t})$, were singular, there would be a nonzero polynomial $P \in \operatorname{span} X_j$ that vanishes on $\mathbf{t}$. This means that the number of zeros of P in (α_n, ∞) would be $m+1$. However, each set Ξ_j has cardinality $m+1$ so that the number of sign changes in the sequence of the coefficients of P is at most m. This is a contradiction to the conclusion of the lemma. ∎

The set of points given in this proposition ensures the uniqueness of determining a polynomial by the set of its Radon projections. Condition (2.15) implies that all points are clustered toward one end of the interval $[-1,1]$. This appears to be neither a practical nor a good choice for computation. In fact, the numerical test indicates that some of the matrices Ξ_j tend to have very large condition numbers.

Our second positive result is more interesting. Here the points $\{t_k\}$ are based on the zeros of the Chebyshev polynomials U_{2m}. It is well known that these zeros are given by

$$\eta_{j,2m} := \cos\frac{j\pi}{2m+1}, \qquad j = 1, 2, \ldots, 2m.$$

We state the result for $n = 2m$ first.

THEOREM 2.10. *Let t_0 be any point in $(-1, 1)$ such that $U_{2m}(t_0) \neq 0$. Let $t_j = \eta_{2j,2m}$ for $j = 1, 2, \ldots, m$. Then the matrices $\Xi_1(\mathbf{t}), \Xi_2(\mathbf{t}), \ldots, \Xi_{m-1}(\mathbf{t})$ in Theorem 2.1 are all nonsingular. Consequently, the polynomial $P \in \Pi^2_{2m}$ is uniquely determined by the set (2.12) of its Radon projection.*

Proof. It is easy to see that the set $\mathbf{t}$ is asymmetric. Hence, according to Theorem 2.1, we need only show that $\Xi_j(\mathbf{t})$ is invertible for each $j = 1, 2, \ldots, m-1$. Using the explicit expression of $U_{2m}(t)$, it is easy to see that

$$U_{2m-2j}(t_k) = U_{2m-2j}(\eta_{2k,2m}) = -U_{2j-1}(\eta_{2k,2m}) = -U_{2j-1}(t_k). \tag{2.16}$$

Indeed, since $\sin(2m+1)\eta_{2k,2m} = 0$ and $\cos(2m+1)\eta_{2k,2m} = 1$, it follows from the addition formula that $\sin(2m-2j+1)\eta_{2k,2m} = -\sin 2j\eta_{2k,2m}$, from which the equation follows.

Since $U_{2m}(t_1) = \cdots = U_{2m}(t_m) = 0$, we evidently have

$$\det \Xi_j(\mathbf{t}) = U_{2m}(t_0) \det \widetilde{\Xi}_j(\mathbf{t}),$$

where $\widetilde{\Xi}_j(\mathbf{t})$ is a submatrix of $\Xi_j(\mathbf{t})$ with the column $X_j(t_0)$ and the row containing $U_{2m}(t_j)$ of $\Xi_j(\mathbf{t})$ removed. Using (2.16), we can replace all rows of $\widetilde{\Xi}_j(\mathbf{t})$ that contain even Chebyshev polynomials with rows that contain odd Chebyshev polynomials. More precisely, we replace the rows $(U_{2i}(t_1), \ldots, U_{2i}(t_m))$ for $i = j, j+1, \ldots, m-1$ with rows $(U_{2m-2i-1}(t_1), \ldots, U_{2m-2i-1}(t_m))$, respectively. The new matrix has all elements given in terms of the Chebyshev polynomials of the odd degree; hence,

$$\det \widehat{\Xi}_j(\mathbf{t}) = \varepsilon \det(U_{2i-1}(t_k))_{i=1,k=1}^{m,\ m},$$

where $\varepsilon = \pm 1$. Assume now that this determinant is zero. Then there exists a nonzero polynomial Q of the form

$$Q(x) = \sum_{i=1}^{m} b_i U_{2i-1}(x),$$

which vanishes at $t_1, \ldots, t_m$. Since Q is odd, it vanishes also at $-t_1, \ldots, -t_m$. Observe that the sets $(-t_1, \ldots, -t_m)$ and $(t_1, \ldots, t_m)$ do not overlap. We conclude that the polynomial Q of degree $2m-1$ vanishes at $2m$ distinct points and, consequently, it vanishes identically, which is a contradiction. □

A similar theorem holds for $n = 2m-1$, which we state below.

THEOREM 2.11. *Let $t_j = \eta_{2j,2m}$ for $j = 1, 2, \ldots, m$. Then the matrices $\Xi_1(\mathbf{t})$, $\Xi_2(\mathbf{t}), \ldots, \Xi_m(\mathbf{t})$ in Theorem 2.6 are all nonsingular. Consequently, the polynomial $P \in \Pi^2_{2m-1}$ is uniquely determined by the set (2.12) of its Radon projection.*

The proof is similar to that of Theorem 2.10 and uses

$$U_{2m-2j}(\eta_{2k-1,2m}) = U_{2j-1}(\eta_{2k-1,2m}),$$

which can be verified using elementary trigonometric identities.

We have conducted some numerical tests for identifying other sets of t_k, for which the matrices $\Xi_1(\mathbf{t}), \ldots, \Xi_{m-1}(\mathbf{t})$ are nonsingular, so that the reconstruction of a polynomial from the set (2.12) of its Radon projections is unique. For $n \leq 20$, it turns out that both the equidistant points in $(0, 1)$ and the Chebyshev points in $(0, 1)$ work out. See, however, the discussion at the end of the next section.

3. Reconstruction of polynomials. In this section we set up the algorithm that can be used to compute P_n from the Radon projections. We consider only the case $n = 2m$. Let $\gamma_{j,k}$ be the data

$$\gamma_{j,k} = \mathcal{R}_{\phi_j}(f; t_k) \Big/ \sqrt{1 - t_k^2}, \qquad 0 \le j \le 2m, \quad 0 \le k \le m, \tag{3.1}$$

where $\phi_{j,m}$ are given as in (2.13) and t_k are distinct numbers in $[0, 1)$, chosen in advance, such that the linear systems of equations (3.2) and (3.3) below have unique solutions for all j.

Recall that the Lagrange interpolation by trigonometric polynomials is used in the proof of Theorem 2.1. The Lagrange interpolation based on the points in (2.13) is given explicitly by (see, for example, [19])

$$L_n f(\phi) = \sum_{j=0}^{2m} f(\phi_{j,m}) \ell_j(\phi), \qquad \ell_j(\phi) = \frac{\sin(m + \frac{1}{2})(\phi - \phi_{j,m})}{(2m + 1) \sin \frac{1}{2}(\phi - \phi_{j,m})},$$

where we assume that f is the function being interpolated; that is, $L_n f(\phi_{j,m}) = f(\phi_{j,m})$ for $0 \le j \le 2m$. Using the well-known formula

$$1 + 2\cos\phi + \cdots + 2\cos m\phi = \frac{\sin(m + \frac{1}{2})\phi}{\sin \frac{\phi}{2}},$$

we can write $L_n f$ in the standard form of a trigonometric polynomial,

$$L_n f(\phi) = m_0(f) + 2 \sum_{j=1} \left(m_j^C(f) \cos j\phi + m_j^S(f) \sin j\phi \right),$$

where

$$m_j^C(f) = \frac{1}{2m + 1} \sum_{l=0}^{2m} f(\phi_{l,m}) \cos j\phi_{l,m},$$
$$m_j^S(f) = \frac{1}{2m + 1} \sum_{l=0}^{2m} f(\phi_{l,m}) \sin j\phi_{l,m}.$$

For each k, $0 \le k \le m$, we will use the above formulas with $f(\phi_{j,m}) = \gamma_{j,k}$ and write $m_{j,k}^C = m_j^C(f)$ and $m_{j,k}^S = m_j^S(f)$ for this particular f. This will allow us to determine the values $A_0(t_k)$, $A_j(t_k) + A_{2m-j+1}(t_k)$, and $B_j(t_k) - B_{2m-j+1}(t_k)$ for $1 \le j \le m$.

The next step is to fix j and use the values of $A_0(t_k)$, or $A_j(t_k) + A_{2m-j+1}(t_k)$, or $B_j(t_k) - B_{2m-j+1}(t_k)$ for $k = 0, 1, \ldots, m$ to determine the coefficients of A_j and B_j. For this purpose we consider the following linear systems of equations: For $1 \le j \le m/2$,

$$\sum_{l=j}^{m} d_{j,2l} U_{2l}(t_k) + \sum_{l=m-j+1}^{m} d_{m-j+1,2l} U_{2l-1}(t_k) = m_{2j,k}, \quad 0 \le k \le m, \tag{3.2}$$

and for $1 \le j \le (m + 1)/2$,

$$\sum_{l=j}^{m} d_{j,2l-1} U_{2l-1}(t_k) + \sum_{l=m-j+1}^{m} d_{m-j+1,2l} U_{2l}(t_k) = m_{2j-1,k}, \quad 0 \le k \le m. \tag{3.3}$$

The quantities $m_{j,k}$ will be specified later. It is easy to see that the coefficient matrix of these systems of equations is Ξ_j as in Theorem 2.1. There are a total of m systems of linear equations, each of size $(m+1)\times(m+1)$. Solving these equations with proper $m_{j,k}$ will determine the coefficients of A_j and B_j.

The last step is to use (2.7) and (2.8) to get $c_{j,k}$, which are the coefficients of P_n in (2.2).

ALGORITHM.

Step 1. For $1 \le l \le m$ and $0 \le k \le m$, compute

$$m^C_{l,k} = \frac{1}{2m+1}\sum_{j=0}^{2m}\gamma_{j,k}\cos l\theta_j \qquad \text{and} \qquad m^S_{l,k} = \frac{1}{2m+1}\sum_{j=0}^{2m}\gamma_{j,k}\sin l\theta_j. \tag{3.4}$$

Step 2. Solve the systems of equations (3.2) and (3.3) to get the coefficients $a_{j,k}$ and $b_{k,j}$ in (2.6):

Case 1. Solve the m systems of (3.2) and (3.3) for $m_{j,k} = m^C_{j,k}$ to get

$$a_{j,k} := d_{j,k}, \qquad 1 \le j \le m, \quad 0 \le k \le 2m.$$

Case 2. Solve the m systems of (3.2) and (3.3) for $m_{j,k} = m^S_{j,k}$ to get

$$b_{j,2l} := d_{j,2l}, \quad 1 \le j \le m/2 \quad \text{and} \quad b_{j,2l} := -d_{j,2l}, \quad \frac{m+2}{2} \le j \le m,$$
$$b_{j,2l-1} := d_{j,2l-1}, \quad 1 \le j \le \frac{m+1}{2} \quad \text{and} \quad b_{j,2l-1} := -d_{j,2l-1}, \quad \frac{m+1}{2} \le j \le m,$$

where $1 \le l \le m$.

Step 3. Substitute the outputs $a_{j,k}$ and $b_{j,k}$ from Step 2 into (2.7) and (2.8) to get $c_{j,k}$. Then the polynomial P is given by (2.2).

The output of this algorithm is the polynomial P that satisfies (1.2). The main computation appears to be Step 2. Initial numerical experiments indicate that the linear systems of equations (3.2) and (3.3) are ill-conditioned. However, among the set of points that we tested, the largest condition number of the matrices corresponding to the equidistant points $t_k = (k+1)/(m+2)$, $0 \le k \le m$ of $(0,1)$ is smaller than that corresponding to the Chebyshev points $t_k = \cos(k+1)\pi/(2m+4)$, $0 \le k \le m$, or the points in Theorem 2.10.

As pointed out by a referee, it is perhaps not surprising that the matrices of the linear systems (3.2) and (3.3) are ill-conditioned. Solving these linear systems means inverting the Radon transform while it is known that the Radon reconstruction problem is ill posed. Thus, for m large, the algorithm may not be useful for practical computation. On the other hand, our result shows that it is possible to determine a polynomial of degree n uniquely from a set of $(n+1)(n+2)/2$ Radon data that concurs with the parallel geometry. This appears to be of independent interest.

REFERENCES

[1] B. Bojanov and I. K. Georgieva, *Interpolation by bivariate polynomials based on Radon projections*, Studia Math., 162 (2004), pp. 141–160.

[2] B. Bojanov and Y. Xu, *On a Hermite interpolation by polynomials of two variables*, SIAM J. Numer. Anal., 39 (2002), pp. 1780–1793.

[3] B. Bojanov and Y. Xu, *On polynomial interpolation of two variables*, J. Approx. Theory, 120 (2003), pp. 267–282.

[4] A. S. Cavaretta, C. A. Micchelli, and A. Sharma, *Multivariate interpolation and the Radon transform, Part* I, Math. Z., 174 (1980), pp. 263–279.
[5] A. S. Cavaretta, C. A. Micchelli, and A. Sharma, *Multivariate interpolation.* II. *Some further examples*, in Quantitive Approximation, R. A. DeVore and K. Scherer, eds., Acad. Press, New York, London, 1980, pp. 49–62.
[6] C. F. Dunkl and Y. Xu, *Orthogonal Polynomials of Several Variables*, Cambridge University Press, Cambridge, UK, 2001.
[7] H. Hakopian, *Multivariate divided differences and multivariate interpolation of Lagrange and Hermite type*, J. Approx. Theory, 34 (1982), pp. 286–305.
[8] H. Hakopian, *Multivariate spline functions, B-spline basis and polynomial interpolation* II, Studia Math., 79 (1984), pp. 91–102.
[9] H. Hakopian and S. Ismaeil, *On a bivariate interpolation problem*, J. Approx. Theory, 116 (2002), pp. 76–99.
[10] C. Hamaker and D. C. Solmon, *The angles between the null spaces of X-rays*, J. Math. Anal. Appl., 62 (1978), pp. 1–23.
[11] B. Logan and L. Shepp, *Optimal reconstruction of a function from its projections*, Duke Math. J., 42 (1975), pp. 645–659.
[12] R. Marr, *On the reconstruction of a function on a circular domain from a sampling of its line integrals*, J. Math. Anal. Appl., 45 (1974), pp. 357–374.
[13] F. Natterer, *The Mathematics of Computerized Tomography*, Classics in Appl. Math. 32, SIAM, Philadelphia, 2001.
[14] N. Obrechkoff, *Zeros of Polynomials*, Bulgarian Academic Monographs 7, Marin Drinov Academic Publishing House, Sofia, 2003.
[15] T. J. Rivlin, *Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory*, 2nd ed., Wiley, New York, 1990.
[16] D. C. Solmon, *The X-ray transform*, J. Math. Anal. Appl., 56 (1976), pp. 61–83.
[17] Y. Xu, *Funk-Hecke formula for orthogonal polynomials on spheres and on balls*, Bull. London Math. Soc., 32 (2000), pp. 447–457.
[18] Y. Xu, *Polynomial interpolation on the unit sphere*, SIAM J. Numer. Anal., 41 (2003), pp. 751–766.
[19] A. Zygmund, *Trigonometric Series*, Cambridge University Press, Cambridge, UK, 1959.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 251–275

GLOBAL ATTRACTORS AND STEADY STATES FOR UNIFORMLY PERSISTENT DYNAMICAL SYSTEMS*

PIERRE MAGAL† AND XIAO-QIANG ZHAO‡

Abstract. By appealing to the theory of global attractors on complete metric spaces, we obtain weaker sufficient conditions for the existence of interior global attractors for uniformly persistent dynamical systems, and hence generalize the earlier results on coexistence steady states. We also provide examples to show applicability of our interior fixed point theorem in the case of convex κ-contracting maps, and to prove the existence of discrete- and continuous-time dynamical systems that admit global attractors, but no strong global attractors, which gives an affirmative answer to an open question presented by Sell and You [*Dynamics of Evolutionary Equations*, Springer-Verlag, New York, 2002] in the case of continuous-time semiflows.

Key words. uniform persistence, global attractors, steady states

AMS subject classifications. 37C25, 37C70, 37L05, 37N25

DOI. 10.1137/S0036141003439173

1. Introduction.

Uniform persistence is an important concept in population dynamics since it characterizes the long-term survival of some or all interacting species in an ecosystem. There have been extensive investigations on uniform persistence for discrete- and continuous-time dynamical systems. We refer to [13, 27, 30] for surveys and reviews. Looked at abstractly, uniform persistence is the notion that a closed subset of the state space (e.g., the set of extinction for one or more populations) is repelling for the dynamics on the complementary set. A natural question concerns the existence of "interior" global attractors and "coexistence" steady states for uniformly persistent dynamical systems. The existence of interior global attractors was addressed by Hale and Waltman [10], and the existence of coexistence steady states under a general setting was investigated by Zhao [29]. In [10, 29] the traditional concept of global attractors was employed: a global attractor is a compact, invariant set which attracts every bounded set in the phase space (see, e.g., Hale [7], Temam [24], and Raugel [20]).

Recently, the following weaker concept of global attractors was introduced by Hirsch, Smith, and Zhao [11] and Sell and You [22]: a global attractor is a compact, invariant set which attracts some neighborhood of itself and every point in the phase space. For convenience, we refer to a traditional global attractor as a strong global attractor. With the concept of strong global attractor, Zhao [29, Theorem 2.3] assumed more conditions than necessary for the existence of a coexistence fixed point. However, the proof of [29, Theorem 2.3] needs only the property that the interior attractor attracts every compact set, and hence actually implies a general fixed point theorem that, if a continuous and κ-condensing map T has an interior global attractor, then it has a coexistence fixed point (see Theorem 4.1). So an important problem is to obtain sufficient conditions for the existence of interior global attractors for uniformly

*Received by the editors December 29, 2003; accepted for publication (in revised form) November 17, 2004; published electronically September 20, 2005.
http://www.siam.org/journals/sima/37-1/43917.html

†Faculte des Sciences et Techniques B.P. 540, 76058 Le Havre, France (magal.pierre@wanadoo.fr).
‡Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NF A1C 5S7, Canada (xzhao@math.mun.ca). The research of this author was supported in part by the NSERC of Canada.

persistent dynamical systems. This is a nontrivial problem since the phase space M_0 is an open subset of a complete metric space (M, d).

The main purpose of this paper is to establish the existence of the interior global attractor (i.e., the global attractor for $T : (M_0, d) \to (M_0, d)$) and a fixed point in M_0.

There is the following open question on the weaker concept of global attractor (see pages 55 and 56 of [22]): Does there exist an example of a κ-contracting semiflow that is point dissipative on a complete metric space W in which the global attractor does not attract every bounded set in W? In other words, we expect to find a dynamical system which has a global attractor, but no strong global attractor. In the case of discrete-time semiflows (i.e., maps), such a question has already been answered positively by Cholewa and Hale [4] (see also Raugel [20]) who developed an original result of Cooperman [5] and introduced an appropriate κ-contraction map on the Hilbert space of square summable series. As a byproduct of our investigations on interior global attractors, we will provide examples of both discrete- and continuous-time semiflows to give an affirmative answer to Sell and You's question (see sections 5.2 and 5.3). It is worth pointing out that our continuous-time dynamical systems are solution semiflows associated with a class of evolutionary equations with age structure.

It is obvious that we should start with the development of the theory of strong global attractors into that of global attractors on complete metric spaces. Note that the metric space (M_0, d) is not complete since M_0 is an open subset of the complete metric space (M, d). In order to apply the theory of global attractors to $T : (M_0, d) \to (M_0, d)$, we introduce a new metric $d_0(x, y)$ on M_0 (see equation (2) for its definition) so that (M_0, d_0) is a complete metric space. It turns out that the strongly bounded sets introduced in [10] correspond to the bounded sets in (M_0, d_0). This metric function $d_0(x, y)$ is the key tool for both the existence of a global attractor for $T : (M_0, d) \to (M_0, d)$ and four counterexamples of dynamical systems on the complete metric space (M_0, d_0). The theory of global attractors was already developed for continuous-time semiflows in the book [22], where the concept of κ-contracting maps was introduced (i.e., for each bounded set $B \subset M$, $\kappa(T^n(B)) \to 0$ as $n \to +\infty$). It seems that this strong notion may not be applied to $T : (M_0, d_0) \to (M_0, d_0)$. In fact, if $T : (M_0, d_0) \to (M_0, d_0)$ is κ-contracting, then $T^n(B)$ is strongly bounded for all sufficiently large integers n whenever B is a strongly bounded subset of M_0. But this property may not be satisfied in general in the applications (see the first example in section 5.1). So we will use the concept of asymptotically smooth maps introduced in [7] to establish the existence of global attractors.

Using our established theory of global attractor in M_0, we further investigate the existence of a fixed point of T in M_0. We also generalize the aforementioned coexistence fixed point theorem for κ-condensing maps to convex κ-contracting maps (see Definition 4.3), a new concept motivated by the fixed point theorem of Hale and Lopes [9] and the Poincaré maps associated with periodic age-structured population models. In terms of uniform persistence, our fixed point theorem (see Theorem 4.5) and its corollary (see Corollary 4.6) generalize earlier results due to Browder [2], Nussbaum [18, 19], Zhao [29], and Magal and Arino [15]. Clearly, there are analogues of interior global attractors and fixed point results for continuous-time semiflows (see Remark 3.10 and Theorem 4.7).

This paper is organized as follows. In section 2, we recall some basic concepts and results for dissipative dynamical systems based on the book of Hale [7] and establish sufficient conditions for the existence of global attractors and strong global attractors. In section 3, we prove the existence of a global attractor for $T : (M_0, d) \to (M_0, d)$. In section 4, we present the fixed point theorems and their corollaries. In section 5, we provide four examples to show the existence of discrete- and continuous-time dynamical systems that admit global attractors, but no strong global attractors. A simple periodic age-structured model is also studied in this section to illustrate applicability of Theorem 4.5 in the case of convex κ-contracting maps.

2. Preliminaries. Let (M, d) be a complete metric space. Recall that a set U in M is said to be a neighborhood of another set V provided V is in the interior $\mathrm{int}(U)$ of U. For any subsets A, $B \subset M$ and any $\epsilon > 0$, we define

$$d(x, A) := \inf_{y \in A} d(x, y), \quad \delta(B, A) := \sup_{x \in B} d(x, A),$$

$$N(A, \epsilon) := \{x \in M : d(x, A) < \epsilon\} \text{ and } \overline{N}(A, \epsilon) := \{x \in M : d(x, A) \leq \epsilon\}.$$

The *Kuratowski measure of noncompactness*, κ, is defined by

$$\kappa(B) = \inf\{r : B \text{ has a finite open cover of diameter } \leq r\}$$

for any bounded set B of M. We set $\kappa(B) = +\infty$ whenever B is unbounded.

For various properties of Kuratowski's measure of noncompactness, we refer to [17, 6] and [22, Lemma 22.2]. The proof of the following lemma is straightforward.

LEMMA 2.1. *The following statements are valid:*

(a) *Let $I \subset [0, +\infty)$ be unbounded, and let $\{A_t\}_{t \in I}$ be a nonincreasing family of nonempty closed subsets (i.e., $t \leq s$ implies $A_s \subset A_t$). Assume that $\kappa(A_t) \to 0$, as $t \to +\infty$. Then $A_\infty = \cap_{t \geq 0} A_t$ is nonempty and compact, and $\delta(A_t, A_\infty) \to 0$, as $t \to +\infty$.*

(b) *For each $A \subset M$ and $B \subset M$, we have $\kappa(B) \leq \kappa(A) + \delta(B, A)$.*

Let $T : M \to M$ be a continuous map. We consider the discrete-time dynamical system $T^n : M \to M$ $\forall n \geq 0$, where $T^0 = Id$ and $T^n = T \circ T^{n-1}$ $\forall n \geq 1$. We denote for each subset $B \subset M$, $\gamma^+(B) = \cup_{m \geq 0} T^m(B)$ the *positive orbit of B for T*, and denote

$$\omega(B) = \bigcap_{n \geq 0} \overline{\bigcup_{m \geq n} T^m(B)}$$

the *omega-limit set of B*. A subset $A \subset M$ is *positively invariant* for T if $T(A) \subset A$. A is *invariant* for T if $T(A) = A$. We say that a subset $A \subset M$ *attracts* a subset $B \subset M$ for T if $\lim_{n \to \infty} \delta(T^n(B), A) = 0$.

It is easy to see that B is precompact (i.e., $\overline{B}$ is compact) if and only if $\kappa(B) = 0$. A continuous mapping $T : X \to X$ is said to be *compact* (*completely continuous*) if T maps any bounded set to a precompact set in M.

The theory of attractors is based on the following fundamental result, which is related to [7, Lemmas 2.1.1 and 2.1.2].

LEMMA 2.2. *Let B be a subset of M, and assume that there exists a compact subset $C \subset M$, which attracts B for T. Then $\omega(B)$ is nonempty, compact, invariant for T, and attracts B.*

Proof. Let $I = N$, the set of all nonnegative integers, and

$$A_n = \overline{\bigcup_{m \geq n} T^m(B)} \ \forall n \geq 0.$$

Since C attracts B, from Lemma 2.1(b) we deduce that

$$\kappa(A_n) \leq \kappa(C) + \delta(A_n, C) = \delta(A_n, C) \to 0 \text{ as } n \to +\infty.$$

So the family $\{A_n\}_{n\geq 0}$ satisfies the conditions of assertion (a) in Lemma 2.1, and we deduce that $\omega(B)$ is nonempty, compact, and $\delta(A_n, \omega(B)) \to 0$, as $n \to +\infty$. So $\omega(B)$ attracts B for T. Moreover, we have

$$T\left(\bigcup_{m\geq n} T^m(B)\right) = \bigcup_{m\geq n+1} T^m(B) \; \forall n \geq 0,$$

and since T is continuous, we obtain

$$T(A_n) \subset A_{n+1}, \text{ and } A_{n+1} \subset \overline{T(A_n)} \; \forall n \geq 0.$$

Finally, since $\delta(A_n, \omega(B)) \to 0$, as $n \to +\infty$, we have $T(\omega(B)) = \omega(B)$. □

Definition 2.3. *A continuous mapping* $T : M \to M$ *is said to be* point (compact, bounded) dissipative *if there is a bounded set* B_0 *in* M *such that* B_0 *attracts each point (compact set, bounded set) in* M*;* T *is* κ-condensing (κ-contraction *of order* k, $0 \leq k < 1$) *if* T *takes bounded sets to bounded sets and* $\kappa(T(B)) < \kappa(B)(\kappa(T(B)) \leq k\kappa(B))$ *for any nonempty closed bounded set* $B \subset M$ *with* $0 < \kappa(B) < +\infty$*;* T *is* asymptotically smooth *if, for any nonempty closed bounded set* $B \subset M$ *for which* $T(B) \subset B$*, there is a compact set* $J \subset B$ *such that* J *attracts* B*.*

Clearly, a compact map is a κ-contraction of order 0, and a κ-contraction of order k is κ-condensing. It is well known that κ-condensing maps are asymptotically smooth (see, e.g., [7, Lemma 2.3.5]). By Lemma 2.1, it follows that $T : M \to M$ is asymptotically smooth if and only if $\lim_{n\to\infty} \kappa(T^n(B)) = 0$ for any nonempty closed bounded subset $B \subset M$ for which $T(B) \subset B$.

A positively invariant subset $B \subset M$ for T is said to be *stable* if for any neighborhood V of B there exists a neighborhood $U \subset V$ of B such that $T^n(U) \subset V \; \forall n \geq 0$. We say that A is *globally asymptotically stable* for T if, in addition, A attracts points of M for T.

By the proof that (i) implies (ii) in [7, Theorem 2.2.5], we have the following result.

Lemma 2.4. *Let* $B \subset M$ *be compact and positively invariant for* T*. If* B *attracts compact subsets of one of its neighborhoods, then* B *is stable.*

Definition 2.5. *A nonempty, compact, and invariant set* $A \subset M$ *is said to be an* attractor for T *if* A *attracts one of its neighborhoods; a* global attractor for T *if* A *is an attractor that attracts every point in* M*; and a* strong global attractor for T *if* A *attracts every bounded subset of* M*.*

We remark that the notions of attractor and global attractor were used in [11, 22, 30]. The strong global attractor was defined as a global attractor in [7, 24, 20]. The following result is essentially the same as [8, Theorem 3.2]. Note that the proof of this result was not provided in [8]. For completeness, we state it in terms of global attractors and give an elementary proof below.

Theorem 2.6. *Let* T *be a continuous map on a complete metric space* (M, d)*. Assume that*

(a) T *is point dissipative and asymptotically smooth;*

(b) *positive orbits of compact subsets of* M *for* T *are bounded.*

Then T *has a global attractor* $A \subset M$*. Moreover, for each subset* B *of* M*, if there exists* $k \geq 0$ *such that* $\gamma^+(T^k(B))$ *is bounded, then* A *attracts* B *for* T*.*

Proof. Assume that (a) is satisfied. Since T is point dissipative, we can find a closed and bounded subset B_0 in (M,d) such that, for each $x \in M$, there exists $k = k(x) \geq 0$, $T^n(x) \in B_0 \ \forall n \geq k$. Define

$$J(B_0) := \{y \in B_0 : T^n(y) \in B_0 \ \forall n \geq 0\}.$$

Thus, $T(J(B_0)) \subset J(B_0)$, and for every $x \in M$, there exists $k = k(x) \geq 0$ such that $T^k(x) \in J(B_0)$. Since $J(B_0)$ is closed and bounded, and T is asymptotically smooth, Lemma 2.2 implies that $\omega(J(B_0))$ is compact invariant and attracts points of M.

Assume, in addition, that (b) is satisfied. We claim that there exists an $\varepsilon > 0$ such that $\gamma^+(N(\omega(J(B_0)),\varepsilon))$ is bounded. Assume, by contradiction, that $\gamma^+(N(\omega(J(B_0)), \frac{1}{n+1}))$ is unbounded for each $n > 0$. Let $z \in M$ be fixed. Then we can find a sequence $x_n \in N(\omega(J(B_0)), \frac{1}{n+1})$, and a sequence of integers $m_n \geq 0$ such that $d(z, T^{m_n}(x_n)) \geq n$. Since $\omega(J(B_0))$ is compact, we can always assume that $x_n \to x \in \omega(J(B))$ as $n \to +\infty$. Since $H := \{x_n : n \geq 0\} \cup \{x\}$ is compact, assumption (b) implies that $\gamma^+(H)$ is bounded, which is a contradiction. Let $D = \overline{\gamma^+(N(\omega(J(B_0)),\varepsilon))}$. Then D is closed, bounded, and positively invariant for T. Since $\omega(J(B_0))$ attracts points of M for T, and $\omega(J(B_0)) \subset N(\omega(J(B_0)),\varepsilon) \subset \operatorname{int}(D)$, we deduce that, for each $x \in M$, there exists $k = k(x) \geq 0$ such that $T^k(x) \in \operatorname{int}(D)$. It then follows that, for each compact subset C of M, there exists an integer $k \geq 0$ such that $T^k(C) \subset D$. Thus, the set $A := \omega(D)$ attracts every compact subset of M. Fix a bounded neighborhood V of A. By Lemma 2.4, it follows that A is stable, and hence, there is a neighborhood W of A such that $T^n(W) \subset V \ \forall n \geq 0$. Clearly, the set $U := \cup_{n\geq 0} T^n(W)$ is a bounded neighborhood of A, and $T(\overline{U}) \subset \overline{U}$. Since T is asymptotically smooth, there is a compact set $J \subset \overline{U}$ such that J attracts $\overline{U}$. By Lemma 2.2, $\omega(\overline{U})$ is nonempty, compact, invariant for T, and attracts $\overline{U}$. Since A attracts $\omega(\overline{U})$, then $\omega(\overline{U}) \subset A$. Thus, A is a global attractor for T.

To prove the last part of the theorem, without loss of generality we assume that B is a bounded subset of M and $\gamma^+(B)$ is bounded. We set $K = \overline{\gamma^+(B)}$. Then $T(K) \subset K$. Since K is bounded and T is asymptotically smooth, there exists a compact C which attracts K for T. Note that $T^k(B) \subset T^k(\gamma^+(B)) \subset T^k(K) \ \forall k \geq 0$. Thus, C attracts B for T. By Lemma 2.2, we deduce that $\omega(B)$ is nonempty, compact, invariant for T, and attracts B. Since A is a global attractor for T, it follows that A attracts compact subsets of M. By the invariance of $\omega(B)$ for T, we deduce that $\omega(B) \subset A$, and hence, A attracts B for T. ☐

Remark 2.7. From the first part of the proof of Theorem 2.6, it is easy to see that if T is point dissipative and asymptotically smooth, then there exists a nonempty, compact, and invariant subset C of M for T such that C attracts every point in M for T.

The following lemma provides sufficient conditions for the positive orbit of a compact set to be bounded.

LEMMA 2.8. *Assume that T is point dissipative. If C is a compact subset of M with the property that, for every bounded sequence $\{x_n\}_{n\geq 0}$ in $\gamma^+(C)$, $\{x_n\}_{n\geq 0}$ or $\{T(x_n)\}_{n\geq 0}$ has a convergent subsequence, then $\gamma^+(C)$ is bounded in M.*

Proof. Since T is point dissipative, we can choose a bounded and open subset V of M such that for each $x \in M$ there exists $n_0 = n_0(x) \geq 0$ such that $T^n(x) \in V \ \forall n \geq n_0$. By the continuity of T and the compactness of C, it follows that there exists a positive integer $r = r(C)$ such that for any $x \in C$ there exists an integer $k = k(x) \leq r$ such that $T^k(x) \in V$. Let $z \in M$ be fixed. Assume, by contradiction, that $\gamma^+(C)$ is unbounded. Then there exists a sequence $\{x_p\}$ in $\gamma^+(C)$ such that

$$x_p = T^{m_p}(z_p),\ z_p \in C,\ \text{and}\ \lim_{p\to\infty} d(z, x_p) = \infty.$$

Since T is continuous and C is compact, without loss of generality we can assume that

$$\lim_{p\to\infty} m_p = \infty,\ \text{ and }\ m_p > r,\ x_p \notin V\ \forall p \geq 1.$$

For each $z_p \in C$, there exists an integer $k_p \leq r$ such that $T^{k_p}(z_p) \in V$. Since $x_p = T^{m_p}(z_p) \notin V$, there exists an integer $n_p \in [k_p, m_p)$ such that

$$y_p = T^{n_p}(z_p) \in V,\ \text{ and }\ T^l(y_p) \notin V\ \forall 1 \leq l \leq l_p = m_p - n_p.$$

Clearly, $x_p = T^{l_p}(y_p)\ \forall p \geq 1$, and $\{y_p\}$ is a bounded sequence in $\gamma^+(C)$.

We consider only the case where $\{y_p\}$ has a convergent subsequence, since the proof for the case where $\{T(y_p)\}$ has a convergent subsequence is similar. Thus, without loss of generality we can assume that $\lim_{p\to\infty} y_p = y \in \overline{V}$. In the case where the sequence $\{l_p\}$ is bounded, there exist an integer $\hat{l}$ and sequence $p_k \to \infty$ such that $l_{p_k} = \hat{l}\ \forall k \geq 1$, and hence,

$$d(z, T^{\hat{l}}(y)) = \lim_{k\to\infty} d(z, T^{\hat{l}}(y_{p_k})) = \lim_{k\to\infty} d(z, x_{p_k}) = \infty,$$

which is impossible. In the case where the sequence $\{l_p\}$ is unbounded, there exists a subsequence $l_{p_k} \to \infty$ as $k \to \infty$. Then for each fixed $m \geq 1$ there exists an integer k_m such that $m \leq l_{p_k}\ \forall k \geq k_m$, and hence,

$$T^m(y_{p_k}) \in M \setminus V\ \forall k \geq k_m.$$

Letting $k \to \infty$, we obtain

$$T^m(y) \in M \setminus V\ \forall m \geq 1,$$

which contradicts the definition of V. ∎

The following result on the existence of strong global attractors is implied by [8, Theorems 3.1 and 3.4]. Since the proof of this result was not provided in [8], we include a simple proof of it.

THEOREM 2.9. *Let T be a continuous map on a complete metric space (M, d). Assume that T is point dissipative on M, and one of the following conditions holds:*

(a) *T^{n_0} is compact for some integer $n_0 \geq 1$, or*

(b) *T is asymptotically smooth and, for each bounded set $B \subset M$, there exists $k = k(B) \geq 0$ such that $\gamma^+(T^k(B))$ is bounded.*

Then there is a strong global attractor A for T.

Proof. The conclusion in case (b) is an immediate consequence of Theorem 2.6. In the case of (a), since T^{n_0} is compact for some integer $n_0 \geq 1$, it suffices to show that for each compact subset $C \subset M$, $\cup_{n\geq 0} T^n(C)$ is bounded. By applying Lemma 2.8 to $\widetilde{T} = T^{n_0}$, we deduce that for each compact subset $C \subset M$, $\cup_{n\geq 0}\widetilde{T}^n(C)$ is bounded. So Theorem 2.6 implies that $\widetilde{T}$ has a global attractor $\widetilde{A} \subset M$. We set $\widetilde{B} = \cup_{0\leq k\leq n_0-1} T^k(\widetilde{A})$. By the continuity of T, it then follows that $\widetilde{B}$ is compact and attracts every compact subset of M for T, and hence, the result follows from Theorem 2.6. ∎

Remark 2.10. It is easy to see that a metric space (M, d) is complete if and only if for any subset B of M, $\kappa(B) = 0$ implies that $\overline{B}$ is compact. However, we can prove that Lemmas 2.2 and 2.4 also hold for noncomplete metric spaces by employing the equivalence between the compactness and the sequential compactness for metric spaces. It then follows that Theorems 2.6 and 2.9 are still valid for any metric space. We refer to [3, 20] for the existence of strong global attractors of continuous-time semiflows on a metric space.

3. Persistence and attractors. Let (M, d) be a complete metric space, and let $\rho : M \to [0, +\infty)$ be a continuous function. We define

$$M_0 := \{x \in M : \rho(x) > 0\} \text{ and } \partial M_0 := \{x \in M : \rho(x) = 0\}.$$

A subset $B \subset M_0$ is said to be *ρ-strongly bounded* if B is bounded in (M, d) and $\inf_{x \in B} \rho(x) > 0$. Throughout this section, we always assume that $T : M \to M$ is a continuous map with $T(M_0) \subset M_0$.

DEFINITION 3.1. *T is said to be* ρ-uniformly persistent *if there exists $\varepsilon > 0$ such that* $\liminf_{n\to+\infty} \rho(T^n(x)) \geq \varepsilon$, $\forall x \in M_0$*;* weakly ρ-uniformly persistent *if there exists $\varepsilon > 0$ such that* $\lim_{n\to+\infty} \sup \rho(T^n(x)) \geq \varepsilon\ \forall x \in M_0$. *The set ∂M_0 is said to be* ρ-ejective *for T if there exists $\varepsilon > 0$ such that for every $x \in M$ with $0 < \rho(x) < \varepsilon$ there is $n_0 = n_0(x) \geq 0$ such that $\rho(T^{n_0}(x)) \geq \varepsilon$.*

For a given open subset $M_0 \subset M$, let $\partial M_0 := M \setminus M_0$. Then we can use the continuous function $\rho : M \to [0, \infty)$, defined by $\rho(x) = d(x, \partial M_0)\ \forall x \in M$, to obtain the traditional definition of persistence.

PROPOSITION 3.2. *Assume that there is a compact subset C of M that attracts every point in M for T. Then the following statements are equivalent:*

(1) *T is weakly ρ-uniformly persistent.*

(2) *T is ρ-uniformly persistent.*

(3) *∂M_0 is ρ-ejective for T.*

Proof. The observations (1)⇔(3) and (2)⇒(1) are obvious. Let us prove that (1)⇒(2). Let $\varepsilon > 0$ be fixed such that

$$\lim_{n\to+\infty} \sup \rho(T^n(x)) \geq \varepsilon\ \forall x \in M_0. \tag{1}$$

Then for each $x \in M_0$, and each $n \geq 0$, there exists $p \geq 0$ such that $\rho(T^{n+p}(x)) \geq \varepsilon/2$. Assume that T is not ρ-uniformly persistent. Then we can find a sequence $\{x_m\}_{m\geq 0} \subset M_0$ such that

$$\lim_{n\to+\infty} \inf \rho(T^n(x_m)) \leq \frac{1}{m+1}\ \forall m \geq 0.$$

So there exist $l_m \geq 1$ and $n_m \geq 0$ such that

$$\begin{aligned} &d(T^{n_m}(x_m), C) \leq \frac{1}{m+1},\ \rho(T^{n_m}(x_m)) \geq \varepsilon/2, \\ &\rho\left(T^{n_m+k}(x_m)\right) \leq \varepsilon/2\,\forall\ k = 1, \ldots, l_m, \text{ and} \\ &\rho\left(T^{n_m+l_m}(x_m)\right) \leq \frac{1}{m+1}. \end{aligned}$$

Since C is compact, by taking a subsequence that we denote with the same index, we can always assume that $y_m = T^{n_m}(x_m) \to y \in C$. Since ρ and T are continuous, we deduce that

$$\rho(y) \geq \varepsilon/2, \text{ and } \rho\left(T^k(y)\right) \leq \varepsilon/2\ \ \forall k = 1, \ldots, l,$$

where $l = \lim_{m\to+\infty} \inf l_m$. If $l < +\infty$, we have $\rho\left(T^l(y)\right) = 0$, which is impossible because $T(M_0) \subset M_0$. If $l = +\infty$, we have

$$\lim_{n\to+\infty} \sup \rho\left(T^n(y)\right) \le \varepsilon/2 < \varepsilon,$$

which contradicts (1). ◻

We note that the concept of general ρ-persistence was used in [27, 23, 30]. It was also shown in [27] that the ρ-uniform persistence implies the weak ρ-uniform persistence for nonautonomous semiflows under appropriate conditions. The following result shows that the notion of ρ-uniform persistence is independent of the choice of continuous function ρ.

PROPOSITION 3.3. *Let $\xi : M \to [0,+\infty)$ be a continuous function such that $\partial M_0 = \{x \in M : \xi(x) = 0\}$. Assume that there is a compact subset of M that attracts every point in M. Then T is ρ-uniformly persistent if and only if T is ξ-uniformly persistent.*

Proof. It suffices to prove that ρ-uniform persistence implies ξ-uniform persistence since the problem is symmetric. Let us first remark that T is ρ-uniformly persistent if and only if there exists $\varepsilon > 0$ such that

$$\inf_{x\in M_0} \inf_{y\in\omega(x)} \rho(y) \ge \varepsilon,$$

where $\omega(x)$ is the omega-limit set of the positive orbit of x. Define

$$A_\omega = \cup_{x\in M_0}\omega(x) \text{ and } V = \{y \in M : \rho(y) \ge \varepsilon\}.$$

Then

$$\inf_{x\in M_0} \inf_{y\in\omega(x)} \rho(y) = \inf_{x\in A_\omega} \rho(x) \ge \varepsilon.$$

Clearly, $A_\omega \subset C$, so $\overline{A}_\omega$ is compact. Since A_ω is included in $V \subset M_0$, which is closed, we deduce that $\overline{A}_\omega \subset V \cap C \subset M_0$. So $\overline{A}_\omega \subset M_0$ is compact, and hence, there exists $\eta > 0$ such that $\inf_{x\in\overline{A_\omega}} \xi(x) \ge \eta$, which implies that T is ξ-uniformly persistent. ◻

Let A be a nonempty subset of M. A is said to be *ejective for T* if there exists a neighborhood V of A such that for every $x \in (M \setminus A) \cap V$ there is $n_0 = n_0(x) \ge 0$ such that $T^{n_0}(x) \in M \setminus V$.

PROPOSITION 3.4. *Assume that $\partial M_0 \ne \emptyset$ and that there is a compact subset C of M which attracts every point in M for T. Then the following statements are equivalent:*

(1) *T is ρ-uniformly persistent.*

(2) *∂M_0 is ejective for T.*

Proof. Assume first that (1) is true. Let $\varepsilon > 0$ be fixed such that

$$\lim_{n\to+\infty} \sup \rho\left(T^n(x)\right) \ge \varepsilon \ \forall x \in M_0.$$

Then it is clear that ∂M_0 is ejective for T, with $V = \{x \in M : \rho(x) \le \varepsilon/2\}$.

Conversely, assume that ∂M_0 is ejective for T. Let V be a neighborhood of ∂M_0 such that for every $x \in M_0 \cap V$ there is $n_0 = n_0(x) \ge 0$ such that $T^{n_0}(x) \in M \setminus V$. By Proposition 3.3, it is sufficient to prove that T is ρ-uniformly persistent when

$\rho(x) = d(x, \partial M_0)$. Assume, by contradiction, that T is not ρ-uniformly persistent. Then for each $n \geq 1$ there exists $x_n \in M_0$, such that

$$\lim_{m\to+\infty} \sup \rho\left(T^m(x_n)\right) \leq \frac{1}{n}.$$

By the attractivity of C, it follows that for each $n \geq 1$ there exists $l_n \geq 0$ such that each $y_n := T^{l_n}(x_n) \in M_0$ satisfies

$$d\left(T^k\left(y_n\right), C\right) \leq \frac{2}{n} \text{ and } d\left(T^k\left(y_n\right), \partial M_0\right) \leq \frac{2}{n} \ \forall k \geq 0.$$

Since C is compact and V is a neighborhood of ∂M_0, there exists $\delta > 0$ such that

$$\{x \in M : d(x, C) \leq \delta \text{ and } d(x, \partial M_0) \leq \delta\} \subset V.$$

Let $n_0 \geq 2/\delta$ be fixed. Then we have $y_{n_0} \in M_0$, and

$$d\left(T^k\left(y_{n_0}\right), C\right) \leq \delta \text{ and } d\left(T^k\left(y_n\right), \partial M_0\right) \leq \delta \ \forall k \geq 0.$$

Thus, we obtain

$$y_{n_0} \in M_0 \cap V \text{ and } T^k\left(y_{n_0}\right) \in V \ \forall k \geq 0,$$

which is a contradiction. □

Observe that M_0 is an open subset in (M, d). In order to make M_0 become a complete metric space, we define a new metric function d_0 on M_0 by

$$d_0(x, y) = \left|\frac{1}{\rho(x)} - \frac{1}{\rho(y)}\right| + d(x, y) \ \forall x, y \in M_0. \tag{2}$$

LEMMA 3.5. *(M_0, d_0) is a complete metric space.*

Proof. It is easy to see that d_0 is a metric function. Let $\{x_n\}_{n\geq 0}$ be a Cauchy sequence in (M_0, d_0). Since $d(x, y) \leq d_0(x, y) \ \forall x, y \in M_0$, we deduce that $\{x_n\}_{n\geq 0}$ is a Cauchy sequence in (M, d), and there exists $x \in M$, such that $d(x_n, x) \to 0$ as $n \to +\infty$. To prove that $d_0(x_n, x) \to 0$ as $n \to +\infty$, it is sufficient to show that $x \in M_0$. Given $\varepsilon > 0$, since $\{x_n\}_{n\geq 0}$ is a Cauchy sequence in (M_0, d_0), there exists $n_0 \geq 0$ such that $d_0(x_n, x_p) \leq \varepsilon \ \forall n, p \geq n_0$. In particular, we have $d_0(x_n, x_{n_0}) \leq \varepsilon \ \forall n \geq n_0$. Then

$$\left|\frac{1}{\rho(x_n)} - \frac{1}{\rho(x_{n_0})}\right| \leq \varepsilon \ \forall n \geq n_0.$$

So there exists $r > 0$ such that $\inf_{n\geq 0} \rho(x_n) \geq r$. Since ρ is continuous and $d(x_n, x) \to 0$ as $n \to +\infty$, we deduce that $\rho(x) \geq r$, and hence $x \in M_0$. Thus, (M_0, d_0) is complete. □

For any two subsets $A, B \subset M$, we denote

$$\delta(B, A) = \sup_{x\in B} \inf_{y\in A} d(x, y),$$

and if $A, B \subset M_0$, we denote

$$\delta_0(B, A) = \sup_{x\in B} \inf_{y\in A} d_0(x, y).$$

LEMMA 3.6. *The following two statements are valid:*

(1) *Let* $\{B_t\}_{t\in I}$ *be a family of subsets of* M_0, *where* I *is a unbounded subset of* $[0,+\infty)$. *If* $A \subset M_0$ *is compact in* (M,d) *and* $\lim_{t\to\infty} \delta(B_t, A) = 0$, *then* $\lim_{t\to\infty} \delta_0(B_t, A) = 0$.

(2) *If* T *is asymptotically smooth, then* T *is asymptotically smooth in* (M_0, d_0).

Proof. (1) We denote $k := \frac{1}{2}\inf_{x\in A}\rho(x) > 0$. Assume, by contradiction, that $\lim_{t\to+\infty}\sup \delta_0(B_t, A) > \varepsilon > 0$. Then we can find a sequence $\{t_p\}_{p\geq 0} \subset I$ such that $t_p \to +\infty$, $p \to +\infty$, and a sequence $\left\{x_{t_p}\right\}_{p\geq 0} \subset M_0$ such that $x_{t_p} \in B_{t_p}$, $d_0\left(x_{t_p}, A\right) \geq \varepsilon\ \forall p \geq 0$. Since $d\left(x_{t_p}, A\right) \to 0$, as $p \to +\infty$, without loss of generality we can assume that there exists $x \in A$ such that $d(x_{t_p}, x) \to 0$, as $p \to +\infty$. Since ρ is continuous and $\rho(x) > k$, there exists $p_0 \geq 0$ such that $\rho(x_{t_p}) \geq k\ \forall p \geq p_0$. Thus, we have

$$0 < \varepsilon \leq d_0\left(x_{t_p}, x\right) \leq k^{-2}\left|\rho(x_{t_p}) - \rho(x)\right| + d(x_{t_p}, x) \to 0 \text{ as } p \to +\infty,$$

which is a contradiction.

(2) It is easy to see that $T : (M_0, d_0) \to (M_0, d_0)$ is continuous. Let B be a bounded subset in (M_0, d_0) such that $T(B) \subset B$. Since T is asymptotically smooth, there exists a compact subset $C \subset M$ that attracts B for T. So $C_0 = C \cap \overline{B} \subset M_0$ is compact and attracts B for T. It easily follows that C_0 is also compact in (M_0, d_0). Since C_0 attracts B for T, statement (1) implies that C_0 attracts B for $T : (M_0, d_0) \to (M_0, d_0)$. ∎

The main result of this section is the following theorem.

THEOREM 3.7. *Assume that* T *is asymptotically smooth and* ρ*-uniformly persistent, and that* T *has a global attractor* A. *Then* $T : (M_0, d) \to (M_0, d)$ *has a global attractor* A_0. *Moreover, for each subset* B *of* M_0, *if there exists* $k \geq 0$ *such that* $\gamma^+\left(T^k(B)\right)$ *is* ρ*-strongly bounded, then* A_0 *attracts* B *for* T.

Proof. We consider the continuous map $T : (M_0, d_0) \to (M_0, d_0)$. Since T is point dissipative and ρ-uniformly persistent, T is point dissipative in (M_0, d_0). Moreover, Lemma 3.6 implies that T is asymptotically smooth in (M_0, d_0). Let C be a compact subset in (M_0, d_0), and $\{x_p\}$ a bounded sequence in $\gamma^+(C)$ in (M_0, d_0). Then $x_p = T^{m_p}(z_p)$, $z_p \in C\ \forall p \geq 1$, and the sequence $\{x_p\}$ is ρ-strongly bounded in (M, d). Since C is also compact in (M, d), we have $\lim_{m\to\infty}\delta(T^m(C), A) = 0$. Thus, $\{x_p\}$ has a convergent subsequence $x_{p_k} \to x$ in (M, d) as $k \to \infty$. By the continuity of ρ and the ρ-strong boundedness of $\{x_p\}$, it follows that $\rho(x) > 0$, i.e., $x \in M_0$, and hence, $x_{p_k} \to x$ in (M_0, d_0) as $k \to \infty$. Thus, Lemma 2.8 implies that positive orbits of compact sets are bounded for $T : (M_0, d_0) \to (M_0, d_0)$. Then the conclusion for $T : (M_0, d) \to (M_0, d)$ follows from Theorem 2.6, as applied to $T : (M_0, d_0) \to (M_0, d_0)$. ∎

THEOREM 3.8. *Assume that* T *is point dissipative on* M *and* ρ*-uniformly persistent, and that one of the following conditions holds:*

(a) *There exists some integer* $n_0 \geq 1$ *such that* T^{n_0} *is compact on* M, *and* T^{n_0} *maps* ρ*-strongly bounded subsets of* M_0 *onto* ρ*-strongly bounded sets in* M_0, *or*

(b) T *is asymptotically smooth on* M, *and for every* ρ*-strongly bounded subset* $B \subset M_0$, *there exists* $k = k(B) \geq 0$ *such that* $\gamma^+\left(T^k(B)\right)$ *is* ρ*-strongly bounded in* M_0.

Then $T : (M_0, d) \to (M_0, d)$ *has a global attractor* A_0, *and* A_0 *attracts every* ρ*-strongly bounded subset in* M_0 *for* T.

Proof. Clearly, $T : (M_0, d_0) \to (M_0, d_0)$ is point dissipative. It is easy to see that condition (a) implies that $T^{n_0} : (M_0, d_0) \to (M_0, d_0)$ is compact, and that condition (b) implies that condition (b) of Theorem 2.9 holds for $T : (M_0, d_0) \to (M_0, d_0)$.

By Theorem 2.9, there is a strong global attractor A_0 for $T : (M_0, d_0) \to (M_0, d_0)$. Consequently, A_0 is a global attractor for $T : (M_0, d) \to (M_0, d)$, and A_0 attracts every ρ-strongly bounded subset in M_0 for T. ∎

Remark 3.9. A result similar to Theorem 3.8 was already presented for discrete- and continuous-time dynamical systems in [29] and [10], respectively. The only difference, compared with the earlier results, is that we add a ρ-boundedness assumption for case (a). In fact, this assumption is necessary for the existence of a strong global attractor in M_0 for T (see two examples in section 5.1).

Remark 3.10. A family of mappings $\Phi(t) : M \to M$, $t \geq 0$, is called a continuous-time semiflow if $(x, t) \to \Phi(t)x$ is continuous, $\Phi(0) = Id$, and $\Phi(t) \circ \Phi(s) = \Phi(t + s)$ for $t, s \geq 0$. By similar arguments we can prove the analogues of Theorems 3.7 and 3.8 for a continuous-time semiflow $\Phi(t)$ on M with $\Phi(t)(M_0) \subset M_0$ $\forall t \geq 0$.

4. Coexistence steady states. In this section, we establish the existence of a coexistence steady state (i.e., the fixed point in M_0) for uniformly persistent dynamical systems.

Throughout this section we always assume that M is a closed and convex subset of a Banach space $(X, \|\cdot\|)$, that $\rho : M \to [0, +\infty)$ is a continuous function such that $M_0 = \{x \in M : \rho(x) > 0\}$ is nonempty and convex, and that $T : M \to M$ is a continuous map with $T(M_0) \subset M_0$. For convenience, we set $\partial M_0 := M \setminus M_0$.

Assume that $T : M_0 \to M_0$ has a global attractor A_0. By Definition 2.5, it easily follows that for every compact set $K \subset M_0$ there exists an open neighborhood of K which is attracted by A_0. This property of A_0 is enough to support the arguments in the proof of [29, Theorem 2.3] (see also [30, Theorem 1.3.6]) instead of the property that A_0 attracts ρ-strongly bounded sets in M_0. Thus, the proof of [29, Theorem 2.3] actually implies the following fixed point theorem.

THEOREM 4.1. *Assume that T is κ-condensing. If $T : M_0 \to M_0$ has a global attractor A_0, then T has a fixed point $x_0 \in A_0$.*

Note that a fixed point theorem for κ-condensing maps in [9] was used in the proof of [29, Theorem 2.3]. To generalize Theorem 4.1 to another class of maps, we need the following fixed point theorem, which is a combination of Theorems 3 and 5 in [9] (see also [7, Lemma 2.6.5]).

LEMMA 4.2 (Hale–Lopes fixed point theorem). *Assume that $K \subset B \subset S$ are convex subsets of a Banach space X, with K compact, S closed and bounded, and B open in S. If $T : S \to X$ is continuous, $T^n B \subset S$ $\forall n \geq 0$, and K attracts compact subsets of B, then there exists a closed bounded and convex subset $C \subset S$ such that $C = \overline{co}\left(\cup_{j \geq 1} T^j (B \cap C)\right)$. Moreover, if C is compact, then T has a fixed point in B.*

We should point out that in the above fixed point theorem the claim that T has a fixed point in B follows from the proof of [7, Lemma 2.6.5], where Horn's fixed point theorem [12] was used.

Motivated by Lemma 4.2 and the Poincaré maps associated with age-structured population models, we give the following definition.

DEFINITION 4.3. *Let M be a closed and convex subset of a Banach space X, and let $T : M \to M$ be a continuous map. Define $\widehat{T}(B) = \overline{co}\,(T(B))$ for each $B \subset M$. T is said to be convex κ-contracting if $\lim_{n\to\infty} \kappa(\widehat{T}^n(B)) = 0$ for each bounded subset $B \subset M$.*

Now we are ready to generalize Theorem 4.1 to convex κ-contracting maps.

THEOREM 4.4. *Assume that T is convex κ-contracting. If $T : M_0 \to M_0$ has a global attractor A_0, then T has a fixed point $x_0 \in A_0$.*

Proof. Since A_0 is a global attractor for $T : M_0 \to M_0$, the proof of [29, Theorem

2.3] (see also [30, Theorem 1.3.6]) implies that there are three convex subsets, $K \subset B \subset S \subset M$, such that $K \subset M_0$, $B \subset M_0$, and the assumptions of Lemma 4.2 hold for T. Let C be defined in Lemma 4.2. Define $\widehat{C} := \cup_{j\geq 1} T^j (B \cap C)$. Then we have

$$\widehat{C} = T(B \cap C) \cup T\left(\widehat{C}\right) \text{ and } C = \overline{co}\left(\widehat{C}\right),$$

and hence, $\widehat{C} \subset T(C)$. Thus, we further obtain

$$C \subset \widehat{T}(C) \subset \widehat{T}^2(C) \subset \cdots \subset \widehat{T}^n(C) \; \forall n \geq 0.$$

Since T is convex κ-contracting, it follows that $\kappa(C) \leq \kappa(\widehat{T}^n(C)) \to 0$ as $n \to \infty$. Then $\kappa(C) = 0$, and hence, C is compact. Now Lemma 4.2 implies the existence of a fixed point of T in A_0. ∎

Combining Theorems 2.6, 2.9, 3.7, 4.1, and 4.4 we have the following result on the existence of coexistence steady states for uniformly persistent systems, which is a generalization of [29, Theorem 2.3].

THEOREM 4.5. *Assume that*

(1) *T is point dissipative and ρ-uniformly persistent.*

(2) *One of the following two conditions holds:*

  (2a) *T^{n_0} is compact for some integer $n_0 \geq 1$, or*

  (2b) *for each compact subset $C \subset M$, $\gamma^+(C)$ is bounded.*

(3) *Either T is κ-condensing or T is convex κ-contracting.*

Then $T : M_0 \to M_0$ admits a global attractor A_0, and T has a fixed point in A_0.

Let $A \subset M$ and $B \subset M \setminus A$. A is said to be *ejective for T in B* if there exists a neighborhood V of A such that for each $x \in V \cap B$ there exists $n = n(x) \geq 0$ such that $T^n(x) \in M \setminus V$. A is said to be *ejective for T* if A is ejective for T in $M \setminus A$.

The following corollary is a generalization of [15, Theorem 4.1] on semi-ejective fixed points.

COROLLARY 4.6. *Assume that $T(\partial M_0) \subset \partial M_0$ and that there exists $\overline{x}_\partial \in \partial M_0$, a fixed point of T, which is globally asymptotically stable for $T : \partial M_0 \to \partial M_0$. Assume, in addition, that*

(1) *T is point dissipative and $\overline{x}_\partial$ is ejective for T in M_0.*

(2) *One of the following two conditions holds:*

  (2a) *T^{n_0} is compact for some integer $n_0 \geq 1$, or*

  (2b) *positive orbits of compact subsets of M are bounded.*

(3) *Either T is κ-condensing or convex κ-contracting.*

Then $T : M_0 \to M_0$ admits a global attractor A_0, and T has a fixed point in A_0.

Proof. By [29, Theorem 2.2] (see also [30, Theorem 1.3.1]), we deduce that T is ρ-uniformly persistent with $\rho(x) = d(x, \partial M_0)$. Now Theorem 4.5 completes the proof. ∎

We remark that when $\partial M_0 = \{\overline{x}_\partial\}$ in Corollary 4.6, we obtain a generalization of the classical Browder [2] ejective fixed point theorem.

A point $e \in M$ is said to be an equilibrium of a continuous-time semiflow $\Phi(t)$ on M if $\Phi(t)e = e \; \forall \, t \geq 0$. As a consequence of Theorems 4.1 and 4.4 and the proof of [29, Theorem 2.4] (see also [30, Theorem 1.3.7]), we have the following result on the existence of equilibrium in M_0 for $\Phi(t)$.

THEOREM 4.7. *Let $\Phi(t)$ be a continuous-time semiflow on M with $\Phi(t)(M_0) \subset M_0 \; \forall \, t \geq 0$. Assume that either $\Phi(t)$ is κ-condensing for each $t > 0$, or $\Phi(t)$ is convex κ-contracting for each $t > 0$, and that $\Phi(t) : M_0 \to M_0$ has a global attractor A_0. Then $\Phi(t)$ has an equilibrium $x_0 \in A_0$.*

In the rest of this section, we establish sufficient conditions for T to be convex κ-contracting.

LEMMA 4.8. *Let M be a closed and convex subset of a Banach space X, and let $T : M \to M$ be a continuous map which takes bounded sets to bounded sets. Assume that there exists a sequence of bounded linear operators $\{P_k\}_{k\geq 1} \in \mathcal{L}(X, X)$ such that*

(1) *for each bounded subset $B \subset M$, $(Id - P_1)T(B)$ is relatively compact;*
(2) *one of the following conditions holds:*
  (2a) *There exists $n_0 \geq 0$ such that P_{n_0} is compact, and if $k \in \{1, \ldots, n_0 - 1\}$, $C \subset M$, and $(Id - P_k)C$ is compact, then $(Id - P_{k+1})\,T(C)$ is compact.*
  (2b) *There exists $c \in (0, 1)$ such that $\|P_{k+1}T(x)\| \leq c\,\|P_k x\|$ $\forall\, x \in M$, $\forall\, k \geq 1$, and if $k \geq 1$, $C \subset M$, and $(Id - P_k)C$ is compact, then $(Id - P_{k+1})\,T(C)$ is compact.*

Then T is convex κ-contracting.

Proof. Let $B \subset M$ be a bounded subset of M. Since $(Id - P_1)T(B)$ is relatively compact and P_1 is linear, it follows that

$$(I - P_1)\overline{co}(T(B)) = \overline{co}((I - P_1)T(B)) \text{ is compact,}$$

$$\text{and } (Id - P_1)\overline{co}(T(B)) \text{ is compact.}$$

Thus, $(Id - P_2)\overline{co}(T(\overline{co}(T(B))))$ is compact, and, by induction, $(Id - P_{k+1})\widehat{T}^k(B)$ is compact $\forall k \in \{1, \ldots, n_0 - 1\}$ if (2a) holds, and $\forall\, k \geq 1$ if (2b) holds. If (2a) holds, since P_{n_0} is compact, we deduce that $\widehat{T}^{n_0}(B)$ is compact, and hence, $\kappa(\widehat{T}^n(B)) = 0$ $\forall n \geq n_0$. If (2b) holds, then the boundedness of linear operator P_1 implies that

$$\sup_{y\in\overline{co}(T(B))} \|P_1 y\| = \sup_{x\in T(B)} \|P_1 x\| \leq c \sup_{x\in B} \|x\|\,.$$

Similarly, we have

$$\begin{aligned}\sup_{y\in\overline{co}(T(\widehat{T}(B)))} \|P_2 y\| = \sup_{x\in T(\widehat{T}(B))} \|P_2 x\| &\leq c \sup_{x\in\widehat{T}(B)} \|P_1 x\| \\ &\leq c^2 \sup_{x\in B} \|x\|\,.\end{aligned}$$

By induction, it follows that

$$\sup_{y\in\widehat{T}^k(B)} \|P_k y\| \leq c^k \sup_{x\in B} \|x\| \;\; \forall k \geq 1.$$

Let $\delta_k := c^k \sup_{x\in B} \|x\|$. Since $(Id - P_k)\widehat{T}^k(B)$ is compact, there exists $x_1, \ldots, x_{m(k)} \in (Id - P_k)\widehat{T}^k(B)$ such that

$$(Id - P_k)\widehat{T}^k(B) \subset \cup_{j=1,\ldots,m(k)} B_M\left(x_j, \delta_k\right),$$

where $B_M\left(x_j, \delta_k\right) = \{x \in M : \|x - x_j\| < \delta_k\}$. Thus, we have

$$(Id - P_k)\widehat{T}^k(B) + P_k\widehat{T}^k(B) \subset \cup_{j=1,\ldots,m(k)} B_M\left(x_j, 2\delta_k\right).$$

Since $\widehat{T}^k(B) \subset (Id - P_k)\widehat{T}^k(B) + P_k\widehat{T}^k(B)$, it follows that

$$\kappa(\widehat{T}^k(B)) \leq \kappa\left((Id - P_k)\widehat{T}^k(B) + P_k\widehat{T}^k(B)\right) \leq 2\delta_k \to 0 \text{ as } k \to +\infty.$$

Thus, T is convex κ-contracting. $\square$

We complete this section with an example of convex κ-contracting maps.

Example 4.9. Consider $T : L^1_+(0,c) \to L^1_+(0,c)$, with $c \in (1,+\infty]$, defined by

$$T(\varphi)(a) = \begin{cases} \chi(\varphi)\varphi(a-1) \text{ if } 1 \leq a < c, \\ \lambda \text{ if } a \in (0,1), \end{cases}$$

where $\lambda > 0$, and $\chi : L^1_+(0,c) \to [0,\alpha]$ (with $0 < \alpha$) is a continuous map. We choose for each integer $k \geq 1$, $P_k : L^1(0,c) \to L^1(0,c)$ the operator defined by

$$P_k(\varphi)(a) = \begin{cases} \varphi(a) \text{ if } a \in (0,c) \cap (k,+\infty), \\ 0 \text{ otherwise.} \end{cases}$$

If $c < +\infty$, then (2a) holds. If $c = +\infty$ and $\alpha < 1$, then (2b) holds. Thus, Lemma 4.8 implies that T is convex κ-contracting. Note that in this example we need to impose some additional conditions on χ to show that T is κ-condensing.

5. Five examples. In this section, we first provide four examples of discrete- and continuous-time semiflows which admit global attractors, but no strong global attractors, in the complete metric spaces (M_0, d_0) introduced in section 3. Then we give an example showing applicability of Theorem 4.5 in the case of a convex κ-contracting map. Our examples are highly motivated by age-structured population models. We refer to Webb [28], Iannelli [14], and Anita [1] for the classical approach and to Thieme [25] and Magal and Thieme [16] (and references therein) for the integrated semigroup approach to this class of evolutionary equations.

5.1. Asymptotically smooth semiflows on (0 0). Let $C([0,1],\mathbb{R})$ be endowed with the usual norm $\|\varphi\|_\infty = \sup_{a\in[0,1]} |\varphi(a)|$. Let $M := C_+([0,1],\mathbb{R})$ be endowed with the metric $d(x,y) = \|x-y\|$, and $T : M \to M$ be defined by

$$T(\varphi) = \delta \frac{\mathcal{F}_\beta(\varphi)}{1+\mathcal{F}_\beta(\varphi)} 1_{[0,1]},$$

where $1_{[0,1]}(a) = 1\ \forall a \in [0,1]$, and $\mathcal{F}_\beta(\varphi) = \int_0^1 \beta(a)\varphi(a)da\ \forall \varphi \in X$. We assume that

(A1) $\delta > 1, \beta \in C([0,1],\mathbb{R}),\ \int_0^1 \beta(a)da = 1,\ \beta(a) > 0\ \forall a \in [0,1)$, and $\beta(1) = 0$.

Consider the following discrete-time system on M:

$$u_{n+1} = T(u_n)\ \ \forall n \geq 0, \text{ and } u_0 \in M.$$

It is easy to see that the map T is continuous and maps bounded sets into compact sets of M. Note that $T(M) \subset [0,\delta]\, 1_{[0,1]} = \left\{\alpha 1_{[0,1]} : \alpha \in [0,\delta]\right\}$ is bounded. So T is compact and point dissipative and has a strong global attractor in M. Set

$$\partial M_0 = \{0\}, \quad M_0 = M \setminus \{0\}, \quad \rho(x) = \|x\|_\infty.$$

Clearly, $T(M_0) \subset M_0$, $T(\partial M_0) \subset \partial M_0$, and the fixed points of T are 0 and $\overline{u} = (\delta-1)1_{[0,1]}$. Then it is easy to see that for each $\varphi \in M_0$, $T^m(\varphi) \to \overline{u}$, as $m \to +\infty$. So T is ρ-uniformly persistent. Let $\overline{\alpha} = (\delta - 1)$ and $B := \{x \in M : \|x\|_\infty = \overline{\alpha}\}$. Since $\beta(1) = 0$, we have $\mathcal{F}_\beta(B) = (0,\overline{\alpha}]$. Moreover, $T(B) = \{\alpha 1_{[0,1]} : \alpha \in (0,\overline{\alpha}]\}$, and $T^n(B) = T(B)\ \forall n \geq 1$. Thus, there exists no compact subset in M_0 that attracts B for T. In particular, there is no strong global attractor for $T : (M_0,d_0) \to (M_0,d_0)$, where d_0 is defined as in (2).

Next we consider the continuous-time semiflow $\{U(t)\}_{t\geq 0}$ on $M := L^1_+(0,1)$, which is generated by the following age-structured model:

$$(3)\qquad \begin{cases} \frac{\partial u(t)}{\partial t} + \frac{\partial u(t)}{\partial a} = -\mu(a)u(t)(a) - \mathcal{F}_\Gamma(u(t))\,u(t)(a), \quad a \in (0,1), \\ u(t,0) = \mathcal{F}_\beta(u(t)), \\ u(0) = \varphi \in L^1_+(0,1), \end{cases}$$

where for each $\chi \in L^\infty(0,1)$, and each $\varphi \in L^1(0,1)$, $\mathcal{F}_\chi(\varphi) = \int_0^1 \chi(a)\varphi(a)da$. We assume that

(A2) $\beta \in (0,+\infty)$, $\mu \in L^1_{loc}[0,1)$, $\mu \geq 0$, $\lim_{a\to 1^-}\int_0^a \mu(r)dr = +\infty$, $\int_0^1 \beta\exp(-\int_0^a \mu(s)ds)da > 1$, and

$$\Gamma(a) = \frac{1}{\int_0^1 \exp(-\int_0^s \mu(r) + \lambda_0 dr)ds}\int_a^1 \exp\left(-\int_a^s \mu(r) + \lambda_0 dr\right)ds \;\forall a \in [0,1],$$

where $\lambda_0 > 0$ is the unique solution of $\int_0^1 \beta\exp\left(-\int_0^a \mu(s) + \lambda_0 ds\right)da = 1$.

Let $\{T(t)\}_{t\geq 0}$ be the C_0-semigroup of bounded linear operators generated by $A : D(A) \subset L^1(0,1) \to L^1(0,1)$ with

$$A\varphi = -\frac{d\varphi}{da} - \mu\varphi \;\forall\; \varphi \in D(A),$$
$$D(A) = \left\{\varphi \in W^{1,1}(0,1) : \mu\varphi \in L^1(0,1) \text{ and } \varphi(0) = \int_0^1 \beta\varphi(a)da\right\}.$$

Let $P : L^1(0,1) \to L^1(0,1)$ be the bounded linear operator of projection defined by

$$P(\varphi)(a) = \int_0^1 \Gamma(s)\varphi(s)ds\chi(a) \;\forall\varphi \in L^1(0,1),$$

where $\chi(a) = \alpha\exp\left(-\int_0^a \mu(s) + \lambda_0 ds\right)$, and

$$\alpha = \left(\int_0^1 \Gamma(s)\exp\left(-\int_0^a \mu(s) + \lambda_0 ds\right)ds\right)^{-1}.$$

Then $PT(t) = T(t)P = e^{\lambda_0 t}P \;\forall t \geq 0$, and there exist $\delta > 0$ and $M \geq 1$ such that

$$\|(Id - P)T(t)\| \leq Me^{(\lambda_0-\delta)t} \;\forall t \geq 0.$$

Moreover, we have

$$U(t)x = \frac{T(t)x}{1 + \int_0^t \mathcal{F}_\Gamma(T(s)x)\,ds} \;\forall t \geq 0, \forall x \in M.$$

It is easy to see that for each $\varphi \in M_0$, $U(t)\varphi \to \lambda_0\chi$, as $t \to +\infty$. Since $T(t)$ is compact for $t \geq 2$, $U(t)$ is compact for $t \geq 2$. So $\{U(t)\}_{t\geq 0}$ has a strong global attractor. Set

$$\partial M_0 = \{0\}, \quad M_0 = M \setminus \{0\}, \quad \rho(\varphi) = \|\varphi\|_{L^1(0,1)} \;\forall\varphi \in M.$$

Since $T(t)$ is irreducible, we have $U(t)(\partial M_0) \subset \partial M_0$, and $U(t)M_0 \subset M_0 \;\forall t \geq 0$. Since $U(t)\varphi \to \lambda_0\chi$, as $t \to +\infty$, we deduce that $U(t)$ is ρ-uniformly persistent. So $U(t) : (M_0, d) \to (M_0, d)$ has a global attractor. Let

$$B := \left\{\varphi \in L^1_+(0,1) : \|\varphi\|_{L^1(0,1)} = 1\right\}.$$

Then B is ρ-strongly bounded. Since $\Gamma(1) = 0$, we deduce that there exists $c > 0$ such that $(0, c] \subset \mathcal{F}_\Gamma(B)$. We further claim that for each $\varepsilon > 0$ and $t_0 > 0$, there exist $t_1 > t_0$ and $\varphi \in B$ such that $\|U(t_1)\varphi\|_{L^1(0,1)} < \varepsilon$. Indeed, given $\varepsilon > 0$ and $t_0 > 0$, we can choose $t_1 > t_0$ such that $Me^{-\delta t_1} \leq \varepsilon/2$. Then for every $\varphi \in B$, we have

$$\|U(t_1)\varphi\| \leq \frac{\mathcal{F}_\Gamma(\varphi)\|\chi\|}{\left[1 - \frac{\mathcal{F}_\Gamma(\varphi)}{\lambda_0}\right] e^{-\lambda_0 t_1} + \frac{\mathcal{F}_\Gamma(\varphi)}{\lambda_0}} + \frac{\varepsilon}{2},$$

and hence, by choosing $\varphi \in B$ with $\mathcal{F}_\Gamma(\varphi)$ small enough, we obtain $\|U(t_1)\varphi\| \leq \varepsilon$. This claim shows that for each $t_0 > 0$, $\cup_{t \geq t_0} U(t)B$ is not ρ-strongly bounded. So there exists no compact set in M_0 that attracts B for $U(t)$. In particular, there exists no strong global attractor for the semiflow $U(t) : (M_0, d_0) \to (M_0, d_0)$, where d_0 is defined as in (2).

5.2. -contracting maps on (0 0). In this subsection, we construct κ-contracting maps on (M_0, d_0) such that they admit a global attractor, but no strong global attractor.

We set

$$X = L^1((0, +\infty), \mathbb{R}) \times \mathbb{R}, \; X_+ = L^1_+((0, +\infty), \mathbb{R}) \times \mathbb{R}_+$$

and endow X with the product norm $\|(\varphi, y)\| = \|\varphi\|_{L^1} + |y|$. Define $1_{[0,1]} \in X$ by $1_{[0,1]}(l) = 1 \; \forall l \in (0, 1)$, and $1_{[0,1]}(l) = 0 \; \forall l \in [1, \infty)$. Let a, b, and c be three real numbers. Define $T : X_+ \to X_+$ by $T(\varphi, y) = (T_1(\varphi, y), T_2(\varphi, y))$ with

$$T_1(\varphi, y) = a\varphi(\cdot + 1) + \left[a \int_0^1 \varphi(l)dl + c \frac{\int_0^1 \varphi(l)dl}{1 + \|(\varphi, y)\|}\right] 1_{[0,1]},$$
$$T_2(\varphi, y) = ay + b \frac{\|(\varphi, y)\|}{1 + \|(\varphi, y)\|}.$$

We assume that

(A3) $a \in (0, 1)$, $b > 0$, $c > 0$, $\sqrt{a} < a + b < 1$, and $a + c > 1$.

Consider the discrete-time system

$$x_{n+1} = T(x_n) \; \forall n \geq 0, \text{ and } x_0 \in X_+.$$

It is easy to see that $T^n(0, y) \to 0$, as $n \to +\infty$. Clearly, T is not uniformly persistent for $X_+ \setminus \{0\}$. We will find a closed subset M of X_+, such that it contains 0 and is positively invariant for T, and show that T is uniformly persistent for $M \setminus \{0\}$.

LEMMA 5.1. *There exists a nondecreasing and right-continuous function $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that $f(0) = 0, f(x) > 0 \; \forall x > 0$, $\lim_{x \to 0} f(x) = 0$, and the set $M := \{(\varphi, y) \in X_+ : y \leq f(\|\varphi\|)\}$ is positively invariant for T.*

Proof. We define $F : \mathbb{R}^2_+ \to \mathbb{R}^2_+$ by

$$F(x_1, x_2) = \left(ax_1, ax_2 + b \frac{x_1 + x_2}{1 + x_1 + x_2}\right) \; \forall x = (x_1, x_2) \in \mathbb{R}^2_+.$$

Then F is nondecreasing on $\mathbb{R}^2_+$. Set

$$\chi(t) = (ta + (1 - t), 1) \; \forall t \in [0, 1].$$

By induction, we define $\chi : \mathbb{R}_+ \to \mathbb{R}_+^2$ by

$$\chi(t) = F(\chi(t-1)) \ \forall t \in (n, n+1], \ \ \forall n \geq 1.$$

Note that $\chi(1)_1 = F\left(\chi(0)\right)_1$ and $a < 1$. Then the function $t \to \chi(t)_1$ is strictly decreasing and continuous. Since $F\left(1,1\right) \leq \left(a,1\right)$, the function $t \to \chi(t)_2$ is nonincreasing and left-continuous. Moreover, since $a+b<1$, we have $\lim_{t\to+\infty} \chi(t) = 0$. We further set

$$\chi(t) = (1-t, 1) \ \ \forall t \in (-\infty, 0]\,.$$

Since $\chi(t)_1$ is strictly decreasing in $t \in \mathbb{R}$, we can define

$$f(x) = \begin{cases} \chi(\chi(x)_1^{-1})_2 \text{ if } x > 0, \\ 0 \text{ if } x = 0. \end{cases}$$

It is easy to see that f has the desired properties.

Let $D := \left\{(x_1, x_2) \in \mathbb{R}_+^2 : x_2 \leq f(x_1)\right\}$. Since f is nondecreasing and right-continuous, it easily follows that D is closed. Now we show that $F(D) \subset D$. Let $x = (x_1, x_2) \in D$; then $x_2 \leq f\left(x_1\right)$. If $x_1 = 0$, there is nothing to prove because $F(0) = 0$. Assume that $x_1 > 0$; then there exists $t \in \mathbb{R}$ such that $\chi(t)_1 = x_1$, and hence, $x_2 \leq f(x_1) = \chi(t)_2$. Clearly, $x = (x_1, x_2) \leq \chi(t)$, and $F(x) \leq F(\chi(t))$. In the case where $t \geq 0$, we have

$$\chi(t+1)_1 = F\left(\chi(t)\right)_1 = F\left(x\right)_1,$$

and hence,

$$f(F(x)_1) = \chi(t+1)_2 = F\left(\chi(t)\right)_2 \geq F\left(x\right)_2,$$

which implies that $F(x) \in D$. In the case where $t \leq 0$, we have

$$x_1 \geq F\left(x\right)_1 = F\left(\chi(t)\right)_1 = a\chi(t)_1 = a\left(1-t\right) \geq a = \chi(1)_1,$$

and hence, there exists $s \in [t, 1]$ such that $\chi(s)_1 = F\left(x\right)_1$. It then follows that

$$f(F(x)_1) = \chi(s)_2 = 1 \geq F(\chi(t))_2 \geq F(x)_2,$$

which implies that $F(x) \in D$. This proves that $F(D) \subset D$.

Finally, we prove that $T(M) \subset M$. For any $(\varphi, y) \in M$, we have $(\|\varphi\|, y) \in D$, and hence, the positive invariance of D for F implies that $F(\|\varphi\|, y)_2 \leq f(F(\|\varphi\|, y)_1)$. Note that $\|T_1(\varphi, y)\| \geq a\|\varphi\| = F(\|\varphi\|, y)_1$ and $T_2(\varphi, y) = F(\|\varphi\|, y)_2$. By the monotonicity of f, it then follows that

$$T_2(\varphi, y) = F(\|\varphi\|, y)_2 \leq f(F(\|\varphi\|, y)_1) \leq f(\|T_1(\varphi, y)\|),$$

which implies that $T(\varphi, y) \in M$. Thus, M is positively invariant for T. ◻

Now we consider $T : M \to M$, where M is endowed with the usual distance $d(x, \widehat{x}) = \|x - \widehat{x}\|$. We set

$$\partial M_0 = \{0\}\,, \ M_0 = M \setminus \{0\}\,, \text{ and } \rho(x) = \|x\|\,.$$

Since T is the sum of a compact operator and a linear operator with norm being a, we have $\kappa\left(T(B)\right) \leq a\kappa\left(B\right)$ for any bounded set $B \subset M$. Thus, T is a κ-contraction. Moreover, for each $x \in M$, we have $\|T(x)\| \leq a\|x\| + b + c$, and hence

$$\|T^n(x)\| \leq a^n\|x\| + \left(\sum_{i=0}^{n-1} a^i\right)(b+c)\ \forall n \geq 1.$$

It then follows that $B = \{x \in M : \|x\| \leq \frac{b+c}{1-a}\}$ is positively invariant for T, and attracts every bounded subset of M for T. So $T : (M, d) \to (M, d)$ has a strong global attractor.

Let $\varepsilon > 0$ be fixed such that $a + \frac{c}{1+\varepsilon} > 1$. We claim that

$$\limsup_{n\to\infty} \|T^n x\| \geq \varepsilon\ \forall x = (\varphi, y) \in M_0.$$

Assume, by contradiction, that $\limsup_{n\to\infty} \|T^n x\| < \varepsilon$ for some $x = (\varphi, y) \in M_0$. We set $(\varphi_n, y_n) = T^n x\ \forall n \geq 0$. By the definition of M, we have $\varphi \in L^1_+\left((0, +\infty), \mathbb{R}\right) \backslash \{0\}$. It then follows that there exists $n_0 \geq 0$ such that $\int_0^1 \varphi_{n_0}(l)dl > 0$ and

$$\int_0^1 \varphi_{n+1}(l)dl \geq \left(a + \frac{c}{1+\varepsilon}\right)\int_0^1 \varphi_n(l)dl\ \forall n \geq n_0.$$

Thus, we obtain

$$\int_0^1 \varphi_n(l)dl \to +\infty \text{ as } n \to +\infty,$$

which is a contradiction. By Proposition 3.2, we conclude that T is ρ-uniformly persistent. Since $T : (M, d) \to (M, d)$ has a global attractor, it follows from Theorem 3.7 that $T : (M_0, d_0) \to (M_0, d_0)$ has a global attractor.

To avoid possible confusion, we denote by κ_0 the Kuratowski measure of non-compactness on the complete metric (M_0, d_0). We now consider $T : (M_0, d_0) \to (M_0, d_0)$. Let $\varepsilon > 0$ be fixed such that

$$\sqrt{a} < d := a + \frac{b}{1+\varepsilon} < 1.$$

Then for each $x \in M$, we have

$$\|T(x)\| \geq a\|x\| + b\frac{\|x\|}{1+\|x\|} \geq d\min(\varepsilon, \|x\|).$$

Let $B \subset M_0$ be a ρ-bounded set. We set $\rho_0 = \inf_{x\in B}\rho(x)$. Then for each $x \in B$, we obtain

$$\|T(x)\| \geq a\|x\| + b\frac{\|x\|}{1+\|x\|} \geq d\min(\varepsilon, \|x\|).$$

By induction, it follows that

$$\rho\left(T^n\left(x\right)\right) \geq d^n \min\left(\varepsilon, \rho_0\right)\ \forall n \geq 1, \forall x \in B.$$

Thus, for each $x, y \in B$, we have

$$\begin{aligned} d_0\left(T^n(x), T^n(y)\right) &= \left|\frac{1}{\rho\left(T^n\left(x\right)\right)} - \frac{1}{\rho\left(T^n\left(y\right)\right)}\right| + \left\|T^n\left(x\right) - T^n\left(y\right)\right\| \\ &\leq \left[\frac{1}{\rho\left(T^n\left(x\right)\right)\rho\left(T^n\left(y\right)\right)} + 1\right]\left\|T^n\left(x\right) - T^n\left(y\right)\right\| \\ &\leq \left[\frac{1}{d^{2n}\min\left(\varepsilon, \rho_0\right)^2} + 1\right] d(T^n\left(x\right), T^n\left(y\right)), \end{aligned}$$

and hence,

$$\begin{aligned} \kappa_0\left(T^n(B)\right) &\leq \left[\frac{1}{d^{2n}\min\left(\varepsilon, \rho_0\right)^2} + 1\right]\kappa\left(T^n(B)\right) \\ &\leq a^n\left[\frac{1}{d^{2n}\min\left(\varepsilon, \rho_0\right)^2} + 1\right]\kappa\left(B\right). \end{aligned}$$

Since $d > \sqrt{a}$, we obtain $\kappa_0\left(T^n(B)\right) \to 0$ as $n \to +\infty$. So $T : (M_0, d_0) \to (M_0, d_0)$ is κ_0-contracting.

It remains to show that $T : (M_0, d_0) \to (M_0, d_0)$ has no strong global attractor. Let $\delta > 0$ be fixed, and consider the ρ-strongly bounded set

$$B_\delta = \left\{x \in M : \rho\left(x\right) = \delta\right\}.$$

For each $m \geq 0$, we set $x^m := (\varphi^m, 0)$ with $\varphi^m = \delta 1_{[m,m+1]}$, and

$$x_n^m := (\varphi_n^m, y_n^m) = T^n\left(x^m\right) \quad \forall n \geq 0.$$

Then for each $m \geq 1$ and each $n \in \{0, \ldots, m-1\}$, we have $\int_0^1 \varphi_n^m(l)dl = 0$, and hence,

$$\begin{cases} \varphi_{n+1}^m(\cdot) = a\varphi_n^m(\cdot + 1) + a\int_0^1 \varphi_n^m(l)dl 1_{[0,1]}(\cdot), \\ y_{n+1}^m = ay_n^m + b\frac{\|x_n^m\|}{1+\|x_n^m\|}. \end{cases}$$

Thus, for each $m \geq 1$ and each $n \in \{0, \ldots, m-1\}$, we obtain

$$\left\|x_{n+1}^m\right\| \leq (a+b)\left\|x_n^m\right\| \leq (a+b)^n\,\delta.$$

It follows that $\inf_{x \in B_\delta} \rho\left(T^n\left(x\right)\right) \to 0$ as $n \to +\infty$. So the κ_0-contracting map $T : (M_0, d_0) \to (M_0, d_0)$ has a global attractor, but no strong global attractor.

5.3. -contracting semiflows on ($_0$ $_0$). In this subsection, we construct continuous-time κ-contracting semiflows on (M_0, d_0) such that they admit a global attractor, but no strong global attractor.

Let X and X_+ be defined as in the previous subsection. Consider the following age-structured model:

$$\begin{cases} \frac{\partial u}{\partial t} + \frac{\partial u}{\partial a} = -\mu u(t,a), \ t \geq 0, a \in (0, \infty), \\ u(t,0) = \frac{\int_0^{+\infty}\beta(a)u(t,a)da}{1+\|(u(t),y(t))\|}, \\ \frac{dy(t)}{dt} = -\mu y(t) + \gamma\frac{\|(u(t),y(t))\|}{1+\|(u(t),y(t))\|}, \\ u(0,.) = u_0 \in L_+^1\left((0,+\infty), \mathbb{R}\right), \ y(0) = y_0 \in \mathbb{R}_+. \end{cases} \tag{4}$$

We assume that

(A4) $\mu > 0$, $\gamma \in (\frac{\mu}{2}, \mu)$, $\beta : \mathbb{R}_+ \to \mathbb{R}_+$ is uniformly continuous, bounded, $\int_0^\infty \beta(a)e^{-\mu a}da > 1$, and there exists a sequence of real numbers $\{a_n\}_{n\geq 0} \subset [0, +\infty)$ such that $a_n < a_{n+1}$ $\forall n \geq 0$, $\lim_{n\to+\infty}(a_{2n+2} - a_{2n+1}) = +\infty$, and

$$\beta(a) > 0 \Leftrightarrow a \in \bigcup_{n\geq 0} (a_{2n}, a_{2n+1}).$$

For each $\chi \in L^\infty((0,+\infty), \mathbb{R})$ and each $\varphi \in L^1((0,+\infty), \mathbb{R})$, we define

$$\mathcal{F}_\chi(\varphi) = \int_0^{+\infty} \chi(s)\varphi(s)ds.$$

Let $\{U(t)\}_{t\geq 0}$ be the solution semiflow on X_+ generated by system (4), and let $(u(t), y(t)) = U(t)(u_0, y_0)$. Then we have the following Volterra formulation of system (4):

$$u(t,a) = \begin{cases} e^{-\mu t}u_0(a-t) \text{ if } a > t, \\ e^{-\mu a}B(t-a) \text{ if } a \leq t, \end{cases}$$

with $B(t) = \frac{\mathcal{F}_\beta(u(t))}{1+\mathcal{F}_1(u(t))+y(t)}$, and for each $t \geq 0$,

$$\begin{cases} \frac{d\mathcal{F}_1(u(t))}{dt} = -\mu\mathcal{F}_1(u(t)) + \frac{\mathcal{F}_\beta(u(t))}{1+\mathcal{F}_1(u(t))+y(t)}, \\ \frac{dy(t)}{dt} = -\mu y(t) + \gamma\frac{\mathcal{F}_1(u(t))+y(t)}{1+\mathcal{F}_1(u(t))+y(t)}, \end{cases} \tag{5}$$

and

$$\mathcal{F}_\beta(u(t)) = e^{-\mu t}\int_t^{+\infty} \beta(s)u_0(s-t)ds + \int_0^t \beta(s)e^{-\mu a}\frac{\mathcal{F}_\beta(u(t-a))}{1+\mathcal{F}_1(u(t-a))+y(t-a)}da.$$

LEMMA 5.2. *There exists a continuous and nondecreasing function $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that $f(0) = 0$, $f(x) > 0$ $\forall x > 0$, and the set*

$$M := \{(\varphi, y) \in X_+ : y \leq f(\|\varphi\|)\}$$

is positively invariant for $\{U(t)\}_{t\geq 0}$.

Proof. Let $(\widehat{x}(t), \widehat{y}(t))$ be the unique solution on $[0,\infty)$ of the following cooperative system:

$$\begin{cases} \frac{dx(t)}{dt} = -\mu x(t), \\ \frac{dy(t)}{dt} = -\mu y(t) + \gamma\frac{x(t)+y(t)}{1+x(t)+y(t)} \end{cases} \tag{6}$$

with

$$(\widehat{x}(0), \widehat{y}(0)) = \left(\frac{\gamma}{\mu} + 1, \frac{\gamma}{\mu} + 1\right).$$

Since $\widehat{x}'(0) < 0$ and $\widehat{y}'(0) < 0$, $(\widehat{x}(t), \widehat{y}(t))$ is nonincreasing on some small interval $[0, \epsilon]$. By the monotonicity of the solution semiflow of system (6) on $\mathbb{R}^2_+$, it follows that $(\widehat{x}(t), \widehat{y}(t))$ is nonincreasing on $[0,\infty)$, and $(\widehat{x}(t), \widehat{y}(t)) \to (0,0)$ as $t \to +\infty$. Set

$$\widehat{x}(t) = \frac{\gamma}{\mu} + 1 - t, \text{ and } \widehat{y}(t) = \frac{\gamma}{\mu} + 1 \quad \forall t \in (-\infty, 0].$$

Clearly, $\widehat{x}(t)$ is strictly decreasing in $t \in \mathbb{R}$. Define $f : \mathbb{R}_+ \to \mathbb{R}_+$ by

$$f(\alpha) = \begin{cases} \widehat{y}(\widehat{x}^{-1}(\alpha)) \text{ if } \alpha > 0, \\ 0 \text{ if } \alpha = 0. \end{cases}$$

Then f satisfies the desired properties. Note that the set $D := \{(x, y) \in \mathbb{R}_+^2 : y \leq f(x)\}$ is positively invariant for the solution semiflow of (6). By using the monotonicity of f and the planar vector field associated with (5), one can easily prove that $U(t)M \subset M$ $\forall t \geq 0$. $\square$

Now we consider $U(t) : (M, d) \to (M, d)$, where $d(x, \widehat{x}) = \|x - \widehat{x}\|$. Set

$$\partial M_0 = \{0\}, \ M_0 = M \setminus \{0\}, \text{ and } \rho(x) = \|x\|.$$

Since $\beta : \mathbb{R}_+ \to \mathbb{R}_+$ is uniformly continuous, it follows from [28] that for any bounded set $B \subset M$, we have

$$\kappa(U(t)B) \leq e^{-\mu t} \kappa(B) \ \ \forall t \geq 0.$$

Let $z(t) := \mathcal{F}_1(u(t)) + y(t)$. Then we obtain

$$\frac{dz(t)}{dt} \leq -\mu z(t) + (\|\beta\|_\infty + \gamma) \ \ \forall t \geq 0.$$

Consequently, $U(t) : (M, d) \to (M, d)$ has a strong global attractor.

Let $\varepsilon > 0$ be such that

$$\frac{\int_0^{+\infty} \beta(a) e^{-\mu a} da}{1 + \varepsilon} > 1.$$

We claim that $\limsup_{t\to\infty} \|U(t)x\| \geq \varepsilon$ $\forall x \in M_0$. Assume, by contradiction, that $\limsup_{t\to\infty} \|U(t)x\| < \varepsilon$ for some $x = (u_0, y_0) \in M_0$. Then there exists $t_0 \geq 0$ such that $\|U(t + t_0)x\| < \varepsilon$ $\forall t \geq 0$. By the definition of M, we have $u_0 \neq 0$, and hence, $u(t) \neq 0$ $\forall t \geq 0$. It follows that $u(t + t_0) \geq \widehat{T}(t) u(t_0)$ $\forall t \geq 0$, where $\{\widehat{T}(t)\}_{t\geq 0}$ is the strongly continuous semigroup of bounded linear operators on $L^1((0, +\infty), \mathbb{R})$, which is generated by $\widehat{A}\varphi = -\varphi' - \mu\varphi$ with

$$D(\widehat{A}) = \left\{ \varphi \in W^{1,1}((0, +\infty), \mathbb{R}) : \varphi(0) = \frac{\int_0^{+\infty} \beta(a)\varphi(a) da}{1 + \varepsilon} \right\}.$$

Since $u(t_0) \neq 0$, it follows from [28] that

$$\|u(t + t_0)\|_{L^1} \geq \left\| \widehat{T}(t) u(t_0) \right\|_{L^1} \to +\infty \text{ as } t \to +\infty,$$

which is a contradiction. By the continuous-time version of Proposition 3.2, we deduce that $U(t) : (M, d) \to (M, d)$ is ρ-uniformly persistent, and hence, $U(t) : (M_0, d_0) \to (M_0, d_0)$ has a global attractor (see Theorem 3.7 and Remark 3.10).

We now prove that $U(t) : (M_0, d_0) \to (M_0, d_0)$ is κ_0-contracting. Let $\varepsilon > 0$ be such that $\mu - \frac{2\gamma}{1+\varepsilon} < 0$. Let B be a ρ-strongly bounded set of M_0. We set $\rho_0 = \inf_{x\in B} \rho(x)$. For each $x \in B$, if we set $z(t) = \rho(U(t)x)$ $\forall t \geq 0$, we then have

$$\frac{dz(t)}{dt} \geq -\mu z(t) + \gamma \frac{z(t)}{1 + z(t)} \ \ \forall t \geq 0,$$

and hence,

$$\rho\left(U(t)x\right) \geq e^{\left(-\mu+\frac{\gamma}{1+\varepsilon}\right)t} \min\left(\varepsilon, \rho_0\right) \ \forall t \geq 0.$$

It then follows that for each $x, y \in B$, we have

$$d_0\left(U(t)x, U(t)y\right) \leq \left[\frac{1}{e^{2\left(-\mu+\frac{\gamma}{1+\varepsilon}\right)t} \min\left(\varepsilon, \rho_0\right)^2} + 1\right] d\left(U(t)x, U(t)y\right),$$

and hence,

$$\begin{aligned}\kappa_0\left(U(t)B\right) &\leq \left[\frac{1}{e^{2\left(-\mu+\frac{\gamma}{1+\varepsilon}\right)t} \min\left(\varepsilon, \rho_0\right)^2} + 1\right] \kappa\left(U(t)B\right) \\ &\leq e^{-\mu t}\left[\frac{1}{e^{2\left(-\mu+\frac{\gamma}{1+\varepsilon}\right)t} \min\left(\varepsilon, \rho_0\right)^2} + 1\right] \kappa\left(B\right).\end{aligned}$$

Since $\mu - \frac{2\gamma}{1+\varepsilon} < 0$, we deduce that $\kappa_0\left(U(t)B\right) \to 0$ as $t \to +\infty$. So $U(t) : (M_0, d_0) \to (M_0, d_0)$ is κ_0-contracting.

It remains to show that $U(t) : (M_0, d_0) \to (M_0, d_0)$ has no strong global attractor. We fix a real number $\delta > 0$ and set

$$B := \left\{x \in M : \rho\left(x\right) = \delta\right\}.$$

Let $x^n = \left(u_0^n, 0\right)$, with $u_0^n = \delta 1_{[a_{2n+1}, a_{2n+1}+1]}(\cdot)$, and $(u^n(t), y^n(t)) = U(t)x^n \ \forall t \geq 0$. Then for each $t \geq 0$, we have

$$\begin{aligned}\mathcal{F}_\beta\left(u^n(t)\right) &= e^{-\mu t}\int_t^{+\infty} \beta(s) u_0^n(s-t) ds \\ &+ \int_0^t \beta(s) e^{-\mu a} \frac{\mathcal{F}_\beta\left(u^n(t-a)\right)}{1+\mathcal{F}_1\left(u^n(t-a)\right) + y^n(t-a)} da\end{aligned} \tag{7}$$

and

$$\begin{aligned}\int_t^{+\infty} \beta(s) u_0^n(s-t) ds &= \int_0^{+\infty} \beta(s+t) u_0^n(s) ds \\ &= \delta \int_{a_{2n+1}}^{a_{2n+1}+1} \beta(s+t) ds = \delta \int_{t+a_{2n+1}}^{t+a_{2n+1}+1} \beta(s) ds.\end{aligned}$$

Since $a_{2n+2} - a_{2n+1} \to +\infty$ as $n \to +\infty$, there exists $n_0 \geq 0$ such that $a_{2n+2} - a_{2n+1} > 1 \ \forall n \geq n_0$. Then we have

$$\int_t^{+\infty} \beta(s) u_0^n(s-t) ds = 0 \ \forall t \in [0, a_{2n+2} - (a_{2n+1}+1)], \ \forall n \geq n_0.$$

Since $\mathcal{F}_\beta\left(u^n(t)\right)$ is a solution of (7), we deduce that for each $n \geq n_0$, and $t \in [0, a_{2n+2} - (a_{2n+1}+1)]$, $\mathcal{F}_\beta\left(u^n(t)\right) = 0$. It then follows that $z_n(t) := \|U(t)x^n\|$ satisfies $z_n(0) = \delta$ and

$$\frac{dz_n(t)}{dt} = -\mu z_n(t) + \gamma \frac{z_n(t)}{1+z_n(t)} \ \forall t \in [0, a_{2n+2} - (a_{2n+1}+1)], \ \forall n \geq n_0.$$

Thus, we have

$$z_n(t) \leq e^{(-\mu+\gamma)t}\delta \ \forall t \in [0, a_{2n+2} - (a_{2n+1}+1)],$$

which implies that $\inf_{x\in B} \rho\left(U(t)x\right) \to 0$, as $t \to +\infty$. So $U(t) : (M_0, d_0) \to (M_0, d_0)$ has no strong global attractor.

5.4. A periodic age-structured model. In this subsection, we illustrate applicability of Theorem 4.5 in the case of convex κ-contracting maps.

Consider the 1-periodic nonautonomous age-structured model

$$(8)\quad \begin{cases} \frac{\partial u}{\partial t} + \frac{\partial u}{\partial a} = -\left(\mu + m(t, \int_0^{+\infty} u(t,l)dl)(a)\right) u(t,a),\ t \geq 0,\ a \in (0,+\infty), \\ u(t,0) = \frac{\int_0^{+\infty} \beta(t,a)u(t,a)da}{1+\int_0^{+\infty} u(t,a)da}, \\ u(0,.) = u_0 \in L^1_+ ((0,+\infty), \mathbb{R}). \end{cases}$$

We assume that

(A5) $\mu > 0$ and the following conditions are satisfied:

(a) $\beta : \mathbb{R}^2_+ \to \mathbb{R}_+$ is uniformly continuous, positive, bounded, and $t \to \beta(t,a)$ is 1-periodic.

(b) $m \in C\left(\mathbb{R}^2_+, L^\infty_+ ((0,+\infty), \mathbb{R})\right)$ and the map $t \to m(t,\cdot)$ is 1-periodic.

(c) There exist a bounded and uniformly continuous map $\widehat{\beta} : \mathbb{R}_+ \to \mathbb{R}_+$ and a continuous and bounded map $\widehat{m} : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\beta(t,\cdot) \geq \widehat{\beta}(\cdot) \text{ and } m(t,\cdot) \leq \widehat{m}(\cdot)\ \forall t \in [0,1],$$

and for any $a \geq 0$, there exists $r \geq a$ such that $\widehat{\beta}(r) > 0$ and

$$\int_0^{+\infty} \widehat{\beta}(a) e^{-\int_0^a \mu + \widehat{m}(r)dr} da > 1.$$

Let $Y = L^1_+ ((0,+\infty), \mathbb{R})$ and $Y_+ = L^1_+ ((0,+\infty), \mathbb{R})$, and let $\{U(t,s)\}_{0 \leq s \leq t}$ be the nonautonomous semiflow generated by system (8). Set

$$M = Y_+,\ \partial M_0 = \{0\} \text{ and } M_0 = Y_+ \setminus \{0\}.$$

Then $U(t,s)0 = 0$, and $U(t,s)M_0 \subset M_0\ \forall t \geq s \geq 0$. For a 1-periodic solution of system (8) in $Y_+ \setminus \{0\}$, it suffices to find a fixed point of $T = U(1,0)$. By setting $x(t) := \mathcal{F}_1 (U(t,s)x)$, we have

$$\frac{dx(t)}{dt} \leq -\mu x(t) + \|\beta\|_\infty \frac{x(t)}{1+x(t)},$$

which implies that T is bounded dissipative on M. Moreover, by using the results in [28] and assumptions (A5)(a),(b), we obtain

$$U(t,s) = C(t,s) + N(t,s),$$

where $C(t,s)$ is a compact operator, and

$$\|N(t,s)x\| \leq e^{-\mu(t-s)} \|x\|\ \forall t \geq s \geq 0, \forall x \in M.$$

Thus, T is κ-contracting in the sense that $\kappa(T^n(B)) \to 0$ as $n \to +\infty$ for any bounded set $B \subset M$. It follows from Theorem 2.9 that T has a strong global attractor in M. Using assumption (A5)(c) and comparison arguments, we can further prove that the fixed point 0 of T is ejective. In order to apply Theorem 4.5, we need to verify that T is convex κ-contracting.

Let $V(t,s) = (V_1(t,s), V_2(t,s))$ be the nonautonomous semiflow on $Y_+ \times Y_+$, which is generated by the following system:

$$\begin{cases} \frac{\partial u_1}{\partial t} + \frac{\partial u_1}{\partial a} = -\left(\mu + m(t, \int_0^{+\infty} (u_1 + u_2)(t,l)dl)(a)\right) u_1(t,a), \ a \in (0,+\infty), \\ u_1(t,0) = 0, \\ \frac{\partial u_2}{\partial t} + \frac{\partial u_2}{\partial a} = -\left(\mu + m(t, \int_0^{+\infty} (u_1 + u_2)(t,l)dl)(a)\right) u_2(t,a), \ a \in (0,+\infty), \\ u_2(t,0) = \frac{\int_0^{+\infty} \beta(t,a)(u_1+u_2)(t,a)da}{1+\alpha \int_0^{+\infty} (u_1+u_2)(t,a)da}, \\ (u_1(0,.), u_2(0,.)) = \left(u_0^1, u_0^2\right) \in L_+^1\left((0,+\infty), \mathbb{R}^2\right). \end{cases}$$

We define $P_n : Y \to Y$ by

$$P_n(\varphi) = \varphi 1_{[n,+\infty)} \ \forall n \geq 0.$$

Then for each $n \geq 0$, we have

$$P_{n+1}T(x) = V_1(1,0)(P_n x, (I - P_n)x)$$

and

$$(I - P_{n+1})T(x) = V_2(1,0)(P_n x, (I - P_n)x).$$

Moreover, if B is bounded and $(I - P_n)(B)$ is relatively compact, then

$$\{(I - P_{n+1})T(x) : x \in B\} = \{V_2(1,0)(P_n x, (I - P_n)x) : x \in B\}$$

is relatively compact. Note that for each $x \in M$, we have

$$\|P_{n+1}T(x)\| = \|V_1(1,0)(P_n x, (I - P_n)x)\| \leq e^{-\mu} \|P_n x\|.$$

By Lemma 4.8, it follows that T is convex κ -contracting. Thus, Theorem 4.5 implies that T has a fixed point in M_0, and hence, system (8) admits a nontrivial 1-periodic solution.

Finally, we remark that the similar approach can be applied to more general age-structured models.

Acknowledgments. We would like to express our gratitude to three anonymous referees for their helpful comments and suggestions which led to an important improvement of our original manuscript.

REFERENCES

[1] S. Anita, *Analysis and Control of Age-dependent Population Dynamics*, Math. Model. Theory Appl., 11, Kluwer Academic Publishers, Dordrecht, 2000.
[2] F. Browder, *A further generalization of the Schauder fixed point theorem*, Duke Math. J., 32 (1965), pp. 575–578.
[3] J. W. Cholewa and T. Dlotko, *Global Attractors in Abstract Parabolic Problems*, Cambridge University Press, Cambridge, UK, 2000.
[4] J. W. Cholewa and J. K. Hale, *Some counterexamples in dissipative systems*, Dyn. Contin. Discrete Impuls. Syst., 7 (2000), pp. 159–176.
[5] G. Cooperman, *α-condensing Maps and Dissipative Systems*, Ph.D. Thesis, Brown University, Providence, RI, 1978.
[6] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, New York, 1985.
[7] J. K. Hale, *Asymptotic Behavior of Dissipative Systems*, Math. Surveys Monogr. 25, AMS, Providence, RI, 1988.

[8] J. K. Hale, *Dissipation and attractors*, in Proceedings of the International Conference on Differential Equations (Berlin, 1999), 2 vols., B. Fiedler, K. Groeger, and J. Sprekels, eds., World Scientific, River Edge, NJ, 2000, pp. 622–637.

[9] J. K. Hale and O. Lopes, *Fixed point theorems and dissipative processes*, J. Differential Equations, 13 (1973), pp. 391–402.

[10] J. K. Hale and P. Waltman, *Persistence in infinite-dimensional systems*, SIAM J. Math. Anal., 20 (1989), pp. 388–395.

[11] M. W. Hirsch, H. L. Smith, and X.-Q. Zhao, *Chain transitivity, attractivity, and strong repellors for semidynamical systems*, J. Dynam. Differential Equations, 13 (2001), pp. 107–131.

[12] W. A. Horn, *Some fixed point theorems for compact mappings and flows on a Banach space*, Trans. Amer. Math. Soc., 149 (1970), pp. 391–404.

[13] V. Hutson and K. Schmitt, *Permanence and the dynamics of biological systems*, Math. Biosci., 111 (1992), pp. 293–326.

[14] M. Iannelli, *Mathematical Theory of Age-Structured Population Dynamics*, Giardini Editori e stampatori, Pisa, 1994.

[15] P. Magal and O. Arino, *Existence of periodic solutions for a state dependent delay differential equation*, J. Differential Equations, 165 (2000), pp. 61–95.

[16] P. Magal and H. R. Thieme, *Eventual compactness for a semiflow generated by an age-structured models*, Comm. Pure Appl. Anal., 3 (2004), pp. 695–727.

[17] R. H. Martin, *Nonlinear Operators and Differential Equations in Banach Spaces*, John Wiley and Sons, New York, 1976.

[18] R. D. Nussbaum, *Some asymptotic fixed point theorems*, Trans. Amer. Math. Soc., 171 (1972), pp. 349–375.

[19] R. D. Nussbaum, *Periodic solutions of some nonlinear autonomous functional differential equations*, Ann. Mat. Pura Appl., 101 (1974), pp. 263–306.

[20] G. Raugel, *Global Attractors in Partial Differential Equations*, Handbook of Dynamical Systems 2, North-Holland, Amsterdam, 2002.

[21] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, 1991.

[22] G. R. Sell and Y. You, *Dynamics of Evolutionary Equations*, Springer-Verlag, New York, 2002.

[23] H. L. Smith, and X.-Q. Zhao, *Robust persistence for semidynamical systems*, Nonlinear Anal., 47 (2001), pp. 6169–6179.

[24] R. Temam, *Infinite Dimensional Dynamical Systems in Mechanics and Physics*, Springer-Verlag, New York, 1988.

[25] H. R. Thieme, *Semiflows generated by Lipschitz perturbations of non-densely defined operators*, Differential Integral Equations, 3 (1990), pp. 1035–1066.

[26] H. R. Thieme, *Persistence under relaxed point-dissipativity (with application to an endemic model)*, SIAM J. Math. Anal., 24 (1993), pp. 407–435.

[27] H. R. Thieme, *Uniform persistence and permanence for non-autonomous semiflows in population biology*, Math. Biosci., 166 (2000), pp. 173–201.

[28] G. F. Webb, *Theory of Nonlinear Age-Dependent Population Dynamics*, Marcel Dekker, New York, 1985.

[29] X.-Q. Zhao, *Uniform persistence and periodic coexistence states in infinite-dimensional periodic semiflows with applications*, Canad. Appl. Math. Quart., 3 (1995), pp. 473–495.

[30] X.-Q. Zhao, *Dynamical Systems in Population Biology*, Springer-Verlag, New York, 2003.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 276–290

ASYMPTOTIC BEHAVIOR OF THE ELECTROMAGNETIC FIELD FOR A MICROMAGNETISM EQUATION WITHOUT EXCHANGE ENERGY*

FRANK JOCHMANN†

Abstract. This paper deals with the transient Landau–Lifschitz equations describing ferromagnetic media without exchange interaction coupled with Maxwell's equations. The asymptotic behavior of the solutions to this system is investigated.

Key words. asymptotic behavior of solutions, ferromagnetism, Maxwell's equations, existence of solutions

AMS subject classifications. 35Q60, 35L25, 78A40

DOI. 10.1137/S0036141004443324

1. Introduction. This paper is concerned with a mathematical model for micromagnetism consisting of Maxwell's equations

$$\varepsilon\partial_t\mathbf{E} = \text{curl } \mathbf{H} - \sigma\mathbf{E} - \mathbf{j}, \quad \mu\partial_t\mathbf{H} = -\text{ curl } \mathbf{E} - \mu\partial_t\tilde{\mathbf{M}}, \tag{1.1}$$

coupled with the equation

$$\partial_t\mathbf{M} = F(x, \mathbf{M}) \cdot [\mathbf{H} - A(x)\mathbf{M}] \tag{1.2}$$

for the unknown electromagnetic field quantities $\mathbf{E}, \mathbf{H}$ depending on the time $t \geq 0$ and the space variable $x \in \mathbb{R}^3$. The magnetization $\mathbf{M}$ defined on $\mathbb{R}^+ \times G$ is also an unknown function which solves the ordinary differential equation (1.2); see [1], [6], [5], [8]. Here the set $G \subset \mathbb{R}^3$ represents the ferromagnetic medium. In (1.1) the function $\tilde{\mathbf{M}}$ is the extension of $\mathbf{M}$ on $\mathbb{R}^+ \times \mathbb{R}^3$ defined by zero on the set $\mathbb{R}^+ \times (\mathbb{R}^3 \setminus G)$. This system is supplemented by the initial conditions

$$\mathbf{E}(0, x) = \mathbf{E}_0(x), \quad \mathbf{H}(0, x) = \mathbf{H}_0(x) \tag{1.3}$$

and

$$\mathbf{M}(0, x) = \mathbf{M}_0(x) \text{ on } G. \tag{1.4}$$

From the physical point of view it is reasonable to assume that the initial state for the magnetic induction $\mathbf{B}_0 \stackrel{\text{def}}{=} \mu(\mathbf{H}_0 + \tilde{\mathbf{M}}_0)$ is divergence free.

In (1.2) the term $A(x)\mathbf{M}$ describes a possible anisotropy of the medium, where $A(x)$ is a symmetric positive semidefinite matrix. The precise assumptions on the nonlinear function $F : G \times \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$ will be given in the next section. A physically relevant example for F is

$$F(x, \mathbf{m})\mathbf{h} = -\gamma\mathbf{m} \wedge \mathbf{h} - \alpha|\mathbf{m}|^{-1}\mathbf{m} \wedge (\mathbf{m} \wedge \mathbf{h}), \tag{1.5}$$

including a damping term $\alpha\mathbf{m} \wedge (\mathbf{m} \wedge \mathbf{h})$ with $\alpha \geq 0$.

*Received by the editors April 26, 2004; accepted for publication (in revised form) November 16, 2004; published electronically September 20, 2005.
http://www.siam.org/journals/sima/37-1/44332.html

†Technische Universität Berlin, Fakultät II–Mathematik und Naturwissenschaften, Institut für Mathematik, Strasse des 17. Juni 136, 10623 Berlin, Germany (jochmann@math.tu-berlin.de).

The main topic of this paper is the investigation of the large time asymptotic behavior of the solutions. If F is given by (1.5), one obtains the energy dissipation law

$$\frac{1}{2}\frac{d}{dt}\left(\int_{\mathbb{R}^3}(\varepsilon|\mathbf{E}|^2+\mu|\mathbf{H}|^2)dx+\int_G \mu\mathbf{M}\cdot A\mathbf{M}dx\right) \tag{1.6}$$

$$= -\int_{\mathbb{R}^3}\mathbf{E}\mathbf{j}dx-\int_{\mathbb{R}^3}\sigma|\mathbf{E}|^2dx-\int_G \mu\alpha|\mathbf{M}|^{-1}|\mathbf{M}\wedge[\mathbf{H}-A\mathbf{M}]|^2dx$$

for the system (1.1)–(1.4) including dissipative terms arising from the electrical conductivity σ and the damping for $\partial_t\mathbf{M}$ on the set where $\alpha > 0$. In the autonomous case where $\mathbf{j} = 0$, the energy occurring in (1.6) is a nonincreasing function of the time t. However, no information about the asymptotic behavior of $\mathbf{M}$ and the magnetic field $\mathbf{H}$ (even on sets of nonvanishing electrical conductivity) can be obtained directly from this dissipation law, since there is no direct damping for any component of $\mathbf{H}$. In this paper the damping coefficient α may vanish on some part of the set G, which means that on this subset there is even no damping for $\partial_t\mathbf{M}$.

The main goal of section 3 is to show that the solenoidal part of the magnetic field decays in the weak topology and that the electric field converges for $t \to \infty$ weakly to some asymptotic state which is determined by the prescribed initial data $\mathbf{E}_0, \mathbf{H}_0, \mathbf{M}_0$ and the external current $\mathbf{j}$ (Theorem 2.2).

In section 4 it is shown that this convergence of the electromagnetic field is strong with respect to the energy norm for $t \to \infty$ on bounded sets of nonvanishing electrical conductivity (Theorem 2.4). Here, the main idea is to prove local energy estimates using a vector potential which provides additional dissipative terms for the magnetic field. Since the system (1.1)–(1.4) does not admit strong solutions with bounded derivatives for $t \to \infty$ in general, this strong convergence cannot be obtained from standard embedding results.

The existence of solutions to (1.1)–(1.4) has been proved in [6] for constant coefficients ε, μ and in [5] for nonsmooth coefficients ε, μ. In the spatially one-dimensional case the local decay of the transverse components is shown in [7]. Uniqueness of solutions to (1.1)–(1.4) is still an open question, at least for nonsmooth coefficients as considered here. The quasi-stationary limit for this system is studied in [5], in which the size of the ferromagnetic medium is very small in comparison to the electromagnetic wave length. It is shown in [5] that in this case the magnetic field is governed by the equations of the magnetostatic approximation.

Existence and quasi-stationary limit for the Landau–Lifschitz equation for the magnetic moment with exchange interaction coupled with Maxwell's equations are studied in [2] and [3]; see also [11]. Furthermore it is shown in [2] and [3] that all points of the weak ω-limit set are solutions of the corresponding stationary equations by using the $H^1(G)$ estimate coming from the second order exchange-energy term which does not occur in the system (1.1)–(1.4).

2. Definitions, assumptions, and statement of main results. All assumptions stated in this section shall be fulfilled throughout this paper. Let $G \subset \mathbb{R}^3$ be an open subset of finite measure. The dielectric and magnetic susceptibilities $\varepsilon, \mu \in L^\infty(\mathbb{R}^3)$ are assumed to have a positive lower bound, which means that

$$\varepsilon(x), \mu(x) \geq a_0 \text{ on } \mathbb{R}^3 \text{ with some } a_0 > 0. \tag{2.1}$$

The assumptions on the initial data $(\mathbf{E}_0, \mathbf{H}_0) \in X$ and $\mathbf{j}$ are

$$\mathbf{j} \in L^1((0,\infty), L^2(\mathbb{R}^3)) \cap L^\infty((0,\infty), L^2(\mathbb{R}^3)), \quad \mathbf{E}_0 \in L^2(\mathbb{R}^3), \quad \mathbf{H}_0 \in L^2(\mathbb{R}^3), \tag{2.2}$$

and

$$\mathbf{M}_0 \in L^\infty(G) \text{ with div } [\mu(\mathbf{H}_0 + \tilde{\mathbf{M}}_0)] = 0 \text{ on } \mathbb{R}^3. \tag{2.3}$$

Next, $A \in L^\infty(G, \mathbb{R}^{3\times 3})$ is assumed to be a positive semidefinite matrix-valued function, which means that $A(x) \in \mathbb{R}^{3\times 3}$ is symmetric and

$$y \cdot A(x)y \geq 0 \text{ for all } x \in G,\, y \in \mathbb{R}^3. \tag{2.4}$$

It is assumed that the nonlinear function $F : G \times \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$ satisfies

$$\mathbf{m} \cdot F(x, \mathbf{m})\mathbf{h} = 0 \text{ for all } x \in G,\, \mathbf{h} \in \mathbb{R}^3, \text{ and } \mathbf{m} \in \mathbb{R}^3, \tag{2.5}$$

and that there is some nonnegative function $\beta \in L^\infty(G)$ such that

$$\mathbf{h} \cdot F(x, \mathbf{m})\mathbf{h} \geq \beta(x)|F(x, \mathbf{m})\mathbf{h}|^2 \tag{2.6}$$

for all $x \in G$, $\mathbf{h} \in \mathbb{R}^3$, and $\mathbf{m} \in \mathbb{R}^3$ with $|\mathbf{m}| \leq \|\mathbf{M}_0\|_{L^\infty(G)}$, where $\mathbf{M}_0$ is as in (2.2). Furthermore, F is assumed to be locally Lipschitz-continuous with respect to $\mathbf{M}$; i.e., for $k \in (0,\infty)$ there exists $L_k \in (0,\infty)$ such that

$$|F(x, y) - F(x, \tilde{y})| \leq L_k|y - \tilde{y}| \tag{2.7}$$

for all $x \in G$, $y \in \mathbb{R}^3$, and $\tilde{y} \in \mathbb{R}^3$ with $|y| + |\tilde{y}| \leq k$. Finally,

$$F(\cdot, 0) \in L^\infty(G). \tag{2.8}$$

The conductivity satisfies

$$\sigma \in L^\infty(\mathbb{R}^3) \quad \text{and} \quad \sigma \geq 0. \tag{2.9}$$

Let

$$G_0 \stackrel{\text{def}}{=} \{x \in G : \beta(x) = 0\} \tag{2.10}$$

with the nonnegative function $\beta \in L^\infty(G)$ occurring in assumption (2.6), and let

$$S_\sigma \stackrel{\text{def}}{=} \{x \in \mathbb{R}^3 : \sigma(x) > 0\} \tag{2.11}$$

be the conducting region.

It is assumed that there is a nonempty open set $\mathcal{U} \subset S_\sigma$ such that

$$\overline{G_0} \subset \mathcal{U} \subset S_\sigma. \tag{2.12}$$

Roughly speaking this condition requires that the medium is conducting on a neighborhood of the region of undamped magnetization contained in G_0. Furthermore,

$$\varepsilon(x) = \mu(x) = 1 \text{ on } \mathbb{R}^3 \setminus \mathcal{U}. \tag{2.13}$$

The physical meaning of this condition is that the set $\mathbb{R}^3 \setminus S_\sigma \subset \mathbb{R}^3 \setminus \mathcal{U}$ represents the vacuum region.

Remark 1. If F is given by (1.5) with some nonnegative function $\alpha \in L^\infty(G)$, one obtains

$$\mathbf{h} \cdot F(x, \mathbf{m})\mathbf{h} = -\alpha(x)|\mathbf{m}|^{-1}\mathbf{h} \cdot \mathbf{m} \wedge (\mathbf{m} \wedge \mathbf{h}) = \alpha(x)|\mathbf{m}|^{-1}|\mathbf{m} \wedge \mathbf{h}|^2$$

$$\geq \alpha(x)\|\mathbf{M}_0\|^{-1}_{L^\infty(G)}|\mathbf{m} \wedge \mathbf{h}|^2$$

and

$$|F(x, \mathbf{m})\mathbf{h}|^2 \leq 2\|\gamma\|^2_{L^\infty(G)}|\mathbf{m} \wedge \mathbf{h}|^2 + 2\|\alpha\|^2_{L^\infty(G)}|\mathbf{m} \wedge \mathbf{h}|^2$$

for all $x \in G$, $\mathbf{h} \in \mathbb{R}^3$, and $\mathbf{m} \in \mathbb{R}^3$ with $|\mathbf{m}| \leq \|\mathbf{M}_0\|_{L^\infty(G)}$, whence assumption (2.6). Furthermore, the set G_0 occurring in (2.10) can be chosen as $G_0 = \{x \in G : \alpha(x) = 0\}$ in this case.

Next, let H_{curl} be the space of all $\mathbf{e} \in L^2(\mathbb{R}^3, \mathbb{R}^3)$ with $\operatorname{curl} \mathbf{e} \in L^2(\mathbb{R}^3)$. As in [5] let $1 - P_E$ and $1 - P_H$ be the orthogonal projectors on $H_{curl,0} \stackrel{\text{def}}{=} \{\mathbf{f} \in H_{curl} : \operatorname{curl} \mathbf{f} = 0\}$ with respect to the weighted scalar products

$$\langle \mathbf{e}, \mathbf{f} \rangle_\varepsilon \stackrel{\text{def}}{=} \int_{\mathbb{R}^3} \varepsilon \mathbf{e} \cdot \mathbf{f} dx \quad \text{and} \quad \langle \mathbf{g}, \mathbf{h} \rangle_\mu \stackrel{\text{def}}{=} \int_{\mathbb{R}^3} \mu \mathbf{g} \cdot \mathbf{h} dx,$$

respectively. Furthermore, let P_0 be the orthogonal projector on $H_{div,0} \stackrel{\text{def}}{=} \{\mathbf{f} \in H_{curl} : \operatorname{div} \mathbf{f} = 0\}$ with respect to the standard scalar product. It can be regarded as a pseudodifferential operator of order zero and can be expressed by using a Fourier transform:

$$(1 - P_0)\mathbf{u} = \mathcal{F}^{-1}\left(|k|^{-2} k \cdot \widehat{\mathbf{u}}(k) k\right). \tag{2.14}$$

One of the main results in [5] is the existence of solutions.

Proposition 2.1. *Assume* (2.1)–(2.9). *Then problem* (1.1)–(1.4) *admits a weak solution* $(\mathbf{E}, \mathbf{H}, \mathbf{M})$ *with the properties*

$$(\mathbf{E}, \mathbf{H}) \in C([0, \infty), X) \quad \textit{and} \quad \mathbf{M} \in W^{1,\infty}_{loc}([0, \infty), L^2(G)) \cap L^\infty_{loc}([0, \infty), L^\infty(G)).$$

Here and in what follows, the space variable $x \in \mathbb{R}^3$ is often omitted in the notation for the sake of brevity. The Maxwell system (1.1) is fulfilled in the sense that

$$(\mathbf{E}(t), \mathbf{H}(t)) = \exp(tB)(\mathbf{E}_0, \mathbf{H}_0) \tag{2.15}$$

$$- \int_0^t \exp((t - s)B)\left[\mathcal{R}\partial_t \mathbf{M}(s) + (\varepsilon^{-1}\mathbf{j}(s), 0) + F_\sigma(\mathbf{E}(s), \mathbf{H}(s))\right] ds,$$

where $(\exp(tB))_{t \in \mathbb{R}}$ is the unitary group generated by the skew adjoint operator B in the Hilbert space $X \stackrel{\text{def}}{=} L^2(\mathbb{R}^3, \mathbb{C}^6)$ introduced in [5, sect. 2]; see [9]. The definitions of $F_\sigma : X \to X$ and $\mathcal{R} : L^2(G) \to X$ can also be found in [5, sect. 2].

In what follows let $\mathcal{Z}$ be the set of all $\mathbf{f} \in H_{curl,0}$ with $\mathbf{f}(x) = 0$ for all $x \in S_\sigma$, where S_σ is given in (2.11).

Let P_1 be the orthogonal projector on $\mathcal{Z}$ in $L^2(\mathbb{R}^3)$ with respect to the weighted scalar product $\langle \cdot, \cdot \rangle_\varepsilon$. Now the main result of this paper concerning weak convergence for $t \to \infty$ can be stated.

THEOREM 2.2. *Assume* (2.1)–(2.13). *Then every solution* $(\mathbf{E}, \mathbf{H}, \mathbf{M})$ *(with the previously mentioned properties) to problem* (1.1)–(1.4) *satisfies*

$$\mathbf{E}(t) \stackrel{t\to\infty}{\longrightarrow} P_1 \left(\mathbf{E}_0 - \int_0^\infty \varepsilon^{-1} \mathbf{j}(s) ds \right) \text{ in } L^2(\mathbb{R}^3) \text{ weakly,}$$

$$P_H \mathbf{H}(t) \stackrel{t\to\infty}{\longrightarrow} 0 \text{ and } P_0 \mathbf{H}(t) \stackrel{t\to\infty}{\longrightarrow} 0 \text{ in } L^2(\mathbb{R}^3) \text{ weakly.} \tag{2.16}$$

COROLLARY 2.3. *Suppose that, in addition to the assumptions of Theorem* 2.2*,*

$$\int_{\mathbb{R}^3} \mathbf{D}_0 \mathbf{f} dx = 0 \text{ for all } \mathbf{f} \in \mathcal{Z}, \tag{2.17}$$

where $\mathbf{D}_0 \stackrel{\text{def}}{=} \varepsilon \mathbf{E}_0 - \int_0^\infty \mathbf{j}(s) ds$. *Then every solution* $(\mathbf{E}, \mathbf{H}, \mathbf{M})$ *to problem* (1.1)–(1.4) *satisfies*

$$\mathbf{E}(t) \stackrel{t\to\infty}{\longrightarrow} 0 \text{ in } L^2(\mathbb{R}^3) \text{ weakly.}$$

Since $\nabla\varphi \in \mathcal{Z}$ for all $\varphi \in C_0^\infty(\mathbb{R}^3 \setminus \overline{S_\sigma})$, condition (2.17) includes

$$\text{div } \mathbf{D}_0 = 0 \text{ on } \mathbb{R}^3 \setminus \overline{S_\sigma}.$$

By (1.1) one has

$$\text{div } [\varepsilon \mathbf{E}(t)] = \text{div } \left[\varepsilon \mathbf{E}_0 - \int_0^t \mathbf{j}(s) ds \right] \stackrel{t\to\infty}{\longrightarrow} \text{div } \mathbf{D}_0 = 0 \text{ in } \mathcal{D}'(\mathbb{R}^3 \setminus \overline{S_\sigma})$$

if condition (2.17) is fulfilled. The physical meaning of this is that the space charge $\rho \stackrel{\text{def}}{=} \text{div } [\varepsilon \mathbf{E}(t)]$ determined by the initial state $(\mathbf{E}_0, \mathbf{H}_0)$ and the prescribed current $\mathbf{j}$ decays on the nonconducting region $\mathbb{R}^3 \setminus S_\sigma$ as $t \to \infty$.

The next theorem concerns the local energy decay of $\mathbf{E}$ and $Q_H \mathbf{H}$ on the conducting part of the medium.

THEOREM 2.4. *Suppose that the assumptions of Theorem* 2.2 *are satisfied and that, in addition to assumption* (2.3)*, there exists some* $\mathbf{A}_0 \in H_{curl}$ *with*

$$\text{curl } \mathbf{A}_0 = \mu[\mathbf{H}_0 + \tilde{\mathbf{M}}_0]. \tag{2.18}$$

Let $K \subset \mathbb{R}^3$ *be a compact set such that* σ *has a positive lower bound on some open neighborhood* U *of* K*; i.e., let* $U \subset \mathbb{R}^3$ *be open and* $K \subset U$ *be compact with*

$$\sigma(x) \geq c_0 \text{ for all } x \in U \tag{2.19}$$

with some $c_0 > 0$. *Then every solution* $(\mathbf{E}, \mathbf{H}, \mathbf{M})$ *to problem* (1.1)–(1.4) *satisfies*

$$\|\mathbf{E}(t)\|_{L^2(K)} + \|Q_H \mathbf{H}(t)\|_{L^2(K)} + \|Q_0 \mathbf{H}(t)\|_{L^2(K)} \stackrel{t\to\infty}{\longrightarrow} 0.$$

3. Weak convergence for $\to \infty$, proof of Theorem 2.2. This section is concerned with the proof of Theorem 2.2 concerning the weak convergence of the field quantities for $t \to \infty$. For this purpose the basic energy dissipation law is proved first.

LEMMA 3.1. *Assume* (2.1)–(2.13). *Then every solution* $(\mathbf{E}, \mathbf{H}, \mathbf{M})$ *to problem* (1.1)–(1.4) *satisfies*

$$(\mathbf{E}, \mathbf{H}) \in L^\infty((0,\infty), X), \quad \partial_t \mathbf{M} \in L^\infty((0,\infty), L^2(G)), \tag{3.1}$$

$$\beta^{1/2} \partial_t \mathbf{M} \in L^2((0,\infty), L^2(G)) \tag{3.2}$$

with the nonnegative function $\beta \in L^\infty(G)$ *occurring in assumption* (2.6). *Furthermore,*

$$|\mathbf{M}(t,x)| = |\mathbf{M}_0(x)| \le C_0 \tag{3.3}$$

and

$$\int_0^\infty \int_{\mathbb{R}^3} \sigma |\mathbf{E}(t)|^2 dx dt < \infty, \tag{3.4}$$

in particular

$$F_\sigma(\mathbf{E}(\cdot), \mathbf{H}(\cdot)) \in L^2((0,\infty), X). \tag{3.5}$$

Proof. By assumption (2.5) multiplication of both sides of (1.2) with $\mathbf{M}$ gives $\mathbf{M}\partial_t \mathbf{M} = 0$, whence (3.3). By the standard energy estimate for (2.15) one has

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}&\|(\mathbf{E}(t),\mathbf{H}(t))\|_X^2 \\ &= -\langle (\mathbf{E}(t),\mathbf{H}(t)), F_\sigma(\mathbf{E}(t),\mathbf{H}(t)) + (\varepsilon^{-1}\mathbf{j}(t),0) + \mathcal{R}\partial_t \mathbf{M}(t)\rangle_X \\ &= -\int_{\mathbb{R}^3} \mathbf{E}(t)\left[\sigma \mathbf{E}(t) + \mathbf{j}(t)\right] dx - \int_G \mu \mathbf{H}(t) \cdot \partial_t \mathbf{M}(t) dx.\end{aligned} \tag{3.6}$$

Hence the energy functional

$$\mathcal{E}(t) \stackrel{\text{def}}{=} \frac{1}{2}\left(\|(\mathbf{E}(t),\mathbf{H}(t))\|_X^2 + \int_G \mu \mathbf{M}(t)\cdot A\mathbf{M}(t) dx \right) \tag{3.7}$$

satisfies

$$\begin{aligned}\mathcal{E}'(t) \le \|&\varepsilon^{-1/2}\mathbf{j}(t)\|_{L^2(\mathbb{R}^3}\mathcal{E}(t)^{1/2} \\ &- \int_{\mathbb{R}^3} \sigma |\mathbf{E}(t)|^2 - \int_G \mu [\mathbf{H}(t) - A\mathbf{M}(t)]\cdot \partial_t \mathbf{M}(t) dx.\end{aligned} \tag{3.8}$$

With

$$\begin{aligned}[\mathbf{H} - A\mathbf{M}]\cdot \partial_t \mathbf{M} &= [\mathbf{H} - A\mathbf{M}]\cdot F(x,\mathbf{M})[\mathbf{H} - A\mathbf{M}] \\ &\ge \beta |F(x,\mathbf{M})[\mathbf{H} - A\mathbf{M}]|^2 = \beta |\partial_t \mathbf{M}|^2\end{aligned}$$

by (3.3) and assumption (2.6), it follows from (3.8) that

$$\mathcal{E}'(t) \le -\int_{\mathbb{R}^3} \sigma |\mathbf{E}(t)|^2 dx + \|\varepsilon^{-1/2}\mathbf{j}(t)\|_{L^2}\mathcal{E}(t)^{1/2} - \int_G \mu\beta |\partial_t \mathbf{M}|^2 dx. \tag{3.9}$$

Since $\mathbf{j} \in L^1((0,\infty), L^2(\mathbb{R}^3))$, this completes the proof. ∎

Next the weak ω-limit set of $(\mathbf{E},\mathbf{H})$ will be characterized by means of the following proposition.

PROPOSITION 3.2. *Assume* (2.13) *and let* $\mathbf{g} \in X$ *with*

$$\underline{(\exp(tB)\mathbf{g})}_1 = 0 \text{ on } \mathcal{U} \text{ for all } t \in \mathbb{R}. \tag{3.10}$$

Then $\mathbf{g} \in H_{curl,0} \times H_{curl,0} = \ker B$.

Here and in what follows, $\underline{\mathbf{u}}_1 \in \mathbb{C}^3$ denotes the first three and $\underline{\mathbf{u}}_2 \in \mathbb{C}^3$ denotes the last three components of a vector $\mathbf{u} \in \mathbb{C}^6$.

Proof. In Lemma 2 of [10] this assertion is proved for the scalar wave equation with smooth coefficients and Maxwell's equations with constant coefficients on arbitrary spatial domains. In [4] this has been generalized to discontinuous coefficients ε and μ satisfying (2.13). However, in the whole space case a shorter proof is possible. For this purpose let

$$\mathcal{A}(f)\mathbf{u} \overset{\text{def}}{=} \int_{\mathbb{R}} \hat{f}(t) \exp(-tB)\mathbf{u}dt \tag{3.11}$$

for $f \in C_0^\infty(\mathbb{R})$ and $\mathbf{u} \in X$, where $\hat{f}$ denotes the Fourier transform of f. Then $\mathcal{A}(f)\mathbf{u} \in D(B^k)$, and with $f_k(\lambda) = (-i\lambda)^k f(\lambda)$ one has

$$B^k \mathcal{A}(f)\mathbf{u} = \int_{\mathbb{R}} \hat{f}^{(k)}(t) \exp(-tB)\mathbf{g}dt = \mathcal{A}(f_k)\mathbf{u}. \tag{3.12}$$

Suppose that $\mathbf{g} \in X$ satisfies (3.10). Let $f \in C_0^\infty(\mathbb{R})$, $\mathbf{G} \overset{\text{def}}{=} \mathcal{A}(f)\mathbf{g} \in D(B)$. Then (3.10), (3.11), and (3.12) yield $\underline{(B^k\mathbf{G})}_1 = 0$ on $\mathcal{U}$. Since

$$B^{k+1}\mathbf{G} = \Big(\varepsilon^{-1} \text{ curl } \underline{(B^k\mathbf{G})}_2, -\mu^{-1} \text{ curl } \underline{(B^k\mathbf{G})}_1\Big),$$

this also gives $\underline{(B^k\mathbf{G})}_2 = 0$ on the open set $\mathcal{U}$ and, thus,

$$B^k\mathbf{G} = 0 \text{ on } \mathcal{U} \text{ for all } k \geq 1. \tag{3.13}$$

By assumption (2.13) it follows from (3.13) that

$$B^{k+1}\mathbf{G} = \Big(\varepsilon\underline{(B^{k+1}\mathbf{G})}_1, \mu\underline{(B^{k+1}\mathbf{G})}_2\Big) = \Big(\text{curl } \underline{(B^k\mathbf{G})}_2, -\text{ curl } \underline{(B^k\mathbf{G})}_1\Big) \tag{3.14}$$

for all $k \geq 0$ on $\mathbb{R}^3$ in the sense of distributions. By induction (3.14) yields

$$(1-\Delta)^k B\mathbf{G} = \mathcal{A}(h_k)\mathbf{g} \in L^2(\mathbb{R}^3) \tag{3.15}$$

in the sense of distributions with $h_k(\lambda) \overset{\text{def}}{=} -i\lambda(1+\lambda^2)^k f(\lambda)$. By (3.11) and (3.15) one obtains

$$\begin{aligned} \|(1-\Delta)^k B\mathbf{G}\|_{L^2(\mathbb{R}^3)} &\leq \|\widehat{h_k}\|_{L^1(\mathbb{R})}\|\mathbf{g}\|_X \leq C_1\|(1+\xi^2)\widehat{h_k}(\xi)\|_{L^\infty(\mathbb{R})} \\ &\leq C_2\|h_k - h_k''\|_{L^1(\mathbb{R})} \leq C_3^k \text{ for all } k \in \mathbb{N}. \end{aligned} \tag{3.16}$$

By Sobolev's embedding theorem one obtains from (3.16) $B\mathbf{G} \in C^\infty(\mathbb{R}^3)$ and

$$\|\partial^\alpha B\mathbf{G}\|_{L^\infty} \leq C_4\|\partial^\alpha B\mathbf{G}\|_{H^2(\mathbb{R}^3)} \leq C_5\|(1-\Delta)^{n+1}B\mathbf{G}\|_{L^2(\mathbb{R}^3)} \leq C_6^{n+1} \tag{3.17}$$

for all $n \in \mathbb{N}$ and $|\alpha| \leq 2n$ with $C_1 - C_6 \in (0,\infty)$ independent of n. This yields the real analyticity of $B\mathbf{G}$. By (3.13) this analyticity implies $B\mathbf{G} = 0$ on all of $\mathbb{R}^3$, i.e.,

$$\mathcal{A}(f)\mathbf{g} \in \ker B \text{ for all } f \in C_0^\infty(\mathbb{R}). \tag{3.18}$$

Since $\ker B$ is a closed subspace of $X = L^2(\mathbb{R}^3)$ and there is a sequence $f_n \in C_0^\infty(\mathbb{R})$ with $\mathcal{A}(f_n)\mathbf{g} \overset{n\to\infty}{\longrightarrow} \mathbf{g}$, (3.18) completes the proof. $\square$

In what follows let ω_0 denote the ω-limit set of $(\mathbf{E},\mathbf{H})$ with respect to the weak topology of X, i.e., the set of all $\mathbf{g}\in X$ such that there exists a sequence $t_n \overset{n\to\infty}{\longrightarrow} \infty$ with $(\mathbf{E}(t_n),\mathbf{H}(t_n)) \overset{n\to\infty}{\longrightarrow} \mathbf{g}$ in X weakly. Since $(\mathbf{E},\mathbf{H})\in L^\infty((0,\infty),X)$ by Lemma 3.1 this set is nonempty.

LEMMA 3.3. *Under the assumptions of Theorem* 2.2 *it follows that* $\omega_0 \subset \mathcal{Z}\times H_{curl,0} \subset \ker B$.

Proof. Suppose $\mathbf{g}\in X$ and $t_n \overset{n\to\infty}{\longrightarrow} \infty$ with

$$(\mathbf{E}(t_n),\mathbf{H}(t_n)) \overset{n\to\infty}{\longrightarrow} \mathbf{g} \text{ in } X \text{ weakly.} \tag{3.19}$$

Let $\mathbf{u}^{(n)}(t) \overset{\text{def}}{=} (\mathbf{E}(t_n+t),\mathbf{H}(t_n+t))$ for $t\in(-t_n,\infty)$ and $\mathbf{u}^{(n)}(t)\overset{\text{def}}{=} 0$ for $t\le -t_n$. Next, let $\mathbf{m}^{(n)}(t)\overset{\text{def}}{=}\mathbf{M}(t_n+t)$ for $t\in(-t_n,\infty)$ and $\mathbf{m}^{(n)}(t)\overset{\text{def}}{=}0$ for $t\le -t_n$. After passing to a further subsequence, one has by (3.1)

$$\partial_t \mathbf{m}^{(n)} \overset{n\to\infty}{\longrightarrow} \mathbf{r}^{(\infty)} \text{ in } L^\infty(\mathbb{R},L^2(G)) \text{ weak-}*. \tag{3.20}$$

Let $t\in\mathbb{R}$. By (2.15) one has

$$\begin{aligned}
\mathbf{u}^{(n)}(t) &= \exp(tB)(\mathbf{E}(t_n),\mathbf{H}(t_n)) \\
&\quad - \int_{t_n}^{t_n+t} \exp((t_n+t-\tau)B)\left[F_\sigma(\mathbf{E}(\tau),\mathbf{H}(\tau)) + \mathcal{R}\partial_t\mathbf{M}(\tau) + (\varepsilon^{-1}\mathbf{j}(\tau),0)\right] d\tau \\
&= \exp(tB)(\mathbf{E}(t_n),\mathbf{H}(t_n)) \\
&\quad - \int_0^t \exp((t-s)B)\left[F_\sigma(\mathbf{E}(s+t_n),\mathbf{H}(s+t_n)) + \mathcal{R}\partial_t\mathbf{m}^{(n)}(s) + (\varepsilon^{-1}\mathbf{j}(s+t_n),0)\right] ds
\end{aligned}$$

for all $n\in\mathbb{N}$ with $t_n+t\ge 0$. (In order to apply Proposition 3.2 it is necessary also to consider $t\le 0$.) With $\mathbf{j}\in L^1((0,\infty),L^2(\mathbb{R}^3))$, it follows from (3.5) and (3.19) that

$$\begin{aligned}
\mathbf{u}^{(n)}(t) &\overset{n\to\infty}{\longrightarrow} \mathbf{u}^{(\infty)}(t) \\
&\overset{\text{def}}{=} \exp(tB)\mathbf{g} - \int_0^t \exp((t-s)B)\mathcal{R}\mathbf{r}^{(\infty)}(s)ds \text{ in } X \text{ weakly for all } t\in\mathbb{R}.
\end{aligned} \tag{3.21}$$

In particular, $\mathbf{u}_\infty\in C(\mathbb{R},X)$ is a generalized solution of

$$\partial_t \mathbf{u}^{(\infty)} = B\mathbf{u}^{(\infty)} - \mathcal{R}\mathbf{r}^{(\infty)} \tag{3.22}$$

in the sense of (2.15). For all $a,b\in\mathbb{R}$ with $a<b$ it follows from (3.21) that

$$\int_a^b \sigma\underline{\mathbf{u}^{(n)}}_1(t)dt \overset{n\to\infty}{\longrightarrow} \int_a^b \sigma\underline{\mathbf{u}^{(\infty)}(t)}_1 dt \text{ in } L^2(S_\sigma). \tag{3.23}$$

On the other hand, it follows from (3.5) that

$$\left\|\int_a^b \sigma\underline{\mathbf{u}^{(n)}}_1(t)dt\right\|_{L^2(S_\sigma)} \le (b-a)^{1/2}\left(\int_{a+t_n}^{b+t_n} \|\sigma\mathbf{E}(t)\|^2_{L^2(S_\sigma)}dt\right)^{1/2} \overset{n\to\infty}{\longrightarrow} 0. \tag{3.24}$$

Now (3.23) and (3.24) yield $\int_a^b \sigma \underline{\mathbf{u}^{(\infty)}}_1(t)dt = 0$ on S_σ for all $a, b \in \mathbb{R}$, $a < b$, which implies that

$$\underline{\mathbf{u}^{(\infty)}}_1(t) = 0 \text{ on } S_\sigma \text{ for all } t \in \mathbb{R}. \tag{3.25}$$

Let $\widetilde{\mathbf{r}}^{(\infty)}(t)$ denote the extension of $\mathbf{r}^{(\infty)}(t)$ by zero outside G. Then (3.22) and (3.25) yield, by assumption (2.12),

$$\operatorname{curl} \underline{\mathbf{u}^{(\infty)}}_2 = \partial_t(\varepsilon \underline{\mathbf{u}^{(\infty)}}_1) = 0 \tag{3.26}$$

and

$$\partial_t(\mu \underline{\mathbf{u}^{(\infty)}}_2) + \mu \widetilde{\mathbf{r}}^{(\infty)} = -\operatorname{curl} \underline{\mathbf{u}^{(\infty)}}_1 = 0 \tag{3.27}$$

on $\mathbb{R} \times \mathcal{U}$ in the sense of distributions. Since $\mu^{-1} \in L^\infty(\mathbb{R}^3, \mathbb{R}^{3\times 3})$ is independent of t, it follows easily from (3.27) that

$$\partial_t \underline{\mathbf{u}^{(\infty)}}_2 + \widetilde{\mathbf{r}}^{(\infty)} = 0 \text{ on } \mathbb{R} \times \mathcal{U}.$$

Invoking (3.26) yields

$$\operatorname{curl} \widetilde{\mathbf{r}}^{(\infty)} = -\partial_t \operatorname{curl} \underline{\mathbf{u}^{(\infty)}}_2 = 0 \tag{3.28}$$

on $\mathbb{R} \times \mathcal{U}$ in the sense of distributions.

Next, it follows from (3.2) that for all $a < b$,

$$\left\| \beta^{1/2} \int_a^b \partial_t \mathbf{m}^{(n)}(t)dt \right\|_{L^2(G)} \leq (b-a)^{1/2} \left(\int_{a+t_n}^{b+t_n} \|\beta^{1/2} \partial_t \mathbf{M}(t)\|^2_{L^2(G)} dt \right)^{1/2} \stackrel{n\to\infty}{\longrightarrow} 0,$$

which implies that

$$\int_a^b \mathbf{r}^{(\infty)}(t, x)dt = 0 \text{ for all } x \in G \text{ with } \beta(x) > 0.$$

Hence, one obtains by (2.10)

$$\mathbf{r}^{(\infty)}(t) = 0 \text{ a.e. on } G \setminus G_0$$

and, thus,

$$\widetilde{\mathbf{r}}^{(\infty)}(t) = 0 \text{ a.e. on } \mathbb{R}^3 \setminus \overline{G_0}, \tag{3.29}$$

since $\widetilde{\mathbf{r}}^{(\infty)}(t)$ is the extension of $\mathbf{r}^{(\infty)}(t)$ by zero outside G. Recall that $\overline{G_0} \subset \mathcal{U}$ by assumption (2.12). Hence, (3.28) and (3.29) imply

$$\widetilde{\mathbf{r}^{(\infty)}(t)} \in H_{curl,0} \text{ for all } t \in \mathbb{R}.$$

In particular,

$$\mathcal{R}\mathbf{r}^{(\infty)}(t) \in \ker B \text{ for all } t \in \mathbb{R}. \tag{3.30}$$

Going back to (3.21), one obtains from (3.30)

$$\mathbf{u}^{(\infty)}(t) = \exp(tB)\mathbf{g} - \int_0^t \mathcal{R}\mathbf{r}^{(\infty)}(s)ds \text{ for all } t \in \mathbb{R}.$$

Since $\underline{(\mathcal{R}\mathbf{r}^{(\infty)}(s))}_1 = 0$, one has

$$\underline{\mathbf{u}^{(\infty)}(t)}_1 = \underline{(\exp{(tB)}\mathbf{g})}_1 \text{ for all } t \in \mathbb{R},$$

in particular

$$\underline{(\exp{(tB)}\mathbf{g})}_1 = 0 \text{ on } S_\sigma \text{ for all } t \in \mathbb{R}, \tag{3.31}$$

by (3.25). Now it follows from assumption (2.12), (3.31), and Proposition 3.2 that $\mathbf{g} \in H_{curl,0} \times H_{curl,0}$. Furthermore, $\underline{\mathbf{g}}_1 = 0$ on S_σ by (3.31), whence $\underline{\mathbf{g}}_1 \in \mathcal{Z}$. ◻

Remark 2. The open neighborhood U of G_0 with positive conductivity in assumption (2.12) is necessary to show that $\widetilde{\mathbf{r}}^{(\infty)}(t)$, the extension of $\mathbf{r}^{(\infty)}(t)$ by zero outside G, is curl free on all of $\mathbb{R}^3$. (In particular, the fact that $\mathbf{r}^{(\infty)}(t)$ is supported in G_0 is used here.) If only $G_0 \subset S_\sigma$ instead of (2.12) is assumed, one can only conclude that curl $\mathbf{r}^{(\infty)} = 0$ on $\mathbb{R} \times G_0$, but in general not curl $\widetilde{\mathbf{r}}^{(\infty)} = 0$ on $\mathbb{R} \times \mathbb{R}^3$, which is required for (3.30).

LEMMA 3.4. *Under the assumptions of Theorem* 2.2 *it follows that*

$$\left\| P_1\mathbf{E}(t) - P_1\left[\mathbf{E}_0 - \int_0^\infty \varepsilon^{-1}\mathbf{j}(s)ds\right]\right\|_{L^2(\mathbb{R}^3)} \stackrel{t\to\infty}{\longrightarrow} 0 \tag{3.32}$$

and

$$(1-P_H)\mathbf{H}(t) = -(1-P_H)\tilde{\mathbf{M}}(t) \tag{3.33}$$

for every solution $(\mathbf{E}, \mathbf{H}, \mathbf{M})$ *to problem* (1.1)–(1.4).

Proof. Since

$$\text{ran } P_H = \{\mathbf{h} \in L^2(\mathbb{R}^3) : \text{div } (\mu\mathbf{h}) = 0\},$$

it follows from (2.3) that

$$(1-P_H)(\mathbf{H}_0 + \tilde{\mathbf{M}}_0) = 0. \tag{3.34}$$

Let $\mathbf{a} \in \mathcal{Z} \times H_{curl,0} \subset \ker B$, which means $\mathbf{a} \in \ker B$ and $\underline{\mathbf{a}}_1 = 0$ on S_σ. Then (2.15) and (3.34) yield

$$\begin{aligned} \langle (P_1\mathbf{E}(t), (1-P_H)\mathbf{H}(t)), \mathbf{a}\rangle_X &= \langle (\mathbf{E}(t), \mathbf{H}(t)), \mathbf{a}\rangle_X \\ = \left\langle (\mathbf{E}_0, \mathbf{H}_0) - \int_0^t \left[\mathcal{R}\partial_t\mathbf{M}(s) + F_\sigma(\mathbf{E}(s), \mathbf{H}(s)) + (\varepsilon^{-1}\mathbf{j}(s), 0)\right] ds, \mathbf{a}\right\rangle_X \\ &= \left\langle \left(\mathbf{E}_0 - \int_0^t \varepsilon^{-1}\mathbf{j}(s)ds, \mathbf{H}_0 + \tilde{\mathbf{M}}_0 - \tilde{\mathbf{M}}(t)\right), \mathbf{a}\right\rangle_X \\ &= \left\langle \left(P_1\left[\mathbf{E}_0 - \int_0^t \varepsilon^{-1}\mathbf{j}(s)ds\right], -(1-P_H)\tilde{\mathbf{M}}(t)\right), \mathbf{a}\right\rangle_X, \end{aligned}$$

whence (3.33) and

$$P_1\mathbf{E}(t) = P_1\left[\mathbf{E}_0 - \int_0^t \varepsilon^{-1}\mathbf{j}(s)ds\right], \tag{3.35}$$

from which the assertion follows, since $\mathbf{j} \in L^1(0, \infty, L^2(\mathbb{R}^3))$. ◻

Completion of the proof of Theorem 2.2. Since $1 - P_1$ and P_H are the orthogonal projectors on $\mathcal{Z}^\perp$ and $H^\perp_{curl,0}$, respectively, it follows from Lemma 3.3 that zero is the only possible accumulation point of $((1-P_1)\mathbf{E}(t), P_H\mathbf{H}(t))$ and $((1-P_1)\mathbf{E}(t), P_0\mathbf{H}(t))$ for $t \to \infty$ with respect to the weak topology of X. Hence, by Lemma 3.1,

$$P_0\mathbf{H}(t) \overset{t\to\infty}{\longrightarrow} 0 \text{ in } L^2(\mathbb{R}^3) \text{ weakly}$$

and

$$((1-P_1)\mathbf{E}(t), P_H\mathbf{H}(t)) \overset{t\to\infty}{\longrightarrow} 0 \text{ in } X \text{ weakly.} \tag{3.36}$$

Finally, the assertion follows from Lemma 3.4 and (3.36). ∎

4. Local strong decay, proof of Theorem 2.4. In what follows let $(\mathbf{E}, \mathbf{H}, \mathbf{M})$ be a solution to problem (1.1)–(1.4). The following lemma is based on a result in [5].

LEMMA 4.1. *Let $T > 0$ and $\mathbf{g} \in W^{1,\infty}((0,\infty), L^2(\mathbb{R}^3))$. Then*

$$\int_0^T \chi(s) \int_{\mathbb{R}^3} \psi\mathbf{g}(t+s) \cdot P_H\mathbf{H}(t+s)dxds \overset{t\to\infty}{\longrightarrow} 0 \tag{4.1}$$

for all $\chi \in C_0^\infty(0,T)$ and $\psi \in C_0^\infty(\mathbb{R}^3)$.

Proof. Let $\{t_n\}_{n\in\mathbb{N}}$ with $t_n \overset{n\to\infty}{\longrightarrow} \infty$, $\mathbf{h}_n(t) \overset{\text{def}}{=} P_H\mathbf{H}(t_n + t)$, and $\mathbf{G}_n(t) \overset{\text{def}}{=} \chi(t)\psi\mathbf{g}(t_n + t)$. Then it follows from (2.16) that

$$\mathbf{h}_n \overset{n\to\infty}{\longrightarrow} 0 \text{ in } L^\infty((0,T), L^2(\mathbb{R}^3)) \text{ weak-}*. \tag{4.2}$$

Since ran $(1 - P_H) \subset H_{curl,0}$, one has by (1.1)

$$\int_{\mathbb{R}^3} \mathbf{h}_n(t) \text{ curl } \varphi dx = \int_{\mathbb{R}^3} \mathbf{H}(t_n+t) \text{ curl } \varphi dx = \frac{d}{dt}\int_{\mathbb{R}^3} \mathbf{D}_n(t)\cdot\varphi dx \tag{4.3}$$

for all $\varphi \in C_0^\infty(\mathbb{R}^3)$ with

$$\mathbf{D}_n(t) \overset{\text{def}}{=} \varepsilon\mathbf{E}(t_n+t) + \int_{t_n}^{t_n+t} (\sigma\mathbf{E}(s) + \mathbf{j}(s))ds.$$

By Lemma 3.1 the sequences $\{\mathbf{D}_n\}_{n\in\mathbb{N}}$, $\{\mathbf{G}_n\}_{n\in\mathbb{N}}$, and $\{\mathbf{h}_n\}_{n\in\mathbb{N}}$ are bounded sequences in $L^\infty((0,T), L^2(\mathbb{R}^3))$. Furthermore, $\{\mathbf{G}_n\}_{n\in\mathbb{N}}$ is equicontinuous in the sense that for all $\theta > 0$ there exists some $\delta > 0$ such that

$$\|\mathbf{G}_n(t) - \mathbf{G}_n(s)\|_{L^2(\mathbb{R}^3)} \le \theta \text{ for all } n \in \mathbb{N} \text{ and } s,t \in (0,T) \text{ with } |s-t| \le \delta.$$

Hence, $\{\mathbf{D}_n\}_{n\in\mathbb{N}}$, $\{\mathbf{G}_n\}_{n\in\mathbb{N}}$, and $\{\mathbf{h}_n\}_{n\in\mathbb{N}}$ satisfy all assumptions required for Lemma 3.4 in [5]. Since ψ has compact support, Lemma 3.4 in [5] yields

$$\int_0^T \chi(s) \int_{\mathbb{R}^3} \psi\mathbf{g}(t_n+s)\cdot P_H\mathbf{H}(t_n+s)dxds = \int_0^T\int_{\text{supp }\psi} \psi\mathbf{G}_n(s)\cdot P_H\mathbf{h}_n(s)dxds \overset{n\to\infty}{\longrightarrow} 0,$$

which completes the proof. ∎

Proof of Theorem 2.4. The decay of $P_H\mathbf{H}$ does not follow directly from (3.9), since $\mathbf{H}$ does not occur in the right-hand side of (3.9). In order to obtain a dissipative term also for the magnetic field, a vector potential $\mathbf{A}$ is introduced.

Let $\mathbf{A}_1 \stackrel{\text{def}}{=} P_0\mathbf{A}_0 \in H_{curl} \cap \operatorname{ran} P_0 = H_{curl} \cap H_{div,0}$ with $\mathbf{A}_0 \in H_{curl}$ as in (2.18). Then

$$\text{curl } \mathbf{A}_1 = \text{curl } \mathbf{A}_0 = \mu[\mathbf{H}_0 + \tilde{\mathbf{M}}_0] \tag{4.4}$$

by (2.18). Let

$$\mathbf{A}(t) \stackrel{\text{def}}{=} \mathbf{A}_1 - \int_0^t P_0\mathbf{E}(s)ds \in H_{curl} \cap H_{div,0}. \tag{4.5}$$

Then it follows from (1.1), (4.4), and (4.5) that for all $\mathbf{g} \in H_{curl}$, one has

$$\begin{aligned}\int_{\mathbb{R}^3} \mathbf{A}(t) \text{ curl } \mathbf{g}dx &= \int_{\mathbb{R}^3} \mathbf{A}_1 \text{ curl } \mathbf{g}dx - \int_0^t\int_{\mathbb{R}^3} P_0\mathbf{E}(s)\cdot \text{curl } \mathbf{g}dxds \\ &= \int_{\mathbb{R}^3} \mu[\mathbf{H}_0 + \tilde{\mathbf{M}}_0]\cdot \mathbf{g}dx - \int_0^t\int_{\mathbb{R}^3} \mathbf{E}(s) \text{ curl } \mathbf{g}dxds \\ &= \int_{\mathbb{R}^3} \mu(\mathbf{H}(t) + \tilde{\mathbf{M}}(t))\cdot \mathbf{g}dx.\end{aligned}$$

Hence, $\mathbf{A}(t) \in H_{curl} \cap H_{div,0}$ and, by (3.33),

$$\text{curl } \mathbf{A}(t) = \mu(\mathbf{H}(t) + \tilde{\mathbf{M}}(t)) = \mu P_H(\mathbf{H}(t) + \tilde{\mathbf{M}}(t)). \tag{4.6}$$

By a classical estimate one has

$$\|\mathbf{a}\|_{L^6(\mathbb{R}^3)} \le C_2 \|\text{curl } \mathbf{a}\|_{L^2(\mathbb{R}^3)} \text{ for all } \mathbf{a} \in H_{curl} \cap H_{div,0}. \tag{4.7}$$

Thus, (4.6) yields

$$\|\mathbf{A}(t)\|_{L^6(\mathbb{R}^3)} \le C_2\|P_H(\mathbf{H}(t) + \tilde{\mathbf{M}}(t))\|_{L^2(\mathbb{R}^3)} \le C_3, \tag{4.8}$$

where C_1–C_3 are independent of t. By (4.6) and the estimates (3.1) and (4.8), one has $\mathbf{A} \in L^\infty((0,\infty), L^6(\mathbb{R}^3))$ and curl $\mathbf{A} \in L^\infty((0,\infty), L^2(\mathbb{R}^3))$. Therefore, it follows that

$$\{\mathbf{A}(t) : t \in (0,\infty)\} \text{ is precompact in } L^2(K) \tag{4.9}$$

for all compact subsets $K \subset \mathbb{R}^3$. Now let $U \subset \mathbb{R}^3$ and $K \subset U$ be as in (2.19) and choose $\psi \in C_0^\infty(U)$ with $\psi = 1$ on K. First, it follows from Lemma 3.1 and (2.19) that

$$\int_0^T \|\psi^{1/2}\mathbf{E}(t+s)\|_{L^2(\mathbb{R}^3)}ds \stackrel{t\to\infty}{\longrightarrow} 0 \text{ for all } T > 0. \tag{4.10}$$

Let

$$g(t) \stackrel{\text{def}}{=} \int_{\mathbb{R}^3} \psi\varepsilon\mathbf{E}(t)\mathbf{A}(t)dx. \tag{4.11}$$

By (4.8), Hölder's inequality, and (4.10) one has, for all $T > 0$,

$$\int_0^T |g(t+s)|ds \le C_1 \int_0^T \|\psi\mathbf{E}(t+s)\|_{L^2(\mathbb{R}^3)}ds \stackrel{t\to\infty}{\longrightarrow} 0. \tag{4.12}$$

From (1.1), (3.1) again, (4.6), (4.5), and (4.11) it follows that

$$
\begin{aligned}
(4.13)\qquad g'(t) &= \int_{\mathbb{R}^3} \mathbf{H}(t) \text{ curl } (\psi \mathbf{A}(t)) dx - \int_{\mathbb{R}^3} \psi[\sigma \mathbf{E}(t) + \mathbf{j}(t)]\mathbf{A}(t) dx \\
&\quad - \int_{\mathbb{R}^3} \psi \varepsilon \mathbf{E}(t) \cdot P_0 \mathbf{E}(t) dx \\
&\geq \int_{\mathbb{R}^3} \mathbf{H}(t) \text{ curl } (\psi \mathbf{A}(t)) dx - C_1 \|\psi \mathbf{E}(t)\|_{L^2(\mathbb{R}^3)} - C_1 \|\psi \mathbf{j}(t)\|_{L^2(\mathbb{R}^3)}.
\end{aligned}
$$

With (4.6) one obtains

$$
\begin{aligned}
\int_{\mathbb{R}^3} \mathbf{H}(t) \text{ curl } (\psi \mathbf{A}(t)) dx &= \int_{\mathbb{R}^3} (P_H \mathbf{H}(t)) \text{ curl } (\psi \mathbf{A}(t)) dx \\
&= \int_{\mathbb{R}^3} \psi |\mu^{1/2} P_H \mathbf{H}(t)|^2 dx + \int_{\mathbb{R}^3} \psi \mu \cdot (P_H \mathbf{H}(t)) \cdot P_H \tilde{\mathbf{M}}(t) dx \\
&\quad + \int_{\mathbb{R}^3} (\nabla \psi) \wedge \mathbf{A}(t) \cdot P_H \mathbf{H}(t) dx
\end{aligned}
$$

and hence, by (4.13),

$$
\begin{aligned}
(4.14)\qquad \int_{\mathbb{R}^3} \psi |\mu^{1/2} P_H \mathbf{H}(t)|^2 dx &\leq g'(t) + C_1 \|\psi \mathbf{E}(t)\|_{L^2(\mathbb{R}^3)} + C_1 \|\psi \mathbf{j}(t)\|_{L^2(\mathbb{R}^3)} \\
&\quad - \int_{\mathbb{R}^3} \psi \mu \cdot (P_H \mathbf{H}(t)) \cdot P_H \tilde{\mathbf{M}}(t) dx - \int_{\mathbb{R}^3} (\nabla \psi) \wedge \mathbf{A}(t) \cdot P_H \mathbf{H}(t) dx.
\end{aligned}
$$

It follows from (2.16) and (4.9) that

$$
(4.15)\qquad \int_{\mathbb{R}^3} (\nabla \psi) \wedge \mathbf{A}(t) \cdot P_H \mathbf{H}(t) dx \overset{t \to \infty}{\longrightarrow} 0.
$$

Now a time average is estimated in (4.14). For this purpose let $\chi \in C_0^\infty((0,1),[0,\infty))$. Then Lemma 4.1 yields

$$
(4.16)\qquad \int_0^1 \chi(s) \int_{\mathbb{R}^3} \psi \mu \cdot (P_H \mathbf{H}(t+s)) \cdot P_H \tilde{\mathbf{M}}(t+s) dx ds \overset{t \to \infty}{\longrightarrow} 0,
$$

since $\mathbf{G} \overset{\text{def}}{=} P_H(\tilde{\mathbf{M}}(\cdot)) \in W^{1,\infty}((0,\infty), L^2(\mathbb{R}^3))$ by (3.1). It follows from (4.10), (4.12), and (4.14)–(4.16) that

$$
\begin{aligned}
(4.17)\qquad \limsup_{t \to \infty} \int_0^1 \chi(s) \int_{\mathbb{R}^3} \psi |\mu^{1/2} P_H \mathbf{H}(t+s)|^2 dx ds \\
\leq \limsup_{t \to \infty} \int_0^1 \chi'(t+s) g(t+s) ds = 0.
\end{aligned}
$$

Note that (4.10) and (4.17) do not immediately yield the assertion of Theorem 2.4. The aim of the following considerations is to show that the localized energy $E : [0,\infty) \to \mathbb{R}$ defined by

$$
(4.18)\qquad E(t) \overset{\text{def}}{=} \int_{\mathbb{R}^3} \psi [|\varepsilon^{1/2} \mathbf{E}(t)|^2 + |\mu^{1/2} P_H \mathbf{H}(t)|^2] dx
$$

is uniformly continuous on $[0, \infty)$. With curl $(1 - P_H)\mathbf{g} = 0$ for all $\mathbf{g} \in L^2(\mathbb{R}^3)$, it follows from (1.1) that

$$\partial_t \mathbf{E} = \varepsilon^{-1} \text{ curl } (P_H \mathbf{H}(\cdot)) + \mathbf{f}_1, \quad \partial_t (P_H \mathbf{H}(\cdot)) = -\mu^{-1} \text{ curl } \mathbf{E} + \mathbf{f}_2$$

on $\mathbb{R}^+ \times \mathbb{R}^3$ weakly in the sense of (2.15) with

$$\mathbf{f}_1(t) \stackrel{\text{def}}{=} -\varepsilon^{-1}\sigma \mathbf{E}(t) - \varepsilon^{-1}\mathbf{j}(t), \quad \mathbf{f}_2(t) \stackrel{\text{def}}{=} -P_H \partial_t \tilde{\mathbf{M}}(t).$$

Since $\mathbf{f} \stackrel{\text{def}}{=} (\mathbf{f}_1, \mathbf{f}_2) \in L^\infty((0,\infty), X)$ by Lemma 3.1, the standard energy estimate and (3.1) again yield

$$\begin{aligned} |E'(t)| &= \left| 2 \int_{\mathbb{R}^3} \psi \cdot [\varepsilon \mathbf{E}(t) \mathbf{f}_1(t) + \mu (P_H \mathbf{H}(t)) \mathbf{f}_2(t)] \, dx + 2 \int_{\mathbb{R}^3} (\nabla \psi) \wedge \mathbf{E}(t) \cdot P_H \mathbf{H}(t) dx \right| \\ &\leq C_6 \end{aligned}$$

with some $C_1 \in (0, \infty)$ independent of t. Hence, E satisfies

$$\sup_{s \in (0,a)} |E(t) - E(t+s)| \leq C_3 a \text{ for all } a > 0. \tag{4.19}$$

Now, choose $\chi \in C_0^\infty(0,1)$ with

$$\int_{\mathbb{R}} \chi dt = 1 \tag{4.20}$$

and let $\chi_m(t) \stackrel{\text{def}}{=} m\chi(mt)$ for $m \in \mathbb{N}$. Since $\int_0^1 \chi_m(t) dt = 1$ by (4.20) and supp $\chi_m \subset (0, 1/m)$, it follows that

$$\left| E(t) - \int_0^1 \chi_m(s) E(t+s) ds \right| \leq \sup_{s \in (0,1/m)} |E(t) - E(t+s)| \stackrel{m \to \infty}{\longrightarrow} 0 \tag{4.21}$$

uniformly with respect to t. On the other hand (4.10) and (4.17) yield

$$\int_0^1 \chi_m(s) E(t+s) ds \stackrel{t \to \infty}{\longrightarrow} 0 \text{ for all } m \in \mathbb{N}. \tag{4.22}$$

From (4.21) and (4.22) one obtains $E(t) \stackrel{t \to \infty}{\longrightarrow} 0$, which yields by (4.18)

$$\|\mathbf{E}(t)\|_{L^2(K)} + \|Q_H \mathbf{H}(t)\|_{L^2(K)} \stackrel{t \to \infty}{\longrightarrow} 0. \tag{4.23}$$

It remains to show

$$\|Q_0 \mathbf{H}(t)\|_{L^2(K)} \stackrel{t \to \infty}{\longrightarrow} 0.$$

Since $(1 - Q_H)\mathbf{H}(t) \in H_{curl,0}$, one has

$$\begin{aligned} \|\psi Q_0 \mathbf{H}(t)\|_{L^2(\mathbb{R}^3)} &= \|\psi Q_0 Q_H \mathbf{H}(t)\|_{L^2(\mathbb{R}^3)} = \|\psi Q_0 Q_H \mathbf{H}(t)\|_{L^2(B_R(0))} \\ &\leq \|\psi Q_H \mathbf{H}(t)\|_{L^2(B_R(0))} + \|[\psi, Q_0] Q_H \mathbf{H}(t)\|_{L^2(B_R(0))}, \end{aligned} \tag{4.24}$$

where $R > 0$ with supp $\psi \subset B_R(0)$. By (2.19) assertion (4.23) holds for all compact subsets of U, in particular for supp (ψ). Hence

$$\|\psi Q_H \mathbf{H}(t)\|_{L^2(\mathbb{R}^3)} \stackrel{t \to \infty}{\longrightarrow} 0. \tag{4.25}$$

Since the commutator $[\psi, Q_0]$ is a compact operator from $L^2(\mathbb{R}^3)$ to $L^2(B_R(0))$, it follows from (2.16) that

$$\|[\psi, Q_0]Q_H \mathbf{H}(t)\|_{L^2(B_R(0))} \stackrel{t\to\infty}{\longrightarrow} 0. \tag{4.26}$$

Finally, the assertion follows from (4.24)–(4.26). $\square$

Remark 3. For all compact sets $K \subset \mathbb{R}^3$ satisfying assumption (2.19), one has by Theorem 2.4

$$\|\mathbf{H}(t) - \mathbf{H}^{(q)}(t)\|_{L^2(K)} = \|P_H \mathbf{H}(t)\|_{L^2(K)} \stackrel{t\to\infty}{\longrightarrow} 0, \tag{4.27}$$

where $\mathbf{H}^{(q)}(t) \in L^2(\mathbb{R}^3)$ is given by $\mathbf{H}^{(q)}(t) = (1 - P_H)\mathbf{H}(t) = -(1 - P_H)\tilde{\mathbf{M}}(t)$; i.e.,

$$\operatorname{curl} \mathbf{H}^{(q)}(t) = 0 \text{ and } \operatorname{div} (\mu[\mathbf{H}^{(q)}(t) + \tilde{\mathbf{M}}(t)]) = 0 \text{ on } \mathbb{R}^3.$$

This means that $\mathbf{H}^{(q)}$ is determined by the quasi-stationary approximation obtained in [5]. Therefore, (4.27) provides a further justification of the magnetostatic approximation in which $\mathbf{H}$ is replaced by $\mathbf{H}^{(q)}$.

REFERENCES

[1] F. Brown, *Micromagnetics*, Wiley-Interscience, New York, 1963.

[2] G. Carbou and P. Fabrie, *Time average in micromagnetism*, J. Differential Equations, 147 (1998), pp. 383–409.

[3] G. Carbou, P. Fabrie, and F. Jochmann, *A remark on the weak ω-limit set for micromagnetism equation*, Appl. Math. Lett., 15 (2002), pp. 95–99.

[4] F. Jochmann, *Asymptotic behaviour of solutions to a class of semi-linear hyperbolic systems in arbitrary domains*, J. Differential Equations, 160 (2000), pp. 439–466.

[5] F. Jochmann, *Existence of solutions and a quasi-stationary limit for a hyperbolic system describing ferromagnetism*, SIAM J. Math. Anal., 34 (2002), pp. 315–340.

[6] J. L. Joly, G. Metivier, and J. Rauch, *Global solutions to Maxwell equations in a ferromagnetic medium*, Ann. Henri Poincaré, 1 (2000), pp. 307–340.

[7] P. Joly, A. Komech, and O. Vacus, *On transitions to stationary states in a Maxwell–Landau–Lifschitz–Gilbert system*, SIAM J. Math. Anal., 31 (1999) pp. 346–374.

[8] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media*, Pergamon Press, Elmsford, NY, 1960.

[9] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[10] J. Rauch and M. Taylor, *Penetration into shadow regions and unique continuation properties in hyperbolic mixed problems*, Indiana Univ. Math. J., 22 (1972), pp. 277–285.

[11] A. Visintin, *On Landau-Lifschitz equation for ferromagnetism*, Japan J. Appl. Math., 2 (1985), pp. 69–84.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 291–301

EXISTENCE OF WEAK SOLUTIONS FOR THE MULLINS–SEKERKA FLOW*

MATTHIAS RÖGER†

Abstract. The long-time existence of solutions for the Mullins–Sekerka problem in a new weak formulation is proved. Using a variational approach introduced by Luckhaus and Sturzenhecker [*Calc. Var. Partial Differential Equations*, 3 (1995), pp. 253–271], time-discrete solutions are constructed, satisfying approximate Gibbs–Thomson laws in a BV-formulation. But since the passage to a limit allows a loss of surface area for the phase interfaces, convergence in this setting is in general not true. We consider the surface measure of the phase interfaces and use the theory of varifolds to obtain a rigorous passage to a limit in a suitable weak formulation of the Gibbs–Thomson law.

Key words. free boundaries, Mullins–Sekerka flow, varifolds

AMS subject classifications. 49Q20, 80A22

DOI. 10.1137/S0036141004439647

1. Introduction.

The Mullins–Sekerka flow is a variant of the classical Stefan problem and describes phase transitions, such as melting or solidification processes, where a negligible specific heat of the material under consideration can be assumed. In this situation the energy balance is expressed by a quasi-stationary parabolic equation. A geometric condition on the phase interface, known as Gibbs–Thomson law, accounts for surface tension effects. In contrast to the classical Stefan problem, the Mullins–Sekerka flow allows for superheating and undercooling, i.e., temperatures above the melting point in the solid phase or temperatures below the melting point in the liquid phase.

To state the problem we consider a given time interval $(0, T)$ and an open bounded region $\Omega \subset \mathbb{R}^3$ representing the body of the material. Denote $\Omega_T := (0, T) \times \Omega$. The state variables are the relative temperature

$$u : \Omega_T \to \mathbb{R}$$

($u = 0$ denoting the melting point) and a phase function

$$\mathcal{X} : \Omega_T \to \{0, 1\},$$

which partitions Ω at a time $t \in (0, T)$ into a *liquid phase* $\{\mathcal{X}(t, .) = 1\}$ and a *solid phase* $\{\mathcal{X}(t, .) = 0\}$ separated by the *phase interface*, that is, their common boundary in Ω. The energy balance for the classical Stefan problem reads

$$\partial_t(cu + L\mathcal{X}) - k\Delta u = f \tag{1.1}$$

in the sense of distributions, with a given heat source $f : \Omega \to \mathbb{R}$ and phase independent constants c, L, and k describing the specific heat, latent heat, and heat capacity,

*Received by the editors January 9, 2004; accepted for publication (in revised form) December 28, 2004; published electronically October 7, 2005. This work was supported by DFG Sonderforschungsbereich 611 and by the European Community's Human Potential Programme under contract HPRN-CT-2002-00274, FRONTS-SINGULARITIES.
http://www.siam.org/journals/sima/37-1/43964.html
†Department of Mathematics and Computing Science, Eindhoven University of Technology, Den Dolech 2, 5600 MB Eindhoven, The Netherlands (mroeger@win.tue.nl).

respectively. Assuming that c is negligibly small and setting for simplicity $L = k = 1$, we get the *quasi-stationary energy balance*

$$\partial_t \mathcal{X} - \Delta u = f. \tag{1.2}$$

In the classical Stefan problem the temperature at the phase interface equals the melting temperature, that is,

$$u(t,.) = 0 \quad \text{on the phase interface.} \tag{1.3}$$

Surface tension effects are taken into consideration by the *Gibbs–Thomson law*

$$H(t,.) = u(t,.) \quad \text{on the phase interface,} \tag{1.4}$$

where $H(t,.)$ denotes the scalar mean curvature of the phase interface, which we take as positive for convex liquid phases.

The Mullins–Sekerka problem is given by (1.2) and (1.4). An initial condition for $\mathcal{X}$ and a boundary condition for u on $\partial\Omega$ are prescribed. This model can be seen as a quasi-stationary variant of the *Stefan problem with Gibbs–Thomson law* given by (1.1) and (1.4). Existence of classical solutions for the Mullins–Sekerka problem locally in time was proved by Chen, Hong, and Li [CHY96] and by Escher and Simonett [ES97]. In general, classical solutions can develop singularities. To derive long-time existence results one has to turn to weak formulations. Chen [Che96] obtains solutions globally in time studying the limit of a certain Cahn–Hilliard model. Here the Gibbs–Thomson law is satisfied in a rather weak and complex formulation. In particular, the measures giving the energy density are not necessarily rectifiable.

Luckhaus and Sturzenhecker give in [LS95] another weak formulation of the Mullins–Sekerka problem. Phase and temperature function

$$\mathcal{X} \in \mathrm{L}^\infty(0,T;\mathrm{BV}(\Omega;\{0,1\})), \quad u \in \mathrm{L}^2(0,T;\mathrm{H}^{1,2}(\Omega))$$

satisfy (1.2) in the sense of distributions and (1.4) in the *BV-formulation of the Gibbs–Thomson law*, that is,

$$\int_0^T \int_\Omega \left(\nabla \cdot \xi - \frac{\nabla \mathcal{X}}{|\nabla \mathcal{X}|} \cdot D\xi \frac{\nabla \mathcal{X}}{|\nabla \mathcal{X}|} \right)(t,.)|\nabla \mathcal{X}|(t,.)\,dt = \int_{\Omega_T} \nabla \cdot (u\xi)\mathcal{X}$$

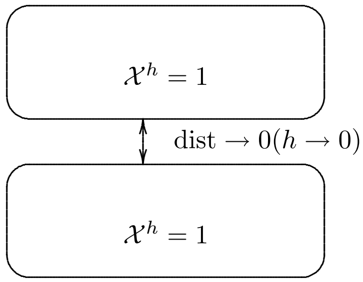
for all $\xi \in C_c^\infty(\Omega_T;\mathbb{R}^3)$.

Whereas this weak formulation is comparatively simple, the convergence of time-discrete approximations $\mathcal{X}^h$ and u^h to correct weak solutions of (1.2) and (1.4) is only shown under an additional condition on the approximations, which reads

$$\lim_{h\to 0} \int_{\Omega_T} |\nabla \mathcal{X}^h| \to \int_{\Omega_T} |\nabla \mathcal{X}|. \tag{1.5}$$

This excludes a loss of surface area for the phase interfaces and allows one to prove the convergence of approximate Gibbs–Thomson laws within the BV-formulation.

In the present paper we use the time-discrete approximation scheme of [LS95] but drop condition (1.5). Difficulties which arise are captured in the following time-independent example. Assume two solid parts of approximations $\mathcal{X}^h$ which merge to one when letting $h \to 0$.



A part of the boundary, indicated by the dashed line, has ceased to separate two different phases. We call this part the *hidden boundary*, whereas the *phase interface* represents the physically relevant part of the boundary. Cusp singularities occur due to the cancellation of phase interfaces. As shown in [Sch97], the BV-formulation of the Gibbs–Thomson law breaks down.

These difficulties are tackled in [Rög03] and [Rög04], where the Stefan problem with Gibbs–Thomson law is treated. Following an idea of Schätzle [Sch01] we consider the surface measure of the phase interfaces to master possible cancellations. In the above example the surface measures $|\nabla \mathcal{X}^h|$ converge with $h \to 0$ to a Radon measure that has double multiplicity on the hidden boundary. This suggests the use of the concept of integral varifolds introduced by Almgren [Alm65]. Geometric measure theory provides a notion of mean curvature for integral varifolds; Schätzle [Sch01] investigates the convergence of approximate Gibbs–Thomson equations in this context. However, the control about the hidden boundaries is quite weak and we have to focus on the physically relevant part of the boundary. For this purpose the following Proposition, which we have proved in an earlier work, will be crucial (for the notations, consult section 2).

Proposition 1.1 (see [Rög04, Proposition 3.1]). *Let $\Omega \subset \mathbb{R}^n$ be open, $E \subset \Omega$, and $\mathcal{X}_E \in \mathrm{BV}(\Omega)$. Assume that there are two integral $(n-1)$-varifolds μ_1, μ_2 on Ω such that for $i = 1, 2$ the following hold:*

$$\partial^* E \subset \operatorname{spt}(\mu_i), \tag{1.6}$$

$$\mu_i \text{ has locally bounded first variation with mean curvature vector } \vec{H}_{\mu_i}, \tag{1.7}$$

$$\vec{H}_{\mu_i} \in \mathrm{L}^s_{\mathrm{loc}}(\mu_i),\ s > n-1,\ s \geq 2. \tag{1.8}$$

Then

$$\vec{H}_{\mu_1}|_{\partial^* E} = \vec{H}_{\mu_2}|_{\partial^* E}$$

is satisfied $\mathcal{H}^{n-1}$-almost everywhere on $\partial^* E$.

This proposition justifies the following definition.

Definition 1.2. *Let $E \subset \Omega$ and $\mathcal{X}_E \in \mathrm{BV}(\Omega)$, and assume that there exists an integral $(n-1)$-varifold μ on Ω satisfying* (1.6)–(1.8). *Then we call*

$$\vec{H} := \vec{H}_\mu|_{\partial^* E}$$

the generalized mean curvature vector of $\partial^* E$ and define a scalar mean curvature by

$$H := \vec{H} \cdot \frac{\nabla \mathcal{X}}{|\nabla \mathcal{X}|} \text{ on } \partial^* E.$$

The essential boundary $\partial^* E$ represents the phase interface, whereas $\mathrm{spt}(\mu) \setminus \partial^* E$ can be seen as a hidden boundary. Proposition 1.1 shows that the varifold's mean curvature restricted to the phase interface is a property of the phase interface itself and independent of the location of hidden boundaries.

Our solutions of the Mullins–Sekerka problem satisfy the Gibbs–Thomson in the sense that for almost all times a generalized mean curvature for the phase interface exists and is given pointwise almost everywhere by (1.4).

THEOREM 1.3. *Let $\Omega \subset \mathbb{R}^3$ be an open bounded set with Lipschitz boundary, Γ_D a $\mathcal{H}^2$-measurable subset of $\partial\Omega$ with $\mathcal{H}^2(\Gamma_D) > 0$, and define $M_0 := \{v \in \mathrm{H}^{1,2}(\Omega) : v = 0$ on $\Gamma_D\}$. For given data*

$$\begin{aligned} \mathcal{X}_0 &\in \mathrm{BV}(\Omega; \{0,1\}), \\ u_D &\in \mathrm{H}^{1,2}(\Omega), \\ f &\in \mathrm{L}^2(\Omega), \end{aligned}$$

there exists functions

$$\begin{aligned} \mathcal{X} &\in \mathrm{L}^\infty(0,T; \mathrm{BV}(\Omega; \{0,1\})), \\ u &\in \mathrm{L}^2(0,T; u_D + M_0), \end{aligned}$$

such that

$$\int_{\Omega_T} \mathcal{X} \partial_t \varphi + \int_\Omega \mathcal{X}_0 \varphi(0,.) - \int_{\Omega_T} \nabla u \cdot \nabla \varphi = - \int_{\Omega_T} f \varphi \tag{1.9}$$

for all $\varphi \in C_c^\infty([0,T) \times \bar{\Omega})$, $\varphi = 0$ on Γ_D. For almost all $t \in (0,T)$ a generalized mean curvature $H(t)$ of $\partial^*\{\mathcal{X}(t,.) = 1\}$ exists and satisfies $\mathcal{H}^2$-almost everywhere on $\partial^*\{\mathcal{X}(t,.) = 1\}$:

$$H(t,.) = u(t,.). \tag{1.10}$$

As shown in [Rög04], (1.10) is well defined since for $u \in \mathrm{L}^2(0,T; \mathrm{H}^{1,2}(\Omega))$ trace values of $u(t,.)$ exist $\mathcal{H}^2$-almost everywhere for almost all $t \in (0,T)$. Furthermore, our formulation of the Gibbs–Thomson law generalizes the BV-formulation. Nevertheless we remark that this solution concept does not include a weak formulation of a boundary condition on the angle between the phase interface and the fixed boundary $\partial\Omega$ (see also the comments in [Rög03]). Moreover, since our focus is on the validation of the Gibbs–Thomson law, we have restricted our investigation to the case of a nonvanishing Dirichlet boundary. In the case of a pure Neumann boundary condition one has to take care of Lagrange multipliers occurring in the approximate Gibbs–Thomson laws. For a discussion of this topic we refer to [BGS98], where a multiphase Mullins–Sekerka system under a pure Neumann boundary condition is investigated.

The proof of Theorem 1.3 is given in sections 3 and 4. Compared to the Stefan problem with Gibbs–Thomson law, the technical difficulty is due to the degeneracy of the energy balance equation and the lack of $\mathrm{L}^1(\Omega_T)$-compactness for the approximate temperature functions.

2. General definitions and notations. We fix some notations and basic definitions. As a general reference for geometric measure theory, see the book by Simon [Sim83].

For functions depending on time and space variables we denote "∇" as the spatial gradient and "$\nabla\cdot$" as the spatial divergence. For a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^n$ and a k-dimensional subspace T of $\mathbb{R}^n$ we define the *divergence restricted to* T as

$$\operatorname{div}_T f(x) := \sum_{i=1}^{k} t_i \cdot Df(x)t_i,$$

where $\{t_i\}_{i=1,\dots,k}$ is any orthonormal basis of T.

For $\Omega \subset \mathbb{R}^n$ open, μ a Radon measure on Ω, $x \in \Omega$, and $\varrho > 0$, define the *scaled measures*

$$\mu_{x,\varrho}(A) := \varrho^{-n+1}\mu(x + \varrho A). \tag{2.1}$$

A k-dimensional subspace $P \subset \mathbb{R}^n$ is called the *k-dimensional tangential plane* of μ in x, denoted by $T_x\mu$, if there is $\theta > 0$ such that

$$\mu_{x,\varrho} \to \theta\mathcal{H}^k\lfloor P \text{ as Radon measures}$$

as ϱ tends to zero. In this case θ is the *multiplicity* of μ in x.

We call μ a *rectifiable* $(n-1)$*-varifold* if for μ-almost all $x \in \Omega$ the $(n-1)$-dimensional tangential plane $T_x\mu$ exists, and an *integral* $(n-1)$*-varifold* if in addition the multiplicity of μ is μ-almost everywhere integer-valued. A *general* $(n-1)$*-varifold* is a Radon measure on the Grassmannian $G^{n-1}\Omega$, which denotes the product of Ω and the space of $(n-1)$-dimensional subspaces of $\mathbb{R}^n$. In the present paper we identify a rectifiable $(n-1)$-varifold μ and the related Radon measure V_μ on $G^{n-1}\Omega$, defined by

$$V_\mu(\zeta) := \int_\Omega \zeta(x, T_x\mu)\, d\mu(x) \quad \text{for } \zeta \in C^0_c(G^{n-1}\Omega).$$

The *first variation* of a rectifiable $(n-1)$-varifold μ is given by

$$\delta\mu(\xi) := \int_\Omega \operatorname{div}_{T_x\mu} \xi(x)\, d\mu(x) \quad \text{for } \xi \in C^1_c(\Omega; \mathbb{R}^n).$$

We say that μ is of *locally bounded first variation* with *mean curvature vector* $\vec{H}_\mu$ if $\vec{H}_\mu \in \mathrm{L}^1_{loc}(\mu)$ and

$$\delta\mu(\xi) = \int_\Omega -\vec{H}_\mu \cdot \xi\, d\mu \quad \text{for all } \xi \in C^1_c(\Omega; \mathbb{R}^n).$$

For a $\mathcal{L}^n$-measurable set $E \subset \Omega$ of finite perimeter let $\partial^* E$ denote the reduced boundary of E in Ω, that is, the subset of $\partial E \cap \Omega$ where a generalized inner normal exists as Radon–Nikodým derivative $\nabla\mathcal{X}_E/|\nabla\mathcal{X}_E|$ with length one. Then $|\nabla\mathcal{X}_E| = \mathcal{H}^{n-1}\lfloor\partial^* E$ is an integral (with density 1) $(n-1)$-varifold on Ω (see [AFP00, paragraph 3.5]).

3. Time discretization. We use the scheme introduced by Luckhaus and Sturzenhecker [LS95] to construct time-discrete approximations. Given a time step $h > 0$, determine iteratively step functions in time

$$u^h : (0,T) \to (u_{\mathrm{D}} + M_0), \quad \mathcal{X}^h : (0,T) \to \mathrm{BV}(\Omega; \{0,1\})$$

by the following procedure. We use the abbreviations $\mathcal{X}_t^h = \mathcal{X}^h(t,.)$, $u_t^h = u^h(t,.)$ and set

$$\mathcal{X}_t^h := \mathcal{X}_0 \quad \text{for } 0 \le t < h.$$

For known $u_{t-h}^h, \mathcal{X}_{t-h}^h$ we define functionals $F_t^h : \mathrm{BV}(\Omega;\{0,1\}) \to \mathbb{R}$,

$$F_t^h(\mathcal{X}) := \int_\Omega |\nabla \mathcal{X}| + \frac{1}{2}\int_\Omega B^h(\mathcal{X} - \mathcal{X}_{t-h}^h - hf)(\mathcal{X} - \mathcal{X}_{t-h}^h - hf), \tag{3.1}$$

where for $v \in \mathrm{L}^2(\Omega)$ the function $B^h(v)$ denotes the solution of

$$-h\Delta B^h(v) = v, \qquad B^h(v)|_{\Gamma_D} = u_{\mathrm{D}}, \quad \nabla B^h(v)\cdot \nu_\Omega|_{\partial\Omega\setminus\Gamma_D} = 0. \tag{3.2}$$

Let $\mathcal{X}_t^h$ be a global minimizer of F_t^h and define

$$u_t^h := -B^h(\mathcal{X}_t^h - \mathcal{X}_{t-h}^h - hf). \tag{3.3}$$

(3.3) yields the approximate energy-balance

$$\partial_t^{-h}\mathcal{X}^h = \Delta u^h + f. \tag{3.4}$$

Since the first variation of F_t^h vanishes in $\mathcal{X}_t^h$, we get for all $\xi \in C_c^1(\Omega,\mathbb{R}^3)$

$$\int_\Omega \left(\nabla\cdot\xi - \frac{\nabla\mathcal{X}_t^h}{|\nabla\mathcal{X}_t^h|}\cdot D\xi\frac{\nabla\mathcal{X}_t^h}{|\nabla\mathcal{X}_t^h|}\right)|\nabla\mathcal{X}_t^h| = \int_\Omega \mathcal{X}_t^h\,\nabla\cdot(u_t^h\xi), \tag{3.5}$$

which is the BV-formulation of the approximate Gibbs–Thomson law

$$H_t^h = u_t^h,$$

where H_t^h denotes the mean curvature of $\partial^*\{\mathcal{X}_t^h = 1\}$.

To derive compactness properties for the approximate solutions we first observe $B^h(hv) = B^1(v)$ and

$$\begin{aligned}\int_\Omega B^h(v)v &= h\int_\Omega |\nabla B^h(v)|^2 - h\int_\Omega \nabla B^h(v)\cdot\nabla u_{\mathrm{D}} + \int_\Omega vu_{\mathrm{D}}\\ &\ge \frac{h}{2}\int_\Omega|\nabla B^h(v)|^2 - \frac{h}{2}\int_\Omega|\nabla u_{\mathrm{D}}|^2 + \int_\Omega vu_{\mathrm{D}}.\end{aligned}$$

With $F_\tau^h(\mathcal{X}_\tau^h) \le F_\tau^h(\mathcal{X}_{\tau-h}^h)$, we calculate for all $0<\tau<T$

$$\begin{aligned}&\int_\Omega|\nabla\mathcal{X}_\tau^h| + \frac{h}{4}\int_\Omega|\nabla u_\tau^h|^2 - \frac{h}{4}\int_\Omega|\nabla u_{\mathrm{D}}|^2 + \frac{1}{2}\int_\Omega(\mathcal{X}_\tau^h - \mathcal{X}_{\tau-h}^h - hf)u_{\mathrm{D}}\\ &\le \int_\Omega|\nabla\mathcal{X}_{\tau-h}^h| + \frac{h}{2}\int_\Omega B^1(f)f.\end{aligned} \tag{3.6}$$

Now, for any $t\in(0,T)$, $t\in[Mh,(M+1)h)$ with $M\in\mathbb{N}_0$, we take inequality (3.6) for $\tau_i = ih$ and sum over $i = 1,\dots,M$, which yields

$$\begin{aligned}&\int_\Omega|\nabla\mathcal{X}_t^h| + \frac{1}{4}\int_0^t\int_\Omega|\nabla u^h|^2\\ &\le \int_\Omega|\nabla\mathcal{X}_0| + \frac{T}{4}\int_\Omega|\nabla u_{\mathrm{D}}|^2 + \frac{1}{2}\int_\Omega|u_{\mathrm{D}}|(1+T|f|) + \frac{T}{2}\int_\Omega B^1(f)f.\end{aligned} \tag{3.7}$$

In particular we get

$$\sup_{h>0} \|\mathcal{X}^h\|_{\mathrm{L}^\infty(0,T;\mathrm{BV}(\Omega))} < \infty, \quad \sup_{h>0} \|u^h\|_{\mathrm{L}^2(0,T;\mathrm{H}^{1,2}(\Omega))} < \infty \tag{3.8}$$

and Fatou's lemma ensures

$$\liminf_{h\to 0} \Big(t \mapsto \|u^h(t,.)\|_{\mathrm{H}^{1,2}(\Omega)} \Big) \in \mathrm{L}^2(0,T). \tag{3.9}$$

In [LS95] an estimate for time differences of $\mathcal{X}^h$ is derived from (3.8); that is,

$$\int_\tau^T \int_\Omega |\mathcal{X}^h_t - \mathcal{X}^h_{t-\tau}| \, d\mathcal{L}^3 \, dt \le C\tau^{\frac{1}{4}} \tag{3.10}$$

holds for all $0 < \tau < T$. The Fréchet–Kolmogorov–Riesz theorem (see, for example, [DS88, IV.8, Theorem 21]), the weak*-compactness of $\mathrm{L}^2(0,T;\mathrm{H}^{1,2}(\Omega))$, and (3.8), (3.9), and (3.10) yield the following proposition.

PROPOSITION 3.1. *There exists a subsequence $h \to 0$ and functions $\mathcal{X} \in \mathrm{L}^\infty(0,T;\mathrm{BV}(\Omega))$, $u \in \mathrm{L}^2(0,T;u_D + M_0)$, such that*

$$u^h \rightharpoonup u \quad \textit{weakly in } \mathrm{L}^2(0,T;\mathrm{H}^{1,2}(\Omega)), \tag{3.11}$$

$$\mathcal{X}^h \to \mathcal{X} \quad \textit{in } \mathrm{L}^1(\Omega_T), \tag{3.12}$$

and for almost all $t \in (0,T)$ we have

$$\mathcal{X}^h(t) \to \mathcal{X}(t) \quad \textit{in } \mathrm{L}^1(\Omega), \tag{3.13}$$

$$\sup_{h>0} \|\mathcal{X}^h(t,.)\|_{\mathrm{BV}(\Omega)} < \infty, \tag{3.14}$$

$$\liminf_{h\to 0} \|u^h(t,.)\|_{\mathrm{H}^{1,2}(\Omega)} < \infty. \tag{3.15}$$

4. Convergence to solutions. Let $u, \mathcal{X}$ and a subsequence $h \to 0$ be given as in Proposition 3.1. To prove Theorem 1.3 we pass (3.4) and (3.5) to a limit and show $u, \mathcal{X}$ to be a correct weak solution of the Mullins–Sekerka problem. The energy balance is derived in a standard way. Use any $\varphi \in C_c^\infty([0,T) \times \bar{\Omega})$ with $\varphi|_{\Gamma_D} = 0$ as a test function in (3.4), perform a discrete partial integration, and use (3.11) and (3.12) to obtain (1.9).

The main effort is the passage to a limit in the approximate Gibbs–Thomson law (3.5). To use a convergence result of Schätzle [Sch01] we argue pointwise in time. Due to the lack of strong $\mathrm{L}^1(\Omega_T)$-compactness of the approximate temperatures, we have to consider any limit point of $(u^h(t,.))_{h>0}$ in the weak-$\mathrm{H}^{1,2}(\Omega)$ topology and identify their traces on $\partial^*\{\mathcal{X}(t,.) = 1\}$ with the trace of the weak limit u in (3.11).

Let us denote by μ^h_t the integral 2-varifolds with density one associated to the surface measure of the phase interfaces

$$\mu^h_t(\eta) := \int_\Omega \eta \, |\nabla \mathcal{X}^h_t| \qquad \text{for } \eta \in C^0_c(\Omega).$$

For the first variation of μ^h_t we obtain, recalling (3.5),

$$\delta\mu^h_t(\xi) = \int_\Omega \left(\nabla\cdot\xi - \frac{\nabla\mathcal{X}^h_t}{|\nabla\mathcal{X}^h_t|} \cdot D\xi \frac{\nabla\mathcal{X}^h_t}{|\nabla\mathcal{X}^h_t|} \right) |\nabla\mathcal{X}^h_t| = \int_\Omega \mathcal{X}^h_t \, \nabla\cdot(u^h_t \xi)$$

for all $\xi \in C^1_c(\Omega,\mathbb{R}^3)$. Finally, we define $T^h : \mathrm{L}^2(0,T;C^1_c(\Omega,\mathbb{R}^3))^* \to \mathbb{R}$ by

$$T^h(\xi) = \int_0^T \delta\mu^h_t(\xi(t,.))\,dt \tag{4.1}$$

and observe from (3.5) and (3.8) that

$$\|\delta\mu^h_t\|_{C^1_c(\Omega,\mathbb{R}^3)^*} \le C(\Omega)\|u^h_t\|_{H^{1,2}(\Omega)}, \tag{4.2}$$

$$\|T^h\|_{\mathrm{L}^2(0,T;C^1_c(\Omega,\mathbb{R}^3))^*} \le C'(\Omega), \tag{4.3}$$

where $C(\Omega), C'(\Omega)$ are independent of $h > 0$.

In a first step we will prove that the phase interfaces $\partial^*\{\mathcal{X}(t,.) = 1\}$ have a generalized mean curvature. Due to Proposition 1.1 this mean curvature is determined by the phase interface itself and thus given by the strong convergence in (3.12) and (3.13). Even more, any limit of the first variations $(\delta\mu^h_t)_{h>0}$ is determined by (3.12).

LEMMA 4.1. *For almost all $t \in (0,T)$ the phase interface $\partial^*\{\mathcal{X}(t,.) = 1\}$ has a generalized mean curvature $\vec{H}(t)$ in the sense of Definition* 1.2 *with*

$$\vec{H}(t) \in \mathrm{L}^4_{\mathrm{loc}}(|\nabla\mathcal{X}(t,.)|). \tag{4.4}$$

For any subsequence $(h_i)_{i\in\mathbb{N}}$, $h_i \to 0 (i \to \infty)$, with $\sup_{i\in\mathbb{N}} \|u^{h_i}(t,.)\|_{\mathrm{H}^{1,2}(\Omega)} < \infty$ *and for all $\xi \in C^1_c(\Omega,\mathbb{R}^3)$ we obtain*

$$\delta\mu^{h_i}_t(\xi) \;\to\; \int_\Omega -\vec{H}(t)\cdot\xi\,|\nabla\mathcal{X}(t,.)|. \tag{4.5}$$

Defining the functionals $T(t) \in C^1_c(\Omega,\mathbb{R}^3)^*$ by

$$\langle \xi, T(t)\rangle \;:=\; \int_\Omega -\vec{H}(t)\cdot\xi\,|\nabla\mathcal{X}(t,.)|, \tag{4.6}$$

we finally observe that

$$T \in \mathrm{L}^2(0,T;C^1_c(\Omega,\mathbb{R}^3))^* \tag{4.7}$$

holds.

Proof. Fix any $t \in (0,T)$ such that (3.13)–(3.15) hold and let $(h_i)_{i\in\mathbb{N}}$ be an arbitrary subsequence with

$$\sup_{i\in\mathbb{N}} \|u^{h_i}(t,.)\|_{\mathrm{H}^{1,2}(\Omega)} \;<\; \infty. \tag{4.8}$$

Define $V^h \in C^0_c(G^2\Omega)^*$ as the general varifold associated to the density-one integer varifold $|\nabla\mathcal{X}^h(t,.)|$. Recalling (3.14) and (4.8), there is a subsequence $(\tilde{h}_i)_{i\in\mathbb{N}}$, a function $v \in \mathrm{H}^{1,2}(\Omega)$, and a Radon measure $V \in C^0_c(G^2\Omega)^*$ such that

$$u^{\tilde{h}_i}(t,.) \rightharpoonup v \qquad \text{weakly in } \mathrm{H}^{1,2}(\Omega), \tag{4.9}$$

$$V^{\tilde{h}_i} \overset{*}{\rightharpoonup} V \qquad \text{weakly* in } C^0_c(G^2\Omega)^*. \tag{4.10}$$

With (3.5), (3.12), and (4.9), (4.10) all assumptions of Theorem 1.1 in [Sch01] are fulfilled and we obtain that

$$V \;=\; V_\mu \quad \text{for an integral} \quad 2\text{-varifold } \mu, \tag{4.11}$$

$$\mu \text{ has locally bounded first variation with mean curvature vector } \vec{H}_\mu, \tag{4.12}$$

$$\vec{H}_\mu \;\in\; \mathrm{L}^4_{\mathrm{loc}}(\mu), \tag{4.13}$$

and that

$$\vec{H}_\mu = v\nu(t,.) \tag{4.14}$$

holds μ-almost everywhere, with

$$\nu(t,.) = \begin{cases} \frac{\nabla\mathcal{X}(t,.)}{|\nabla\mathcal{X}(t,.)|} & \text{on } \partial^*\{\mathcal{X}(t,.)=1\}, \\ 0 & \text{elsewhere.} \end{cases}$$

According to Definition 1.2 the phase interface $\partial^*\{\mathcal{X}(t,.)=1\}$ has the generalized mean curvature vector $\vec{H}(t,.) = \vec{H}_\mu|_{\partial^*\{\mathcal{X}(t,.)=1\}}$ and (4.13) yields (4.4). Moreover, due to [Sch01, Theorem 1.2], we have

$$\vec{H}(t,.) = 0 \quad \mu\text{-almost everywhere in } \{\theta^2(\mu,.) \neq 1\}. \tag{4.15}$$

From (4.10) and (4.11) we obtain the convergence of $\mu_t^{\tilde{h}_i}$ to μ and therefore also the convergence of their first variations. Using (4.12), (4.14), and (4.15) we calculate

$$\begin{aligned} \lim_{i\to\infty} \delta\mu_t^{\tilde{h}_i}(\xi) &= \delta\mu(\xi) \\ &= \int_\Omega -\vec{H}_\mu \cdot \xi \, d\mu = \int_\Omega -\vec{H}(t,.) \cdot \xi \, |\nabla\mathcal{X}(t,.)| = \langle \xi, T(t,.)\rangle. \end{aligned}$$

Since $\vec{H}(t)$ is determined by $\mathcal{X} \in \mathrm{BV}(\Omega;\{0;1\})$ the functional $T(t)$ does not depend on the choices of v, V, and $(\tilde{h}_i)$. Thus we deduce

$$\delta\mu_t^{h_i} \overset{*}{\rightharpoonup} T(t)$$

for the whole sequence $(h_i)_{i\in\mathbb{N}}$ and for all subsequences of $h \to 0$ for which (4.8) holds. We choose $h_i \to 0$ such that

$$\lim_{i\to\infty} \|u^{h_i}(t,.)\|_{\mathrm{H}^{1,2}(\Omega)} = \liminf_{h\to 0} \|u^h(t,.)\|_{\mathrm{H}^{1,2}(\Omega)} < \infty$$

and the lower semicontinuity of the norm with respect to weak*-convergence, and thus (4.2) yields

$$\begin{aligned} \|T(t,.)\|_{C_c^1(\Omega;\mathbb{R}^3)^*} &\le \liminf_{i\to\infty} \|\delta\mu_t^{h_i}\|_{C_c^1(\Omega;\mathbb{R}^3)^*} \\ &\le \liminf_{i\to\infty} C(\Omega)\|u^{h_i}(t,.)\|_{\mathrm{H}^{1,2}(\Omega)} \\ &= C(\Omega) \liminf_{h\to 0} \|u^h(t,.)\|_{\mathrm{H}^{1,2}(\Omega)}. \end{aligned}$$

Recalling (3.9) we get $T \in \mathrm{L}^2(0,T;C_c^1(\Omega;\mathbb{R}^3))^*$. ∎

To prove that the mean curvature of the phase interface is given as trace of the weak limit u in (3.11), the weak*-convergence of T^h in $L^2(0,T;C_c^1(\Omega))^*$ will be crucial. The difficulty is that the pointwise convergence (4.5) holds only for time-dependent subsequences. We use an argument similar to a refined dominated convergence theorem in [PS93].

Lemma 4.2. *There is a subsequence $h \to 0$ such that*

$$T^h \overset{*}{\rightharpoonup} T \quad \text{in } \mathrm{L}^2(0,T;C_c^1(\Omega;\mathbb{R}^3))^*. \tag{4.16}$$

Proof. Fix an arbitrary $\xi \in \mathrm{L}^2(0,T;C^1_c(\Omega;\mathbb{R}^3))$. For $\alpha > 0$ we define functions $T^h_\alpha : (0,T) \to C^1_c(\Omega;\mathbb{R}^3)^*$:

$$T^h_\alpha(t) := \begin{cases} \delta\mu^h_t & \text{if} \quad \|u^h(t,.)\|_{\mathrm{H}^{1,2}(\Omega)} \le \alpha, \\ T(t) & \text{if} \quad \|u^h(t,.)\|_{\mathrm{H}^{1,2}(\Omega)} > \alpha. \end{cases} \tag{4.17}$$

For all $t \in (0,T)$, for which the assertions of Lemma 4.1 hold, the definition of T^h_α and (4.5) yield

$$\langle \xi(t,.), T^h_\alpha(t)\rangle \to \langle \xi(t,.), T(t)\rangle \quad (h \to 0).$$

The estimate

$$|\langle \xi(t,.), T^h_\alpha(t)\rangle| \le \|\xi(t,.)\|_{C^1_c(\Omega;\mathbb{R}^3)} \Big(C\alpha + \|T(t)\|_{C^1_c(\Omega;\mathbb{R}^3)^*}\Big)$$

and (4.7) give a $\mathrm{L}^1(0,T)$-dominator. Thus the Lebesgue theorem ensures

$$\int_0^T \langle \xi(t,.), T^h_\alpha(t)\rangle \, dt \to \int_0^T \langle \xi(t,.), T(t)\rangle \, dt \quad (h \to 0). \tag{4.18}$$

Next, consider the sets $A^h := \{t \in (0,T) \,:\, \|u^h(t,.)\|_{\mathrm{H}^{1,2}(\Omega)} > \alpha\}$ and observe

$$\begin{aligned} &\Big| \int_0^T \big\langle \xi(t,.), \delta\mu^h_t - T^h_\alpha(t)\big\rangle \, dt \Big| \\ \le &\int_{A^h} \Big| \big\langle \xi(t,.), \delta\mu^h_t - T(t)\big\rangle \Big| \, dt \\ \le &\Big(\int_{A^h} \|\xi(t,.)\|^2_{C^1_c(\Omega;\mathbb{R}^3)} dt\Big)^{\frac{1}{2}} \Big(\|T^h\|_{\mathrm{L}^2(0,T;C^1_c(\Omega;\mathbb{R}^3)^*)} + \|T\|_{\mathrm{L}^2(0,T;C^1_c(\Omega;\mathbb{R}^3)^*)}\Big). \end{aligned}$$

By (4.3) and (4.7) the norms $\|T^h\|_{\mathrm{L}^2(0,T;C^1_c(\Omega;\mathbb{R}^3)^*)}$ and $\|T\|_{\mathrm{L}^2(0,T;C^1_c(\Omega;\mathbb{R}^3)^*)}$ are bounded uniformly in $h > 0$. Estimating

$$|A^h| \le \frac{1}{\alpha^2} \|u^h\|^2_{\mathrm{L}^2(0,T;\mathrm{H}^{1,2}(\Omega))} \le \frac{1}{\alpha^2} C$$

by (3.8), we end up with

$$\int_0^T \langle \xi(t,.), T^h - T^h_\alpha(t)\rangle \, dt \to 0 \quad (\alpha \to \infty) \tag{4.19}$$

uniformly in $h > 0$. Putting this together with (4.18) proves the lemma. ∎

By establishing the Gibbs–Thomson law we finish the proof of Theorem 1.3.

LEMMA 4.3. *For all* $\xi \in \mathrm{L}^2(0,T;C^1_c(\Omega;\mathbb{R}^3))$ *we obtain*

$$\int_0^T \langle \xi(t,.), T(t)\rangle \, dt = \int_{\Omega_T} \mathcal{X} \, \nabla \cdot (u\xi).$$

In particular, for almost all $t \in (0,T)$,

$$H(t,.) = u(t,.)$$

holds $\mathcal{H}^2$*-almost everywhere on* $\partial^*\{\mathcal{X}(t,.) = 1\}$.

Proof. We get from (3.5), (3.11), (4.1) and Lemma 4.2

$$\begin{aligned}\int_{\Omega_T} \mathcal{X}\nabla\cdot(u\xi) &= \lim_{h\to 0}\int_{\Omega_T}\mathcal{X}^h\nabla\cdot(u^h\xi)\\ &= \lim_{h\to 0}\int_0^T \langle \xi(t,.),\delta\mu_t^h\rangle\,dt\\ &= \int_0^T \langle \xi(t,.),T(t)\rangle\,dt.\end{aligned}$$

Since no time derivative is involved we deduce that for almost all $t\in(0,T)$ and all $\xi\in C_c^1(\Omega;\mathbb{R}^3)$,

$$\int_\Omega \mathcal{X}(t,.)\nabla\cdot(u(t,.)\xi) = \langle \xi, T(t)\rangle$$

holds. The Gauss–Green theorem [EG92, Theorem 5.8.1] and (4.6) yield

$$\int_\Omega -u(t,.)\nu(t,.)\cdot\xi\,|\nabla\mathcal{X}(t,.)| = \int_\Omega -\vec{H}(t,.)\cdot\xi\,|\nabla\mathcal{X}(t,.)|,$$

with $\nu(t,.)=\nabla\mathcal{X}(t,.)/|\nabla\mathcal{X}(t,.)|$ on $\partial^*\{\mathcal{X}(t,.)=1\}$. This proves the Gibbs–Thomson law. ∎

Acknowledgment. I thank R. Schätzle for many fruitful discussions and for referring me to geometric measure theory.

REFERENCES

[AFP00] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Science Publications, Oxford, UK, 2000.

[Alm65] F. J. Almgren, *The Theory of Varifolds*, Princeton notes, Princeton, NJ, 1965.

[BGS98] L. Bronsard, H. Garcke, and B. Stoth, *A multi-phase Mullins-Sekerka system: Matched asymptotic expansions and an implicit time discretisation for the geometric evolution problem*, Proc. Roy. Soc. Edinburgh Sect. A, 128 (1998), pp. 481–506.

[Che96] X. Chen, *Global asymptotic limit of solutions of the Cahn-Hilliard equation*, J. Differential Geom., 44 (1996), pp. 262–311.

[CHY96] X. Chen, J. Hong, and F. Yi, *Existence, uniqueness, and regularity of classical solutions of the Mullins-Sekerka problem*, Comm. Partial Differential Equations, 21 (1996), pp. 1705–1727.

[DS88] N. Dunford and J. T. Schwartz, *Linear Operators. Part* I, John Wiley, New York, 1988.

[EG92] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[ES97] J. Escher and G. Simonett, *Classical solutions for Hele-Shaw models with surface tension*, Adv. Differential Equations, 2 (1997), pp. 619–642.

[LS95] S. Luckhaus and T. Sturzenhecker, *Implicit time discretization for the mean curvature flow equation*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 253–271.

[PS93] P. I. Plotnikov and V. N. Starovoitov, *The Stefan problem with surface tension as a limit of a phase field model*, Differential Equations, 29 (1993), pp. 395–404.

[Rög03] M. Röger, *Lösungen für das Stefan Problem mit Gibbs-Thomson Gesetz bei einer lokalen Minimierung*, Ph.D. thesis, Universität Bonn, 2003.

[Rög04] M. Röger, *Solutions for the Stefan problem with Gibbs-Thomson law by a local minimisation*, Interfaces Free Bound., 6 (2004), pp. 105–133.

[Sch97] R. Schätzle, *A counterexample for an approximation of the Gibbs–Thomson law*, Adv. Math. Sci. Appl., 7 (1997), pp. 25–36.

[Sch01] R. Schätzle, *Hypersurfaces with mean curvature given by an ambient Sobolev function*, J. Differential Geom., 58 (2001), pp. 371–420.

[Sim83] L. Simon, *Lectures on Geometric Measure Theory*, in Proceedings of the Centre for Mathematical Analysis, Vol. 3, Australian National University Canberra, 1983.

SIAM J. Math. Anal.
Vol. 37, No. 1, pp. 302–320

ANALYTICITY OF DIRICHLET–NEUMANN OPERATORS ON HÖLDER AND LIPSCHITZ DOMAINS*

BEI HU† AND DAVID P. NICHOLLS†

Abstract. In this paper we take up the question of analyticity properties of Dirichlet–Neumann operators with respect to boundary deformations. In two separate results, we show that if the deformation is sufficiently small and lies either in the class of $C^{1+\alpha}$ (any $\alpha > 0$) or Lipschitz functions, then the Dirichlet–Neumann operator is analytic with respect to this deformation. The proofs of both results utilize the "domain flattening" change of variables recently advocated by Nicholls and Reitich for the stable, high-order numerical simulation of Dirichlet–Neumann operators. We extend their analyticity results through the use of more specialized function spaces, and our new theorems are *optimal* in terms of boundary regularity. In the case of $C^{1+\alpha}$ boundary perturbations the underlying field also lies in the Hölder class $C^{1+\alpha}$ and the theorem follows by appealing to familiar Schauder theory arguments. In contrast, for Lipschitz deformations the field must lie in an L^p-based Sobolev space ($W^{1,p}$), so the relevant elliptic estimates come from Sobolev theory. Additionally, in the case of Lipschitz domains, the Dirichlet–Neumann operator must be reformulated weakly in order to accommodate the lack of regularity at the boundary which these Sobolev-class fields possess.

Key words. Dirichlet–Neumann operators, geometric perturbations, free-boundary problems, boundary value problems, Hölder regularity, Sobolev regularity

AMS subject classifications. 35J25, 31B20, 76B07, 78A45

DOI. 10.1137/S0036141004444810

1. Introduction. Many problems of fundamental importance in engineering and the sciences are posed in terms of partial differential equations formulated on irregular and/or moving boundaries. In many instances the differential equations are quite simple (linear and constant coefficient); however, the nonlinearity of the boundary conditions and/or the geometrical difficulties of the domain usually prevent analytic solution of these problems. Classical examples of such problems are the free-surface evolution of an ideal fluid [15], scattering of electromagnetic radiation from an irregular grating [2], and precipitate growth [13]. For these problems a simplification and reduction in dimension can be achieved by considering surface quantities and, if applicable, the shape of the boundary as fundamental variables. Then, if desired, bulk quantities can be recovered from these boundary measurements via appropriate integral formulas. In general this procedure is complicated by the necessity of normal derivatives of field quantities at the boundary. Therefore, Dirichlet–Neumann operators (DNOs), which deliver normal derivatives ("Neumann data") given boundary measurements ("Dirichlet data"), play a crucial role.

Among the many ways in which the DNO can be simulated numerically (e.g., boundary integrals/elements, finite differences, finite elements, etc.), methods based upon boundary perturbations are particularly appealing. These approaches view the shape of the domain as a (small) deformation of a separable geometry (e.g., disk, torus, infinite strip) and seek solutions as a Taylor series expanded in powers of this small parameter. Aside from being highly accurate within their domain of applicability,

*Received by the editors June 29, 2004; accepted for publication (in revised form) February 4, 2005; published electronically October 7, 2005.
http://www.siam.org/journals/sima/37-1/44481.html

†Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556 (Hu.1@nd.edu, Nicholls.2@nd.edu). The second author's research was supported by NSF grants DMS-0196452, DMS-0139822, and DMS-0406007.

a particularly appealing property of these methods is that, in contrast with most alternative approaches, the spatial dimension of the problem does not affect their implementation or performance. See [18, 19, 20] for a complete discussion of these issues and presentation of numerical results.

Since perturbation algorithms play such a crucial role in the study of DNOs we take up the mathematical question of their analyticity with respect to boundary perturbations, i.e., with respect to ε, which measures the *size* of the perturbation. The first results along these lines can be derived from the work of Calderón [4] and of Coifman and Meyer [6], who showed that if the upper boundary of a two-dimensional domain is a (one-dimensional) Lipschitz curve, then the DNO maps H^1 to L^2 and is analytic in ε (sufficiently small). Next, Craig, Schanz, and Sulem [10] showed that the DNO maps $W^{k+1,p}$ to $W^{k,p}$ for $k \geq 0$ and is analytic in ε (sufficiently small) for three-dimensional domains provided that the two-dimensional upper boundary is C^1; Craig and Nicholls [8] extended this result to general d dimensions ($(d-1)$–dimensional upper boundary) by the same techniques but, due to the application at hand, also required the boundary deformation to be in the class $W^{k+1,p}$ for $k \geq 0$.

These results are the most general to date but rely heavily on an implicit boundary integral formulation for the DNO which, from a numerical standpoint, undermines the computational advantages of boundary perturbation approaches. With this consideration in mind, Nicholls and Reitich studied analyticity through the transformed field expansion (TFE) approach [18, 19, 20]. While this method did not deliver the sharpest results from a theoretical standpoint (the boundary deformation was required to be in the class $C^{3/2+\delta}$ for any $\delta > 0$), it did produce a new, *stabilized*, high-order numerical procedure for the approximation of DNOs with all the advantages of boundary perturbation methods (e.g., ease of implementation, dimension independent performance) without the shortcomings of classical implementations (e.g., cancellations and high-order instability); please see [18, 19, 20] for a complete discussion and [21, 22] for recent advancements in the setting of acoustic and electromagnetic scattering applications. Finally, we mention the recent work of Buffoni [3] who, in the setting of an existence theory for two-dimensional traveling capillary-gravity waves, utilized the DNO in Zakharov's formulation [23] of surface water wave evolution. However, since other techniques prevailed, the analyticity of the DNO with respect to boundary perturbations was not used.

The goal of this paper is to show that the TFE approach can, in addition to providing a stabilized numerical approach, be used to realize the most general analyticity results possible (in terms of boundary regularity) in *arbitrary* dimension. Concerning smoothness of the boundary, this matches the theorems of Calderón [4] and Coifman and Meyer [6] in two dimensions. However, the underlying function spaces are quite different being based upon L^p-Sobolev spaces rather than L^2-Sobolev spaces. Our results extend those of Craig, Schanz, and Sulem [10] and Craig and Nicholls [8] in higher dimensions. Of course, our method can be extended to spaces with higher regularity if greater smoothness is assumed on the boundary deformation and Dirichlet data. We begin by showing that the TFE method analyzed with Schauder theory in Hölder spaces gives a simple and elegant analyticity theorem for surface deformations in the class $C^{1+\alpha}$ for any $\alpha > 0$. We then follow this analysis with a more involved calculation in $W^{k,p}$ spaces using Sobolev theory and demonstrate that, in fact, the regularity of the surface shape can be reduced to Lipschitz.

The paper is organized as follows: In section 2 we introduce the TFE change of variables and state our main results. In section 3.1 we work in the classical Hölder spaces via Schauder theory and conclude analyticity for boundary deformations of class $C^{1+\alpha}$ for any $\alpha > 0$, and we show that the DNO will map $C^{1+\alpha}$ Dirichlet data to C^{α} Neumann data and is uniformly analytic in ε. In section 3.2 we utilize the Sobolev theory of $W^{k,p}$ spaces and show that, in fact, the regularity of the boundary deformation can be reduced to Lipschitz in *any* spatial dimension; in this case, the DNO is analytic in ε and maps $W^{1-1/p,p}$ Dirichlet data to $W^{-1/p,p}$ Neumann data (see section 2 for the precise definition of $W^{-1/p,p}$). In Appendix A, we review the key elliptic estimates which enable our analysis of the DNO.

2. Problem statement and change of variables. To focus upon a particular problem we consider the classical free-boundary problem of the evolution of a d-dimensional ideal fluid under the effects of gravity. The fluid sits above the bottom of a flat ocean bed at mean depth h and is bounded above by the free surface $\eta(x,t)$, giving the domain

$$S_{h,\eta} = \{(x,y) \in \mathbf{R}^{d-1} \times \mathbf{R} \mid -h < y < \eta\}.$$

The fundamental variables for this problem are the shape of the free surface, η, and the velocity potential $\varphi(x,y,t)$ which gives the velocity of the fluid from $\vec{v} = \nabla\varphi$. The equations of motion are [15]

$$\Delta\varphi = 0 \qquad \text{in } S_{h,\eta}, \tag{2.1a}$$

$$\partial_y \varphi(x,-h) = 0, \tag{2.1b}$$

$$\partial_t \eta + \nabla_x \varphi \cdot \nabla_x \eta - \partial_y \varphi = 0 \qquad \text{at } y = \eta, \tag{2.1c}$$

$$\partial_t \varphi + \frac{1}{2}|\nabla\varphi|^2 + g\eta = 0 \qquad \text{at } y = \eta. \tag{2.1d}$$

These equations must be supplemented with initial conditions and lateral boundary conditions, which we discuss later.

In a fundamental paper on stability of free-surface ocean waves, Zakharov [23] noted that the Euler equations, (2.1), could be stated as a Hamiltonian system in terms of the canonical variables $(\eta(x,t), \xi(x,t) \equiv \varphi(x,\eta(x,t),t))$. This observation, coupled with the solvability of Laplace's equation on the domain $S_{h,\eta}$ given ξ, leads to the realization that (2.1) can be equivalently stated *at the surface* of the domain $S_{h,\eta}$. The restatement was first made by Craig and Sulem [11] as

$$\partial_t \eta = G(\eta)\,\xi, \tag{2.2a}$$

$$\partial_t \xi = -g\eta - \frac{1}{2(1+|\nabla_x\eta|^2)}\Big[|\nabla_x\xi|^2 - (G(\eta)\,\xi)^2$$

$$-2(G(\eta)\,\xi)\nabla_x\xi\cdot\nabla_x\eta + |\nabla_x\xi|^2|\nabla_x\eta|^2 - (\nabla_x\xi\cdot\nabla_x\eta)^2\Big], \tag{2.2b}$$

where $G(\eta)\,\xi$ is the DNO. This set of equations, (2.2), has been useful in a variety of analytical [8, 7] and numerical [16, 17, 9, 14] treatments of the Euler equations, and clearly a detailed understanding of the DNO is at the heart of these analyses.

Inspired by the geometry of the Euler equations (2.1) and the reduction of Craig and Sulem, we study the DNO, $G(\eta)$, and its associated boundary value problem:

$$\Delta v(x,y) = 0 \qquad \text{in } S_{h,\eta}, \tag{2.3a}$$

$$\partial_y v(x,-h) = 0, \tag{2.3b}$$

$$v(x,\eta(x)) = \xi(x). \tag{2.3c}$$

Upon the solution of (2.3) the DNO is defined as

$$G(\eta)\,\xi = \nabla v|_{y=\eta} \cdot N_\eta = [-\nabla_x \eta \cdot \nabla_x v + \partial_y v]_{y=\eta}\,, \tag{2.4}$$

where the normal $N = (-\nabla_x \eta, 1)^T$ (not of unit length) is chosen to simplify the restatement of the kinematic condition (2.1c) as (2.2a). Regarding lateral boundary conditions, it is well known that bounded solutions to (2.3) are unique. Thus, $v(x, y)$ is periodic in x if $\eta(x)$ and $\xi(x)$ are periodic in x; similarly, the behavior of $v(x)$ as $x \to \pm\infty$ will be uniquely determined by the behavior of $\xi(x)$ near infinity. In this way we incorporate quite general boundary conditions into the definition of the DNO.

In order to work with more general Lipschitz boundaries, we now derive a weak formulation of the DNO: Take any test function $\psi \in T^1_R(\overline{S_{h,\eta}})$, where

$$T^1_R(\overline{S_{h,\eta}}) = \left\{ f \in C^1(\overline{S_{h,\eta}}) \mid f = 0 \text{ on } \{|x| > R\} \text{ for some large } R \right\}.$$

Then

$$\begin{aligned}
0 &= \int_{S_{h,\eta}} (\Delta v)\psi \, \mathrm{d}V \\
&= \int_{y=\eta(x)} (\partial_\nu v)\psi \, \mathrm{d}S - \int_{S_{h,\eta}} (\nabla_x v \cdot \nabla_x \psi + \partial_y v \ \partial_y \psi) \, \mathrm{d}V \\
&= \int_{\mathbf{R}^{d-1}} \frac{G(\eta)\ \xi}{\sqrt{1 + |\nabla_x \eta|^2}} \ \psi(x, \eta(x)) \ \sqrt{1 + |\nabla_x \eta|^2} \, \mathrm{d}x - \int_{S_{h,\eta}} (\nabla_x v \cdot \nabla_x \psi + \partial_y v \ \partial_y \psi) \, \mathrm{d}V.
\end{aligned}$$

Thus

$$\int_{\mathbf{R}^{d-1}} (G(\eta)\ \xi)\psi(x, \eta(x)) \, \mathrm{d}x = \int_{S_{h,\eta}} \nabla_x v \cdot \nabla_x \psi + \partial_y v \ \partial_y \psi \, \mathrm{d}V. \tag{2.5}$$

For any $\psi \in T^1_R(\overline{S_{h,\eta}})$ we can always approximate ψ with $\psi_j \in C^1(\overline{S_{h,\eta}})$ such that

$$\begin{aligned}
\psi_j &\to \psi && \text{strongly in } C(\overline{S_{h,\eta}}), \\
\nabla \psi_j &\to \nabla \psi && \text{weak* in } L^\infty(S_{h,\eta})^d.
\end{aligned}$$

Using this approximation, we find that (2.5) also extends to functions $\psi \in T^{0,1}_R(\overline{S_{h,\eta}})$, where

$$T^{0,1}_R(\overline{S_{h,\eta}}) = \left\{ f \in C^{0,1}(\overline{S_{h,\eta}}) \mid f = 0 \text{ on } \{|x| > R\} \text{ for some large } R \right\}.$$

Using the notation

$$\langle a, b \rangle = \int_{\mathbf{R}^{d-1}} a(x)\ b(x) \, \mathrm{d}x,$$

we restate (2.5) as follows: For any $\psi \in T^{0,1}_R(\overline{S_{h,\eta}})$,

$$\langle G(\eta)\ \xi, \psi(x, \eta(x)) \rangle = \int_{S_{h,\eta}} (\nabla_x v \cdot \nabla_x \psi + \partial_y v \ \partial_y \psi) \, \mathrm{d}V. \tag{2.6}$$

It is clear that the right-hand side of this equality requires v only to be $W^{1,1}_{loc}(\overline{S_{h,\eta}})$.

It has been discovered [18, 19, 20] that an effective technique for establishing the analyticity of DNOs is to make a "domain flattening" change of variables

$$x' = x, \qquad y' = h\frac{y-\eta}{h+\eta}, \tag{2.7}$$

which maps $S_{h,\eta}$ to $S_{h,0}$. Considering the transformed field

$$u(x', y') = v(x', (h+\eta)y'/h+\eta), \tag{2.8}$$

the change of variables induces the formulas

$$(h+\eta)\nabla_x = (h+\eta)\nabla_{x'} - (h+y')(\nabla_{x'}\eta)\partial_{y'}, \tag{2.9a}$$
$$(h+\eta)\ \mathrm{div}_x = (h+\eta)\ \mathrm{div}_{x'} - (h+y')(\nabla_{x'}\eta)\cdot\partial_{y'}, \tag{2.9b}$$
$$(h+\eta)\partial_y = h\partial_{y'}, \tag{2.9c}$$

which include a prefactor of $(h+\eta)$ in order to realize transformed equations with *no quotients involving* η. Upon making this transformation, (2.3) becomes

$$\Delta u(x', y') = F(x', y') \qquad \text{in } S_{h,0}, \tag{2.10a}$$
$$\partial_y u(x', -h) = 0, \tag{2.10b}$$
$$u(x', 0) = \xi(x'), \tag{2.10c}$$

where

$$F(x', y') = \mathrm{div}_{x'}\left[F^{(1)}(x', y')\right] + \partial_{y'}F^{(2)}(x', y') + F^{(3)}(x', y'). \tag{2.11}$$

The form for F can be found most easily from the following calculation:

$$\begin{aligned}
0 &= (h+\eta)^2\{\Delta_x v + \partial_y^2 v\} \\
&= (h+\eta)^2\Delta_x v + (h+\eta)^2\partial_y^2 v \\
&= (h+\eta)\mathrm{div}_x\left[(h+\eta)\nabla_x v\right] - \nabla_x\eta\cdot(h+\eta)\nabla_x v + (h+\eta)\partial_y\left[(h+\eta)\partial_y v\right].
\end{aligned}$$

Using (2.9) it is straightforward to show that

$$\begin{aligned}
0 = {} & h^2\Delta'_x u + h^2\partial_{y'}^2 u \\
& +\eta\ \mathrm{div}_{x'}\left[h\nabla_{x'}u\right] + h\ \mathrm{div}_{x'}\left[\eta\nabla_{x'}u\right] + \eta\ \mathrm{div}_{x'}\left[\eta\nabla_{x'}u\right] - h\ \mathrm{div}_{x'}\left[(h+y)\nabla_{x'}\eta\partial_{y'}u\right] \\
& -\eta\ \mathrm{div}_{x'}\left[(h+y)\nabla_{x'}\eta\partial_{y'}u\right] - (h+y)\nabla_{x'}\eta\cdot\partial_{y'}\left[h\nabla_{x'}u\right] \\
& -(h+y)\nabla_{x'}\eta\cdot\partial_{y'}\left[\eta\nabla_{x'}u\right] + (h+y)\nabla_{x'}\eta\cdot\partial_{y'}\left[(h+y')\nabla_{x'}\eta\partial_{y'}u\right] \\
& -h\nabla_{x'}\eta\cdot\nabla_{x'}u - \eta\nabla_{x'}\eta\cdot\nabla_{x'}u + (h+y')\left|\nabla_{x'}\eta\right|^2\partial_{y'}u.
\end{aligned}$$

From this point, several manipulations can be effected to realize the divergence structure of F. Upon dropping primes, this results in

$$F^{(1)} = -\frac{2}{h}\eta\nabla_x u - \frac{1}{h^2}\eta^2\nabla_x u + \frac{h+y}{h}\nabla_x\eta\partial_y u + \frac{(h+y)}{h^2}\eta\nabla_x\eta\partial_y u, \tag{2.12a}$$
$$F^{(2)} = \frac{h+y}{h}\nabla_x\eta\cdot\nabla_x u + \frac{(h+y)}{h^2}\eta\nabla_x\eta\cdot\nabla_x u - \frac{(h+y)^2}{h^2}\left|\nabla_x\eta\right|^2\partial_y u, \tag{2.12b}$$
$$F^{(3)} = \frac{1}{h}\nabla_x\eta\cdot\nabla_x u + \frac{1}{h^2}\eta\nabla_x\eta\cdot\nabla_x u - \frac{(h+y)}{h^2}\left|\nabla_x\eta\right|^2\partial_y u, \tag{2.12c}$$

where $F^{(1)}$, $F^{(2)}$, and $F^{(3)}$ are all $\mathcal{O}(\eta)$. At this point we note that, as claimed above, the right-hand side of (2.10a) contains no quotients involving η.

Of course, we are primarily concerned with the DNO; formula (2.4) transforms as

$$\begin{aligned}(h+\eta)G(\eta)\ \xi \\ &= \{-\nabla_x\eta\cdot(h+\eta)\nabla_x v+(h+\eta)\partial_y v\}\Big|_{y'=0} \\ &= \{h\partial_{y'}u-h\nabla_{x'}\eta\cdot\nabla_{x'}u-\eta\nabla_{x'}\eta\cdot\nabla_{x'}u+(h+y')\left|\nabla_{x'}\eta\right|^2\partial_{y'}u\}\Big|_{y'=0}.\end{aligned}\tag{2.13}$$

Therefore, again dropping primes,

$$G(\eta)\ \xi(x)=\partial_y u(x,0)+J(x),\tag{2.14}$$

where

$$\begin{aligned}J=&-\frac{\eta}{h}G(\eta)\ \xi-\nabla_x\eta\cdot\nabla_x u(x,0)\\ &-\frac{1}{h}\eta\nabla_x\eta\cdot\nabla_x u(x,0)+\left|\nabla_x\eta\right|^2\partial_y u(x,0),\end{aligned}$$

and clearly $J=\mathcal{O}(\eta)$. The weak statement of the DNO, (2.6), transforms as

$$\begin{aligned}&\langle G(\eta)\ \xi,\psi(x,0)\rangle\\ &=\int_{S_{h,0}}\left\{\left(\nabla_x u-\frac{h+y}{h+\eta}(\nabla_x\eta)\partial_y u\right)\cdot\left(\nabla_x\psi-\frac{h+y}{h+\eta}(\nabla_x\eta)\partial_y\psi\right)\right.\\ &\qquad\left.+\frac{h^2}{(h+\eta)^2}(\partial_y u)\partial_y\psi\right\}\frac{h+\eta}{h}\,\mathrm{d}V\end{aligned}\tag{2.15}$$

for any $\psi\in T_R^{0,1}(\overline{S_{h,0}})$. Finally, we point out that sometimes it is more convenient to write the DNO in the following form:

$$G(\eta)\ \xi(x)=\ -\nabla_x\eta\cdot\nabla_x\xi+\left.\frac{h(1+\left|\nabla_x\eta\right|^2)}{h+\eta}\partial_y u\right|_{y=0},\tag{2.16}$$

where we have used the fact that $u(x,0)=\xi(x)$.

In the spirit of the boundary perturbation methods we alluded to in the Introduction, we now suppose that we are considering η to be a *small* perturbation of a flat geometry, i.e., $\eta(x)=\varepsilon f(x)$. In this case, for future reference, (2.16) becomes

$$G(\eta)\ \xi(x)=\ -\varepsilon\nabla_x f\cdot\nabla_x\xi+\left.\frac{h(1+\left|\varepsilon\nabla_x f\right|^2)}{h+\varepsilon f}\partial_y u\right|_{y=0}.\tag{2.17}$$

We show the following theorem in section 3.1.

THEOREM 2.1. *Let $f,\xi\in C^{1+\alpha}(\mathbf{R}^{d-1})$, $0<\alpha<1$. Let $v(x,y)$ be the solution of* (2.3) *in the region $S_{h,\eta}$ with $\eta=\varepsilon f$ and define*

$$u(x,y,\varepsilon)=v\left(x,\frac{(h+\varepsilon f)y}{h}+\varepsilon f\right),\quad -\infty<x<\infty,-h<y<0$$

(cf. (2.8)*). Define the DNO $G(\varepsilon f)$ by* (2.14) *with $\eta=\varepsilon f$. Then both the solution $u(x,y,\varepsilon)$ and the DNO $G(\varepsilon f)$ are analytic as functions of ε; i.e., they can be expressed*

as the convergent series

$$u(x,y,\varepsilon)=\sum_{n=0}^{\infty}u_n(x,y)\ \varepsilon^n,\quad G(\varepsilon f)=\sum_{n=0}^{\infty}G_n(f)\ \varepsilon^n, \tag{2.18}$$

for small ε, where u_n and $G_n(f)$ satisfy, for some constants B and C independent of ε,

$$|u_n|_{C^{1+\alpha}(S_{h,0})}\le CB^n\,|\xi|_{C^{1+\alpha}(\mathbf{R}^{d-1})}\,,\quad |G_n(f)|_{\mathcal{L}(C^{1+\alpha}(\mathbf{R}^{d-1}),C^{\alpha}(\mathbf{R}^{d-1}))}\le CB^n.$$

This theorem implies that the DNO maps $C^{1+\alpha}$ Dirichlet data to C^α Neumann data.

By working in L^p-based Sobolev spaces, $W^{k,p}$ $(p>d)$, we can refine this result by requiring the boundary to be only Lipschitz continuous. In dealing with these Sobolev spaces, we must appeal to the trace operator and its mapping properties (see [1, Chapter 7 (e.g., Theorem 7.53)] for trace theorems); in particular, if $\partial\Omega\in C^k$, then the trace operator $W^{k,p}(\Omega)\to W^{k-1/p,p}(\partial\Omega)$ is continuous and surjective.

To state the next result with complete accuracy we first define a pair of function spaces. We denote by $B_r(x^*)$ the ball of radius r centered at x^*, and for $p>1$ define

$$X^p=\{\xi\mid \xi\in W^{1-1/p,p}(B_1(x^*))\text{ for any }x^*\in\mathbf{R}^{d-1}\}.$$

For $\xi\in X^p$ we define

$$\|\xi\|_{X^p}=\sup_{x^*\in\mathbf{R}^{d-1}}\|\xi\|_{W^{1-1/p,p}(B_1(x^*))}.$$

Recall that [1, Chapter 7]

$$\|\xi\|_{W^{1-1/p,p}(B_1(x^*))}=\inf\|\zeta\|_{W^{1,p}(B_1(x^*)\times[-h,0])}\,,$$

where the infimum is taken over all functions $\zeta\in W^{1,p}(B_1(x^*)\times[-h,0])$ such that $\zeta(x,0)=\xi(x)$ in the trace sense; i.e., for any C^∞ function $\gamma(x,y)$ such that $\gamma=0$ on

$$\{\partial B_1(x^*)\times[-h,0]\}\cup\{B_1(x^*)\times\{y=-h\}\}\,,$$

$\gamma(\zeta-\xi)\in W_0^{1,p}((B_1(x^*)\times(-h,0)))$. It is clear that with this definition

$$\|\xi\|_{X^p}\le\sup_{x^*\in\mathbf{R}^{d-1}}\|\xi\|_{W^{1-1/p,p}(B_2(x^*))}\le 2^{(d-1)/p}\,\|\xi\|_{X^p}\,.$$

We also define

$$Y^{k,p}=\{u\mid u\in W^{k,p}(B_1(x^*)\times[-h,0])\text{ for any }x^*\in\mathbf{R}^{d-1}\}$$

and

$$\|u\|_{Y^{k,p}}=\sup_{x^*\in\mathbf{R}^{d-1}}\|u\|_{W^{k,p}(B_1(x^*)\times[-h,0])}\,.$$

In the case of boundary data in X^p, the solution $u(x,y,\varepsilon)$ will only be in the space $W^{1,p}$ in the domain. Therefore, the first order derivative ∇u will only be an L^p function in the domain and the trace operator in (2.14) is not well defined. Thus we shall use the weak formulation (2.15). Since the DNO is local in nature, we shall discuss the DNO only in a neighborhood of an arbitrarily fixed point $\hat{x} \in \mathbf{R}^{d-1}$. We will establish the following in section 3.2.

THEOREM 2.2. *If $f \in C^{0,1}(\mathbf{R}^{d-1})$, $\xi \in X^p$, $p > d$. Let $v(x,y)$ be the solution of* (2.3) *in the region $S_{h,\eta}$ with $\eta = \varepsilon f$ and define*

$$u(x,y,\varepsilon) = v\left(x, \frac{(h+\varepsilon f)y}{h} + \varepsilon f\right), \quad -\infty < x < \infty, -h < y < 0$$

(cf. (2.8)*). Define the DNO $G(\varepsilon f)$ by* (2.15) *with $\eta = \varepsilon f$. Then both the solution $u(x,y,\varepsilon)$ and the DNO $G(\varepsilon f)$ are analytic as functions of ε; i.e., they can be expressed as the convergent series*

$$u(x,y,\varepsilon) = \sum_{n=0}^{\infty} u_n(x,y)\ \varepsilon^n, \quad G(\varepsilon f) = \sum_{n=0}^{\infty} G_n(f)\ \varepsilon^n,$$

for small ε, where u_n and $G_n(f)$ satisfy, for some constants B and C independent of ε,

$$\|u_n\|_{Y^{1,p}} \le CB^n \|\xi\|_{X^p}, \quad \|G_n(f)\|_{\mathcal{L}(X^p,(X_c^q(\hat{x}))^*)} \le CB^n \tag{2.19}$$

for any fixed $\hat{x} \in \mathbf{R}^{d-1}$. In these formulas, q is the conjugate of p, i.e., $q = p/(p-1)$, and $(X_c^q(\hat{x}))^*$ is the dual space of $X_c^q(\hat{x})$:

$$\begin{aligned} X_c^q(\hat{x}) &= \{\varphi \in X^q \mid \varphi = 0 \text{ for } |x - \hat{x}| > 1\} \\ &\cong W_0^{1-1/q,q}(B_1(\hat{x})). \end{aligned}$$

Remark. Roughly speaking, X^p behaves like $W^{1-1/p,p}$ and X^q behaves like $W^{1-1/q,q}$. Thus, the dual space of $W^{1-1/q,q}$ behaves locally like $W^{-(1-1/q),p} = W^{-1/p,p}$. Therefore, the above theorem states that the DNO "loses one spatial derivative" and is analytic in ε. This is the *optimal* regularity that one can expect for the DNO.

Remark. Theorem 2.2 concerns a field, v, in $W^{1,p}$ with boundary trace, ξ, in $W^{1-1/p,p}$. Such assumptions were made to enable a proof which demands the *weakest* possible regularity on the boundary perturbation. Of course, if the boundary deformation and Dirichlet data are more regular, then the field and DNO will be smoother as well. Results mentioned in the Introduction (e.g., Calderón [4], Coifman and Meyer [6], Craig, Schanz, and Sulem [10], Craig and Nicholls [8], and Nicholls and Reitich [18, 20]) provide such results in a wide array of function spaces.

Remark. We introduced the spaces X^p and $Y^{k,p}$ in order to include quite general behavior at infinity. For instance, we can accommodate periodicity or convergence (at infinity) to a constant. If we specialize to periodic boundary conditions, say on the period cell $Q \subset \mathbf{R}^{d-1}$, we can simplify the statements of the theorem by replacing X^p with $W^{1-1/p,p}(Q)$, $Y^{1,p}$ with $W^{1,p}(Q \times [-h,0])$, and $B_1(\hat{x})$ with Q in Theorem 2.2.

Remark. Finally, a direct, "method of majorants" approach could be pursued to derive these results; cf. [18, 19, 20]. This would involve (for Theorem 2.1) inserting the expansions (2.18) into (2.10) and (2.14), finding equations satisfied by the u_n and G_n, and then estimating them directly in an appropriate function space. Since our purpose is to simply establish analyticity in ε (rather than *joint* analyticity in x, y, and ε; cf. [20]), we have found that a *complexification* approach greatly simplifies the argument while delivering the most general result possible.

3. Analyticity. In this section we establish analyticity of $u(x, y, \varepsilon)$ in ε via a complexification argument. Of course, in the original system (2.3) we cannot allow ε to be complex-valued as ε measures the magnitude of the (real) deformation of the domain. On the other hand, in the transformed system (2.10) ε has no such interpretation and we are free to allow $\varepsilon = \varepsilon_1 + \mathrm{i}\varepsilon_2 \in \mathbf{C}$ and to look for *complex* solutions, u. The advantage of this approach is the availability of the formulas of complex analysis which readily deliver analyticity provided that straightforward estimates are established. Once this is accomplished we may set $\varepsilon_2 = 0$ and obtain the series expansion for u which must be real-valued.

The complexification approach requires us to simply show that $u(x, y, \varepsilon)$ is differentiable in $\varepsilon = \varepsilon_1 + \mathrm{i}\varepsilon_2$ for $|\varepsilon|$ sufficiently small. To this end we define the finite difference operator as follows:

$$T_\delta[u](x, y, \varepsilon) = \frac{1}{\delta}[u(x, y, \varepsilon + \delta) - u(x, y, \varepsilon)], \quad \delta = \delta_1 + \mathrm{i}\delta_2.$$

A simple computation shows that

$$T_\delta[u \cdot w](x, y, \varepsilon) = T_\delta[u](x, y, \varepsilon) \cdot w(x, y, \varepsilon) + u(x, y, \varepsilon + \delta) \cdot T_\delta[w](x, y, \varepsilon). \tag{3.1}$$

In the next two subsections we show that $T_\delta[u](x, y, \varepsilon)$ converges as $\delta \to 0$, for ε in a small disk. This is done in Hölder spaces in section 3.1 and in $W^{k,p}$ spaces in section 3.2.

3.1. Hölder estimates. To begin this section we recall the following well-known algebra property of the space C^α.

LEMMA 3.1. *Let $0 \le \alpha \le 1$. For $f \in C^\alpha(\mathbf{R}^{d-1})$, $u \in C^\alpha(S_{h,0})$, the product $fu \in C^\alpha(S_{h,0})$, and*

$$|fu|_{C^\alpha} \le |f|_{C^\alpha} |u|_{C^\alpha}.$$

For convenience, we often use $C^{k+\alpha}$ to denote either $C^{k+\alpha}(\mathbf{R}^{d-1})$ or $C^{k+\alpha}(S_{h,0})$; the meaning should be clear from the context. Now, recalling that since $\varepsilon \in \mathbf{C}$, solutions u of (2.10) will generally be complex-valued (with the real and imaginary parts individually satisfying (2.10)), we establish the following lemma regarding existence and uniqueness of solutions.

LEMMA 3.2. *Given $f, \xi \in C^{1+\alpha}$ for any $\alpha \in (0, 1)$, there exists $c_0 > 0$ such that* (2.10) *(with the right-hand side of* (2.10) *given by* (2.12)*) has a unique solution $u \in C^{1+\alpha}$ for all ε in the disk $|\varepsilon| \le c_0$. Furthermore,*

$$|u|_{C^{1+\alpha}} \le C|\xi|_{C^{1+\alpha}}, \tag{3.2}$$

where the constant C is independent of ε.

Proof. The contraction mapping principle will be utilized. Consider the space

$$X = \{u \in C^{1+\alpha} \mid u(x, 0) = \xi, \partial_y u(x, -h) = 0\}$$

and the map Φ, defined by the following steps: For $u \in X$, compute $R(x, y) = F(x, y, u(x, y))$ from (2.11) and (2.12), and find the solution of

$$\begin{aligned} &\Delta w(x, y) = R(x, y) && \text{in } S_{h,0}, \\ &w(x, 0) = \xi(x), \\ &\partial_y w(x, -h) = 0, \end{aligned}$$

guaranteed by Theorem A.2. Setting $\eta = \varepsilon f$, we note that if $u \in C^{1+\alpha}$, then

$$
\begin{aligned}
\left|F^{(1)}\right|_{C^\alpha} &\le \frac{2}{h}\left|\varepsilon f \nabla_x u\right|_{C^\alpha} + \frac{1}{h^2}\left|\varepsilon^2 f^2 \nabla_x u\right|_{C^\alpha} + \left|\frac{h+y}{h}\varepsilon(\nabla_x f)\partial_y u\right|_{C^\alpha} \\
&\quad + \left|\frac{(h+y)}{h^2}\varepsilon^2 f(\nabla_x f)\partial_y u\right|_{C^\alpha} \\
&\le \frac{2|\varepsilon|}{h}\left|f\right|_{C^\alpha}\left|u\right|_{C^{1+\alpha}} + \frac{|\varepsilon|^2}{h^2}\left|f\right|_{C^\alpha}^2\left|u\right|_{C^{1+\alpha}} + \frac{Y|\varepsilon|}{h}\left|f\right|_{C^{1+\alpha}}\left|u\right|_{C^{1+\alpha}} \\
&\quad + \frac{Y|\varepsilon|^2}{h^2}\left|f\right|_{C^\alpha}\left|f\right|_{C^{1+\alpha}}\left|u\right|_{C^{1+\alpha}} \\
&\le |\varepsilon| K_{1,1}\left|f\right|_{C^{1+\alpha}}\left|u\right|_{C^{1+\alpha}} + |\varepsilon|^2 K_{1,2}\left|f\right|_{C^{1+\alpha}}^2\left|u\right|_{C^{1+\alpha}},
\end{aligned}
$$

where we have used Lemma 3.1, and Y is defined by

$$
\left|(h+y)u\right|_{C^\alpha} \le Y\left|u\right|_{C^\alpha}.
$$

Similarly, it can be shown that

$$
\begin{aligned}
\left|F^{(2)}\right|_{C^\alpha} &\le |\varepsilon| K_{2,1}\left|f\right|_{C^{1+\alpha}}\left|u\right|_{C^{1+\alpha}} + |\varepsilon|^2 K_{2,2}\left|f\right|_{C^{1+\alpha}}^2\left|u\right|_{C^{1+\alpha}}, \\
\left|F^{(3)}\right|_{L^\infty} &\le |\varepsilon| K_{3,1}\left|f\right|_{C^1}\left|u\right|_{C^1} + |\varepsilon|^2 K_{3,2}\left|f\right|_{C^1}^2\left|u\right|_{C^1},
\end{aligned}
$$

so that from (A.1) of Theorem A.2, $w \in C^{1+\alpha}$. Thus, $\Phi : X \to X$ defined by $w = \Phi u$ is well-defined.

Now, if we choose $u, \tilde{u} \in X$, this will generate $w, \tilde{w} \in X$, respectively. Furthermore,

$$
\begin{aligned}
\left|w - \tilde{w}\right|_{C^{1+\alpha}} &\le C_e\left[\left|R^{(1)} - \tilde{R}^{(1)}\right|_{C^\alpha} + \left|R^{(2)} - \tilde{R}^{(2)}\right|_{C^\alpha} + \left|R^{(3)} - \tilde{R}^{(3)}\right|_{L^\infty}\right] \\
&\le |\varepsilon| K_{4,1}\left|f\right|_{C^{1+\alpha}}\left|u - \widetilde{u}\right|_{C^{1+\alpha}} + |\varepsilon|^2 K_{4,2}\left|f\right|_{C^{1+\alpha}}^2\left|u - \widetilde{u}\right|_{C^{1+\alpha}} \\
&\le \gamma\left|u - \widetilde{u}\right|_{C^{1+\alpha}}
\end{aligned}
$$

for $\gamma < 1$ if

$$
|\varepsilon| \le c_0 \equiv \max\left\{\frac{\gamma}{2K_{4,1}\left|f\right|_{C^{1+\alpha}}}, \frac{\sqrt{\gamma}}{\sqrt{2K_{4,2}}\left|f\right|_{C^{1+\alpha}}}\right\}.
$$

Clearly the estimate is uniformly valid for all ε in the disk $|\varepsilon| \le c_0$, and thus the contraction mapping principle gives existence and uniqueness of solutions. Repeating the above estimation procedure we find that (3.2) is valid. ∎

We next establish differentiability of u in ε.

LEMMA 3.3. *By shrinking the constant c_0 in Lemma 3.2 if necessary, we have*

$$
\left|T_\delta[u]\right|_{C^{1+\alpha}} \le C \quad \textit{for } |\varepsilon| \le c_0, |\delta| \le c_0, \tag{3.3}
$$

where the constant C is independent of ε and δ.

Proof. We begin by applying the difference operator T_δ to (2.10) as follows:

$$
\Delta T_\delta[u] = \operatorname{div}_x\left[T_\delta[F^{(1)}]\right] + \partial_y T_\delta[F^{(2)}] + T_\delta[F^{(3)}]. \tag{3.4}
$$

The product rule (3.1) can be used to derive

$$
\begin{aligned}
&T_\delta[F^{(1)}] \\
&= -\frac{2}{h}\varepsilon f \nabla_x T_\delta[u](\varepsilon) - \frac{2}{h} f \nabla_x u(\varepsilon+\delta) - \frac{\varepsilon^2}{h^2} f^2 \nabla_x T_\delta[u](\varepsilon) - \frac{2\varepsilon}{h^2} f^2 \nabla_x u(\varepsilon+\delta) \\
&\quad -\frac{\delta}{h^2} f^2 \nabla_x u(\varepsilon+\delta) + \frac{\varepsilon(h+y)}{h} \nabla_x f \partial_y T_\delta[u](\varepsilon) + \frac{(h+y)}{h} \nabla_x f \partial_y u(\varepsilon+\delta) \\
&\quad +\frac{\varepsilon^2(h+y)}{h^2} f \nabla_x f \partial_y T_\delta[u](\varepsilon) + \frac{2\varepsilon(h+y)}{h^2} f \nabla_x f \partial_y u(\varepsilon+\delta) \\
&\quad +\frac{\delta(h+y)}{h^2} f \nabla_x f \partial_y u(\varepsilon+\delta).
\end{aligned} \tag{3.5}
$$

Estimating this in C^α we find

$$
\begin{aligned}
\left|T_\delta[F^{(1)}]\right|_{C^\alpha} \leq& \left\{K_{5,1}|\varepsilon|\,|f|_{C^{1+\alpha}} + K_{5,2}|\varepsilon|^2\,|f|^2_{C^{1+\alpha}}\right\} |T_\delta[u](\varepsilon)|_{C^{1+\alpha}} \\
&+K_6\left\{|f|_{C^{1+\alpha}} + |\varepsilon|\,|f|_{C^{1+\alpha}} + |\delta|\,|f|_{C^{1+\alpha}}\right\} |u(\cdot,\cdot,\varepsilon+\delta)|_{C^{1+\alpha}}\,;
\end{aligned}
$$

similar expressions for $\left|T_\delta[F^{(2)}]\right|_{C^\alpha}$ and $\left|T_\delta[F^{(3)}]\right|_{L^\infty}$ can be found. These results coupled with Theorem A.2 imply that

$$
\begin{aligned}
|T_\delta[u]|_{C^{1+\alpha}} &\leq C_e\left[\left|T_\delta[F^{(1)}]\right|_{C^\alpha} + \left|T_\delta[F^{(2)}]\right|_{C^\alpha} + \left|T_\delta[F^{(3)}]\right|_{L^\infty}\right] \\
&\leq C_e\left[\left\{K_{7,1}\,|\varepsilon|\,|f|_{C^{1+\alpha}} + K_{7,2}\,|\varepsilon|^2\,|f|^2_{C^{1+\alpha}}\right\}|T_\delta[u]|_{C^{1+\alpha}}\right. \\
&\qquad \left. +K_8\left\{|f|_{C^{1+\alpha}} + |\varepsilon|\,|f|^2_{C^{1+\alpha}} + |\delta|\,|f|^2_{C^{1+\alpha}}\right\}|u|_{C^{1+\alpha}}\right].
\end{aligned}
$$

Clearly, if ε and δ are chosen sufficiently small, then $|T_\delta[u]|_{C^{1+\alpha}}$ is bounded independently of ε and δ. □

In the next step we show that this difference quotient converges to the derivative of u with respect to ε.

LEMMA 3.4. *There exists a small positive constant c_0 such that, in the disk $\{|\varepsilon| \leq c_0\}$, the complexified solution u of* (2.10) *is differentiable in the complex variable ε in the space $C^{1+\beta}$, for any $\beta \in (0,\alpha)$, i.e.,*

$$T_\delta[u] \to \partial_\varepsilon u \quad \textit{as } |\delta| \to 0.$$

Proof. For any $\beta \in (0,\alpha)$, we can use the compactness of $C^{1+\alpha}$ to conclude that there exists a subsequence $\delta_n \to 0$ such that

$$T_{\delta_n}[u] \to w \quad \text{in } C^{1+\beta}(\overline{S_{0,h}} \cap \{|x| \leq K\})$$

for any $K > 1$. By passing δ_n to 0 in the equation, we find that w satisfies

$$\Delta w = \operatorname{div}_x\left[H^{(1)}\right] + \partial_y H^{(2)} + H^{(3)},$$

where

$$
\begin{aligned}
H^{(1)} = &-\frac{2}{h}\varepsilon f \nabla_x w - \frac{2}{h} f \nabla_x u - \frac{\varepsilon^2}{h^2} f^2 \nabla_x w - \frac{2\varepsilon}{h^2} f^2 \nabla_x u \\
&+\frac{\varepsilon(h+y)}{h} \nabla_x f \partial_y w + \frac{(h+y)}{h} \nabla_x f \partial_y u \\
&+\frac{\varepsilon^2(h+y)}{h^2} f \nabla_x f \partial_y w + \frac{2\varepsilon(h+y)}{h^2} f \nabla_x f \partial_y u,
\end{aligned}
$$

and similar expressions hold for $H^{(2)}$ and $H^{(3)}$. Similar to the proof of Lemma 3.2, the $C^{1+\beta}$ solution w to such a system is unique. This uniqueness implies that the convergence is independent of the subsequence of δ_n. $\square$

At this point we can prove Theorem 2.1.

Proof of Theorem 2.1. By Cauchy's formula, for $|\varepsilon| < c_0$,

$$u(x,y,\varepsilon) = \frac{1}{2\pi \mathrm{i}} \int_{|\zeta|=c_0} \frac{u(x,y,\zeta)}{\zeta - \varepsilon} \, \mathrm{d}\zeta = \sum_{n=0}^{\infty} u_n(x,y) \, \varepsilon^n,$$

where

$$u_n(x,y) = \frac{1}{2\pi i} \int_{|\zeta|=c_0} \frac{u(x,y,\zeta)}{\zeta^{n+1}} \, \mathrm{d}\zeta.$$

From this formula, we obtain the estimates on u_n from the estimates for u as follows:

$$|u_n|_{C^{1+\alpha}} \le \frac{1}{c_0^{n+1}} \max_{|\zeta|=c_0} |u(\cdot,\cdot,\zeta)|_{C^{1+\alpha}} \le CB^n \, |\xi|_{C^{1+\alpha}} \,,$$

where $B = 1/c_0$. Since $G(\varepsilon f) \; \xi(x)$ is expressed in terms of u and its first order derivatives (see (2.17)), we can extend $G(\varepsilon f) \; \xi(x)$ to complex ε. Using the (complex) analyticity of u in ε, we immediately have the differentiability of $G(\varepsilon f) \; \xi$ with respect to ε and

$$|G(\varepsilon f) \; \xi|_{C^\alpha} \le C \, |u|_{C^{1+\alpha}} \le C \, |\xi|_{C^{1+\alpha}} \,.$$

Thus, for $|\varepsilon| < c_0$,

$$G(\varepsilon f) \; \xi = \frac{1}{2\pi \mathrm{i}} \int_{|\zeta|=c_0} \frac{G(\zeta f) \; \xi}{\zeta - \varepsilon} \, \mathrm{d}\zeta = \sum_{n=0}^{\infty} (G_n(f) \; \xi) \; \varepsilon^n,$$

where

$$G_n(f) \; \xi = \frac{1}{2\pi \mathrm{i}} \int_{|\zeta|=c_0} \frac{G(\zeta f) \; \xi}{\zeta^{n+1}} \, \mathrm{d}\zeta.$$

From this, we obtain

$$|G_n(f) \; \xi|_{C^\alpha} \le \frac{1}{c_0^{n+1}} \max_{|\zeta|=c_0} |G(\zeta f) \; \xi|_{C^\alpha} \le \frac{C}{c_0^{n+1}} \max_{|\zeta|=c_0} |u(\cdot,\cdot,\zeta)|_{C^{1+\alpha}} \le CB^n \, |\xi|_{C^{1+\alpha}} \,.$$

This implies

$$|G_n(f)|_{\mathcal{L}(C^{1+\alpha}(\mathbf{R}^{d-1}), C^\alpha(\mathbf{R}^{d-1}))} \le CB^n.$$

The theorem is proved. $\square$

3.2. ^{1,} **estimates.** Using $W^{1,p}(S_{h,\eta})$ ($W^{1-1/p,p}$ on the boundary) estimates, we will extend the result of the previous section to Lipschitz boundaries; i.e., we will assume

$$f \in C^{0,1}, \quad \xi \in X^p \quad (p > d),$$

and approximate f and ξ by smooth functions where necessary. The key result which allows the estimation of Lipschitz boundaries is the following.

LEMMA 3.5. *For $f \in C^{0,1}$, $u \in Y^{0,p}$, the product $(\nabla_x f)u \in Y^{0,p}$, and*

$$\|(\nabla_x f)u\|_{Y^{0,p}} \leq |f|_{C^{0,1}} \|u\|_{Y^{0,p}} .$$

Given this result we can prove the following lemma.

LEMMA 3.6. *Given $f \in C^{0,1}$, $\xi \in X^p$ $(p > d)$, there exists $c_0 > 0$ such that* (2.10) *has a unique solution $u \in Y^{1,p}$ for all ε in the disk $|\varepsilon| \leq c_0$. Furthermore,*

$$\|u\|_{Y^{1,p}} \leq C \|\xi\|_{X^p} , \tag{3.6}$$

where the constant C is independent of ε.

Proof. The proof is the same as in Lemma 3.2, with the $C^{1+\alpha}$ Hölder estimate (Theorem A.2) replaced by the $W^{1,p}$ estimate (Theorem A.3) given in Appendix A. For instance, the key estimate which guaranteed the contraction property in Lemma 3.2 now reads

$$\begin{aligned}
\left\|F^{(1)}\right\|_{Y^{0,p}} &\leq \frac{2}{h} \|\varepsilon f \nabla_x u\|_{Y^{0,p}} + \frac{1}{h^2} \left\|\varepsilon^2 f^2 \nabla_x u\right\|_{Y^{0,p}} + \left\|\frac{h+y}{h} \varepsilon (\nabla_x f) \partial_y u\right\|_{Y^{0,p}} \\
&\quad + \left\|\frac{(h+y)}{h^2} \varepsilon^2 f (\nabla_x f) \partial_y u\right\|_{Y^{0,p}} \\
&\leq \frac{2|\varepsilon|}{h} |f|_{C^{0,1}} \|u\|_{Y^{1,p}} + \frac{|\varepsilon|^2}{h^2} |f|^2_{C^{0,1}} \|u\|_{Y^{1,p}} \\
&\quad + \frac{\tilde{Y}|\varepsilon|}{h} |f|_{C^{0,1}} \|u\|_{Y^{1,p}} + \frac{\tilde{Y}|\varepsilon|^2}{h^2} |f|^2_{C^{0,1}} \|u\|_{Y^{1,p}} \\
&\leq |\varepsilon| \tilde{K}_{1,1} |f|_{C^{0,1}} \|u\|_{Y^{1,p}} + |\varepsilon|^2 \tilde{K}_{1,2} |f|^2_{C^{0,1}} \|u\|_{Y^{1,p}} ,
\end{aligned}$$

where $\tilde{Y}$ is defined by

$$\|(h+y)u\|_{Y^{0,p}} \leq \tilde{Y} \|u\|_{Y^{0,p}} .$$

From this calculation, using Lemma 3.5, we see the *explicit* appearance of the Lipschitz norm on the boundary deformation $f(x)$. ☐

To establish the differentiability in complex ε, we apply the finite difference operator, $T_\delta[\cdot]$.

LEMMA 3.7. *By shrinking the positive constant c_0 in Lemma* 3.6 *if necessary, we have*

$$\|T_\delta[u]\|_{Y^{1,p}} \leq C \quad \textit{for } |\varepsilon| \leq c_0, |\delta| \leq c_0, \tag{3.7}$$

where the constant C is independent of ε.

Proof. Again, the proof is essentially the same as for Lemma 3.3 with the $C^{1+\alpha}$ Hölder estimate (Theorem A.2) replaced by the $W^{1,p}$ estimate (Theorem A.3) given in Appendix A. ☐

Now we are ready to establish the differentiability in complex ε.

LEMMA 3.8. *If $|\varepsilon| \leq c_0$ and u is the solution of* (2.10)*, then u is differentiable in ε as a complex function almost everywhere; i.e.,*

$$T_\delta[u] \to \partial_\varepsilon u \quad \textit{as } |\delta| \to 0.$$

Proof. The proof is similar to that of Lemma 3.4. However, since we no longer have compactness for the first order derivatives, the subsequential convergence as

$\delta_n \to 0$ must be replaced by the following:

$$T_{\delta_n}[u] \to w \quad \text{strongly in } C^0(\{|x|<K\}\times[-h,0]) \text{ for any } K>1,$$
$$(\nabla_x,\partial_y)T_{\delta_n}[u] \to (\nabla_x,\partial_y)w \quad \text{weakly in } [L^p(\{|x|<K\}\times[-h,0])]^d \text{ for any } K>1.$$

We note that $T_{\delta_n}[u]$ satisfies equation (3.4) from Lemma 3.3. The terms in (3.5) are all *linear* in $T_{\delta_n}[u]$ and its first order of derivatives; furthermore, all the coefficients are in L^∞ since we assume that f is Lipschitz continuous. These key facts allow us to use weak convergence to take the limit $\delta_n \to 0$. Thus we obtain the equation for w. The rest of the proof remains the same as that of Lemma 3.4. □

Proof of Theorem 2.2. The analyticity of $u(x,y,\varepsilon)$ in ε, and the corresponding estimate for u_n in $Y^{1,p}$, can be obtained in the same manner as in the proof of Theorem 2.1. However, the estimates on $G(\varepsilon f)$ must be modified since we are only permitted the weak formulation of the DNO in this case. It is clear that this weak formulation, (2.15), allows the complexification in ε. To use this definition, however, we have to show that (2.15) defines a DNO in the appropriate space also for complex ε. Namely, we have to show that the value on the right-hand side of (2.15) is *independent* of the way the function $\psi(x,0)$ is extended to $\mathbf{R}^{d-1}\times[-h,0]$.

Since

$$T_R^\infty(\overline{S_{h,0}}) = \big\{ f\in C^\infty(\overline{S_{h,0}}) \mid f=0 \text{ on } \{|x|>R\} \text{ for some large } R \big\}$$

is dense in $T_R^{0,1}(\overline{S_{h,0}})$, we only need to show that the right-hand side of (2.15) is independent of the extension for such ψ; namely, we need to show

$$\begin{aligned}\int_{S_{h,0}} \bigg\{ \Big(\nabla_x u - \frac{h+y}{h+\eta}(\nabla_x\eta)\partial_y u\Big)\cdot\Big(\nabla_x\psi - \frac{h+y}{h+\eta}(\nabla_x\eta)\partial_y\psi\Big) \\ + \frac{h^2}{(h+\eta)^2}(\partial_y u)\partial_y\psi\bigg\}\frac{h+\eta}{h}\,\mathrm{d}V = 0,\end{aligned} \tag{3.8}$$

for any $\psi\in C^\infty(\overline{S_{h,0}})$ such that $\psi(x,0)\equiv 0$ for all $x\in\mathbf{R}^{d-1}$ and $\psi(x,y)\equiv 0$ for $|x|>R$ for some $R>1$. Under our assumptions, all boundary terms vanish upon utilization of integration by parts in (3.8), so we can establish (3.8) by using the weak formulation of the complexified equation for u.

We next proceed to establish the estimates for $G(\varepsilon f)$. As in the proof of Theorem 2.1,

$$\langle G(\varepsilon f)\ \xi,\psi(x,0)\rangle = \frac{1}{2\pi\mathrm{i}}\int_{|\zeta|=c_0}\frac{\langle G(\zeta f)\ \xi,\psi(x,0)\rangle}{\zeta-\varepsilon}\,\mathrm{d}\zeta = \sum_{n=0}^\infty \langle G_n(f)\ \xi,\psi(x,0)\rangle\ \varepsilon^n,$$

where $\psi(x,0)\in C_c^{0,1}(\mathbf{R}^{d-1})$, and

$$\langle G_n(f)\ \xi,\varphi(x,0)\rangle = \frac{1}{2\pi\mathrm{i}}\int_{|\zeta|=c_0}\frac{\langle G(\zeta f)\ \xi,\varphi(x,0)\rangle}{\zeta^{n+1}}\,\mathrm{d}\zeta.$$

Thus the conclusion of our theorem will follow if we can establish the estimate

$$\|G(\varepsilon f)\|_{\mathcal{L}(X^p,(X_c^q(\hat{x}))^*)} \le C \tag{3.9}$$

for some C, independent of ε, and for all $|\varepsilon|\le c_0$.

For any $\Psi \in X_c^q(\hat{x})$, we use the definition of $X_c^q(\hat{x})$ to extend Ψ to a function $\psi \in W_{loc}^{1,q}(\mathbf{R}^{d-1} \times [-h, 0])$ such that

$$\psi(x,0) = \Psi(x) \quad \text{in the trace sense,} \tag{3.10a}$$

$$\psi(x,y) = 0 \quad \text{for } |x-\hat{x}| > 1, \quad -h < y < 0, \tag{3.10b}$$

$$\|\psi\|_{W^{1,q}(B_1(\hat{x})\times(-h,0))} \le C \|\Psi\|_{X^q}. \tag{3.10c}$$

Since we have already established a $W^{1,p}(B_1(\hat{x}) \times (-h,0))$ estimate for u, we can approximate ψ with $C^{0,1}$ functions so that its first order derivatives converge weakly in $L^q(B_1(\hat{x}) \times (-h,0))$. Thus the test function defined in (3.10) can be used in (2.15). Using (2.15) we find that, for all $\Psi \in X_c^q(\hat{x})$,

$$\begin{aligned} |\langle G(\varepsilon f)\ \xi, \Psi\rangle| &\le C \|u\|_{W^{1,p}(B_1(\hat{x})\times(-h,0))} \|\nabla\psi\|_{L^q(B_1(\hat{x})\times(-h,0))} \\ &\le C \|\xi\|_{X^q} \|\Psi\|_{X^q}. \end{aligned}$$

This implies that

$$\|G(\varepsilon f)\ \xi\|_{(X_c^q(\hat{x}))^*} \le C \|\xi\|_{X^p};$$

i.e., the estimate (3.9) is valid. □

Appendix A. Elliptic estimates. In this appendix we present the statements (together with brief proofs) of the elliptic estimates which are at the heart of the analyticity results, Theorems 2.1 and 2.2. Of course, the great simplification of our approach was the use of the "domain flattening" change of variables, (2.7), which maps the domain $S_{h,\eta}$ to the strip $S_{h,0}$. Consequently, it is sufficient to analyze (inhomogeneous) elliptic equations on a much simpler geometry. This, in turn, allows the simple establishment of the following results which, we point out, are true on much more general domains (e.g., see [5]).

We begin with the "comparison principle" on a domain, which implies the uniqueness of bounded solutions.

THEOREM A.1. *If w is bounded and satisfies (in the weak sense)*

$$\begin{aligned} &-\Delta w(x,y) \ge 0 && \textit{in } S_{h,0}, \\ &-\partial_y w(x,-h) \ge 0, \\ &w(x,0) \ge 0, \end{aligned}$$

then

$$w(x,y) \ge 0 \qquad \textit{in } S_{h,0}.$$

Proof. Since we can only use weak comparison in the bounded domain, we choose $M = |w|_{L^\infty}$ and let

$$\Phi = \frac{2(d-1)}{R} M \Big[\frac{x_1^2 + \cdots + x_{d-1}^2}{2(d-1)} - \frac{(y+h)^2}{2} + \frac{h^2}{2} \Big] + w.$$

It is clear that

$$\begin{aligned} &-\Delta\Phi \ge 0 && \text{in } S_{h,0}, \\ &-\partial_y \Phi(x,-h) \ge 0, \\ &\Phi \ge 0 && \text{on } \{y=0\} \cup \{x_1^2 + \cdots + x_{d-1}^2 = R^2\}. \end{aligned}$$

We can now apply the comparison principle on the bounded domain $\{x \mid x_1^2 + \cdots + x_{d-1}^2 < R^2\} \times (-h, 0)$ to conclude that $\Phi \geq 0$ there. If we fix (x, y) and let $M \to \infty$, we obtain $w > 0$ on $S_{h,0}$. ◻

We next state the Hölder estimate used in section 3.1.

THEOREM A.2. *For any $\alpha \in (0,1)$ there exists a constant C_e such that for any $R^{(1)}, R^{(2)} \in C^\alpha$, $R^{(3)} \in L^\infty$, and $\xi \in C^{1+\alpha}$ there exists a unique solution $w(x,y)$ of*

$$\begin{aligned} &\Delta w = div_x \left[R^{(1)}\right] + \partial_y R^{(2)} + R^{(3)} && \text{in } S_{h,0},\\ &\partial_y w(x,-h) = 0, &&\\ &w(x,0) = \xi(x), && \end{aligned}$$

which satisfies

$$|w|_{C^{1+\alpha}} \leq C_e \left\{ \left|R^{(1)}\right|_{C^\alpha} + \left|R^{(2)}\right|_{C^\alpha} + \left|R^{(3)}\right|_{L^\infty} + |\xi|_{C^{1+\alpha}} \right\}. \tag{A.1}$$

Proof. The uniqueness is a corollary of the comparison principle. The existence can be proved using a continuation argument once we obtain the estimate, (A.1), in this theorem. This estimate is a special case of the general $C^{1+\alpha}$ theory for elliptic systems in divergence form which is established using Campanato spaces $\mathcal{L}^{p,\mu}$ (see [5, Theorems 2.6 and 2.7, pp. 152–154]).

Since our system is of constant coefficients and in a special domain, we provide a short proof here. We write

$$w = \sum_{j=1}^{d-1} \partial_{x_j} w_j^{(1)} + \partial_y w^{(2)} + w^{(3)} + w^{(4)} + w^{(5)},$$

where

$$\Delta w_j^{(1)} = R_j^{(1)} \quad \text{in } S_{h,0}, \tag{A.2a}$$

$$\partial_y w_j^{(1)}(x,-h) = 0, \tag{A.2b}$$

$$w_j^{(1)}(x,0) = 0; \tag{A.2c}$$

$$\Delta w^{(2)} = R^{(2)} \quad \text{in } S_{h,0}, \tag{A.3a}$$

$$\partial_y w^{(2)}(x,-h) = 0, \tag{A.3b}$$

$$w^{(2)}(x,0) = 0; \tag{A.3c}$$

$$\Delta w^{(3)} = R_j^{(3)} \quad \text{in } S_{h,0}, \tag{A.4a}$$

$$\partial_y w^{(3)}(x,-h) = 0, \tag{A.4b}$$

$$w^{(3)}(x,0) = 0; \tag{A.4c}$$

$$\Delta w^{(4)} = 0 \quad \text{in } S_{h,0}, \tag{A.5a}$$

$$\partial_y w^{(4)}(x,-h) = 0, \tag{A.5b}$$

$$w^{(4)}(x,0) = \xi(x); \tag{A.5c}$$

and, finally,

$$\Delta w^{(5)} = 0 \qquad \text{in } S_{h,0}, \tag{A.6a}$$

$$\partial_y w^{(5)}(x,-h) = -\partial_{yy} w^{(2)}(x,-h) = -R^{(2)}(x,-h), \tag{A.6b}$$

$$w^{(5)}(x,0) = 0. \tag{A.6c}$$

We can apply standard Schauder theory [12] to $w^{(1)}$ and $w^{(2)}$ to obtain $C^{2+\alpha}$ estimates for $w^{(1)}$ and $w^{(2)}$. We can apply $W^{2,p}$ estimates to $w^{(3)}$ for $p > d/(1-\alpha)$ and then use an embedding theorem to obtain the $C^{1+\alpha}$ estimate for $w^{(3)}$. Since the Dirichlet boundary data for $w^{(4)}$ is $C^{1+\alpha}$, we obtain $C^{1+\alpha}$ estimates for $w^{(4)}$. Finally, if we let

$$z(x,y) = \int_{-h}^{y} w^{(5)}(x,s)\, \mathrm{d}s,$$

then

$$\begin{aligned} &\Delta z = w_y^{(5)}(x,-h) = -R^{(2)}(x,-h) \qquad \text{in } S_{h,0},\\ &\partial_y z(x,-h) = 0,\\ &z(x,0) = 0. \end{aligned}$$

Since $R^{(2)}$ is in C^α, we can apply the Schauder $C^{2+\alpha}$ estimate for z and obtain an $C^{1+\alpha}$ estimate for $w^{(5)} = \partial_y z$. ◻

Finally, we state the $W^{k,p}$ estimate used in section 3.2.

THEOREM A.3. *For any $p > d$ there exists a constant $\tilde{C}_e$ such that for any $R^{(1)}, R^{(2)}, R^{(3)} \in Y^{0,p}$, and $\xi \in X^p$ there exists a unique solution $w(x,y)$ of*

$$\Delta w = \mathrm{div}_x R^{(1)} + \partial_y\{(h+y)R^{(2)}\} + R^{(3)} \qquad \text{in } S_{h,0}, \tag{A.7a}$$

$$\partial_y w(x,-h) = 0, \tag{A.7b}$$

$$w(x,0) = \xi(x), \tag{A.7c}$$

which satisfies

$$\|w\|_{Y^{1,p}} \le \tilde{C}_e \left\{ \left\|R^{(1)}\right\|_{Y^{0,p}} + \left\|R^{(2)}\right\|_{Y^{0,p}} + \left\|R^{(3)}\right\|_{Y^{0,p}} + \|\xi\|_{X^p} \right\}.$$

Proof. This estimate is a special case of the general L^p theory for elliptic systems in divergence form which is established in [5] (see page 157, Theorem 2.2 for interior estimates; the boundary estimates can be done in a similar way). In this short proof for our special system, we will assume that the involved functions are smooth since we can always approximate them with smooth functions. The estimate is valid as long as the constants involved are independent of the smoothness. We use the ideas of the earlier proof (Theorem A.2) and divide the proof into two cases.

Case 1: $\xi(x) \equiv 0$. The proof is similar to the proof of Theorem A.2. For any $x^* \in \mathbf{R}^{d-1}$, it suffices to establish estimates on $B_1(x^*) \times (-h,0)$ in terms of norms of $R^{(1)}$, $R^{(2)}$, and $R^{(3)}$ on $B_2(x^*) \times (-h,0)$. We decompose w into $w^{(j)}$ $(j = 1,2,3,4,5)$ as before, and we can then apply the standard $W^{2,p}$ interior-boundary estimates to $w^{(1)}$, $w^{(2)}$, and $w^{(3)}$. Since we have a factor $(h+y)$ on the right-hand side of (A.7) in the $R^{(2)}$ term, $w^{(5)}$ vanishes. Since we have assumed, in this case, that $\xi \equiv 0$, $w^{(4)}$ also vanishes and the estimate is established.

Case 2: General case. We need only estimate $w^{(4)}$; by the maximum principle,

$$\left|w^{(4)}\right|_{L^\infty} \le |\xi|_{L^\infty}.$$

Since $p > d$, we have, by embedding, $|\xi|_{L^\infty} \le C \|\xi\|_{X^p}$. Thus we can use the standard elliptic regularity estimates to derive

$$\left\| w^{(4)} \right\|_{C^2(B_2(x^*)\times[-h,-h/2])} \le C \|\xi\|_{X^p}.$$

For any $x^* \in \mathbf{R}^{d-1}$, we use the definition of X^p to extend the function ξ to a function $\Phi(x,y) \in W^{1,p}_{loc}(B_2(x^*) \times [-h,0]) \cap C^2(\overline{B}_2(x^*) \times [-h,-h/2])$ so that

$$\|\Phi\|_{W^{1,p}(B_2(x^*)\times(-h,0))} \le C \|\xi\|_{X^p}, \tag{A.8}$$

where we understand that $\Phi(\cdot,0) = \xi(\cdot)$ in the trace sense. By using a cut-off function if necessary, we may assume, without loss of generality, that

$$\Phi(x,y) \equiv w^{(4)}(x,y) \quad \text{ for } x \in B_2(x^*), -h \le y \le \frac{-h}{2}.$$

It is clear that $w^{(4)}$ satisfies

$$\begin{aligned}
&\Delta(w^{(4)} - \Phi) = -\mathrm{div}_x \left[\mu_1(y)\nabla_x \Phi\right] - \partial_y\Big((y+h)\frac{\mu_1(y)}{y+h}\partial_y\Phi\Big) \quad \text{ in } S_{h,0},\\
&\partial_y(w^{(4)} - \Phi)(x,-h) = 0,\\
&w^{(4)}(x,0) - \Phi(x,0) = 0,
\end{aligned}$$

where

$$\mu_1(y) = 1 \quad \text{ for } \frac{-h}{2} \le y < 0, \qquad \mu_1(y) = 0 \quad \text{ for } -h \le y < \frac{-h}{2};$$

we point out that, in fact, $w^{(4)} - \Phi \equiv 0$ for $-h \le y \le -h/2$. Using (A.8), we have

$$\left\| \frac{\mu_1(y)}{(h+y)}\partial_y\Phi \right\|_{L^p(B_2(x^*)\times(-h,0))} \le \frac{2}{h} \|\partial_y \Phi\|_{L^p(B_2(x^*)\times(-h,0))} \le C \|\xi\|_{X^p}$$

and

$$\|\mu_1(y)\nabla_x\Phi\|_{L^p(B_2(x^*)\times(-h,0))} \le C \|\xi\|_{X^p}.$$

We can now apply Case 1 to obtain

$$\left\| w^{(4)} \right\|_{W^{1,p}(B_1(x^*)\times(-h,0))} \le C \|\xi\|_{X^p}.$$

Combining all the estimates for $w^{(j)}$ $(j = 1,2,3,4,5)$ and taking the supremum over all $x^* \in \mathbf{R}^{d-1}$, we conclude the theorem. □

Acknowledgment. The authors thank two anonymous referees for their helpful suggestions.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Pure Appl. Math. 65, Academic Press, New York, 1975.

[2] O. P. Bruno and F. Reitich, *Numerical solution of diffraction problems: A method of variation of boundaries*, J. Opt. Soc. Amer. A, 10 (1993), pp. 1168–1175.

[3] B. Buffoni, *Existence and conditional energetic stability of capillary-gravity solitary water waves by minimisation*, Arch. Ration. Mech. Anal., 173 (2004), pp. 25–68.

[4] A. P. Calderón, *Cauchy integrals on Lipschitz curves and related operators*, Proc. Natl. Acad. Sci. USA, 75 (1977), pp. 1324–1327.

[5] Y.-Z. Chen and L.-C. Wu, *Second Order Elliptic Equations and Elliptic Systems*, Transl. Math. Monogr. 174, AMS, Providence, RI, 1998.

[6] R. Coifman and Y. Meyer, *Nonlinear harmonic analysis and analytic dependence*, in Pseudodifferential Operators and Applications (Notre Dame, IN, 1984), AMS, Providence, RI, 1985, pp. 71–78.

[7] W. Craig, P. Guyenne, D. P. Nicholls, and C. Sulem, *Hamiltonian long wave expansions for water waves over a rough bottom*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 461 (2005), pp. 839–873.

[8] W. Craig and D. P. Nicholls, *Traveling two and three dimensional capillary gravity water waves*, SIAM J. Math. Anal., 32 (2000), pp. 323–359.

[9] W. Craig and D. P. Nicholls, *Traveling gravity water waves in two and three dimensions*, Eur. J. Mech. B Fluids, 21 (2002), pp. 615–641.

[10] W. Craig, U. Schanz, and C. Sulem, *The modulation regime of three-dimensional water waves and the Davey-Stewartson system*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 615–667.

[11] W. Craig and C. Sulem, *Numerical simulation of gravity waves*, J. Comput. Phys., 108 (1993), pp. 73–83.

[12] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[13] C. Godrèche, ed., *Solids Far from Equilibrium*, Cambridge University Press, Cambridge, UK, 1992.

[14] P. Guyenne and D. P. Nicholls, *Numerical simulation of solitary waves on plane slopes*, Math. Comput. Simulation, to appear.

[15] H. Lamb, *Hydrodynamics*, 6th ed., Cambridge University Press, Cambridge, UK, 1993.

[16] D. P. Nicholls, *Traveling water waves: Spectral continuation methods with parallel implementation*, J. Comput. Phys., 143 (1998), pp. 224–240.

[17] D. P. Nicholls, *On hexagonal gravity water waves*, Math. Comput. Simulation, 55 (2001), pp. 567–575.

[18] D. P. Nicholls and F. Reitich, *A new approach to analyticity of Dirichlet-Neumann operators*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 1411–1433.

[19] D. P. Nicholls and F. Reitich, *Stability of high-order perturbative methods for the computation of Dirichlet-Neumann operators*, J. Comput. Phys., 170 (2001), pp. 276–298.

[20] D. P. Nicholls and F. Reitich, *Analytic continuation of Dirichlet-Neumann operators*, Numer. Math., 94 (2003), pp. 107–146.

[21] D. P. Nicholls and F. Reitich, *Shape deformations in rough surface scattering: Cancellations, conditioning, and convergence*, J. Opt. Soc. Amer. A, 21 (2004), pp. 590–605.

[22] D. P. Nicholls and F. Reitich, *Shape deformations in rough surface scattering: Improved algorithms*, J. Opt. Soc. Amer. A, 21 (2004), pp. 606–621.

[23] V. Zakharov, *Stability of periodic waves of finite amplitude on the surface of a deep fluid*, J. Appl. Mech. Tech. Phys., 9 (1968), pp. 190–194.

SIAM J. MATH. ANAL.
Vol. 37, No. 1, pp. 321–342

ON BOUND STATES CONCENTRATING ON SPHERES FOR THE MAXWELL–SCHRÖDINGER EQUATION*

TERESA D'APRILE† AND JUNCHENG WEI‡

Abstract. We study the semiclassical limit for the following system of Maxwell–Schrödinger equations:

$$-\frac{\hbar^2}{2m}\Delta v + v + \omega\phi v - \gamma v^p = 0, \quad -\Delta\phi = 4\pi\omega v^2,$$

where $\hbar$, m, ω, $\gamma > 0$, v, $\phi : \mathbb{R}^3 \to \mathbb{R}$, $1 < p < \frac{11}{7}$. This system describes standing waves for the nonlinear Schrödinger equation interacting with the electrostatic field: the unknowns v and ϕ represent the *wave function* associated to the particle and the electric potential, respectively. By using localized energy method, we construct a family of positive radially symmetric bound states $(v_\hbar, \phi_\hbar)$ such that $v_\hbar$ concentrates around a sphere $\{|x| = s_0\}$ when $\hbar \to 0$.

Key words. bound states, Maxwell–Schrödinger equation, finite dimensional reduction

AMS subject classifications. Primary, 35B40, 35B45; Secondary, 35J55, 92C15, 92C40

DOI. 10.1137/S0036141004442793

1. Introduction.

In this paper we consider the following system of Maxwell–Schrödinger equations in the following electrostatic case:

$$\begin{cases} -\hbar^2\Delta v + v + \omega\phi v - \gamma v^p = 0 \text{ in } \mathbb{R}^3, \\ -\Delta\phi = 4\pi\omega v^2 \text{ in } \mathbb{R}^3, \\ v,\ \phi > 0, \quad v,\ \phi \to 0 \text{ as } |x| \to +\infty, \end{cases} \tag{1.1}$$

where $\hbar$, ω, $\gamma > 0$ and $p > 1$. This system was first introduced in [6]; it provides a physical model which describes the interaction of a charged particle with the electrostatic field. In (1.1) we have assumed, for the sake of simplicity, $2m = 1$, where m is the mass of the particle. ω denotes the electric charge of the particle, while $\hbar$ is a constant which is known under the name of *Planck's constant*. The unknowns of the system are the wave function v associated to the particle and the electric potential ϕ. The presence of the nonlinear term in (1.1) simulates the interaction effect among many particles.

The eigenvalue problem for the system (1.1) with $\gamma = 0$ has been studied in [6] (in the case in which the charged particle lies in a bounded space region Ω) and in [8] (under the action of an external nonzero potential). Existence results for (1.1) have been established in [11].

In all the above-mentioned results the size of $\hbar$ is not relevant, hence one may assume, without loss of generality, $\hbar = 1$. This paper deals with the *semiclassical limit* of the system (1.1); i.e., it is concerned with the problem of finding nontrivial solutions and studying their asymptotic behavior when $\hbar \to 0^+$. Hence such solutions

*Received by the editors April 7, 2004; accepted for publication (in revised form) January 21, 2005; published electronically October 7, 2005.
http://www.siam.org/journals/sima/37-1/44279.html
†Dipartimento di Matematica, Universitá degli Studi di Bari, via E. Orabona 4, 70125 Bari, Italy (daprile@dm.uniba.it).
‡Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong (wei@math.cuhk.edu.hk). The research of this author was supported by an Earmarked Grant from RGC of Hong Kong.

are usually referred to as *semiclassical* ones. Letting $\hbar$ go to zero has great physical interest since it formally describes the transition from quantum mechanics to classical mechanics; as one expects, the solutions exhibit some kind of notable behavior in the semiclassical limit, a concentration behavior: their form consists of very sharp peaks which become highly concentrated when $\hbar$ is small.

The study of semiclassical phenomena for nonlinear Schrödinger equations has attracted considerable attention in recent years; in particular, a large number of papers have been devoted to studying single and multiple spike solutions for the following stationary Schrödinger equation in the presence of an external potential V:

$$-\hbar^2\Delta v + V(x)v - \gamma|v|^{p-1}v = 0, \quad x \in \Omega, \quad V : \Omega \to \mathbb{R}, \tag{1.2}$$

where Ω is an open subset of $\mathbb{R}^N$ $(N \geq 1)$. In the case $p \in (1, \frac{N+2}{N-2})$ for $N \geq 3$ and $p > 1$ for $N = 1, 2$, (1.2) admits solutions, called *spike layers*, concentrating at one or multiple points of $\overline{\Omega}$, in the sense that their shape consists of sharp peaks near those points while it vanishes everywhere else as $\hbar \to 0^+$.

Concerning (1.2) in bounded domains Ω with $V \equiv 1$ and Neumann boundary conditions, in [16], [20], [22], [23], [26], [32], [31], [41] the authors look for solutions exhibiting a concentration at one or more points of the boundary which are proved to be critical for the mean curvature of $\partial\Omega$. As regards interior spike solutions, the distance function from the boundary $\partial\Omega$ plays a role similar to that of the mean curvature: in [19], [21], [33], [35], [42] single or multiple spikes are constructed for the Neumann or the Dirichlet problem associated to (1.2) with Ω bounded and $V \equiv 1$, and concentration occurs, roughly speaking, at critical points of the distance function d from the boundary $\partial\Omega$.

When $\Omega \equiv \mathbb{R}^N$ and $V \not\equiv$ const, spike layer solutions are constructed around the critical points of the potential V. The first result in this line seems due to Floer and Weinstein [17]. These authors considered the one-dimensional case and constructed for small $\hbar > 0$ such a concentrating family via a Lyapunov–Schmidt reduction around any nondegenerate critical point of the potential V, under the condition that V is bounded and $p = 3$. This line of research has been extensively pursued in a set of recent papers (we recall, among many others, [1], [12], [14], [13], [15], [18], [24], [25], [36], [37], [38], [40]) and semiclassical states are produced using minimax methods or, under suitable nondegeneracy conditions, by finite dimensional reductions.

In contrast to pointwise concentration, only very recently has concentration at higher dimensional sets been proved. The first progress was the recent paper [29], in which, assuming $V \equiv$ const, the authors prove the existence of positive solutions for (1.2) in a smooth bounded subset $\Omega \subset \mathbb{R}^2$ with the Neumann boundary conditions; such solutions concentrate at the whole boundary $\partial\Omega$ or at some of its components. This result has been extended to higher dimensions in [28]. Under symmetry assumptions, higher dimensional spike layers are proved to exist. In [2] and [4], assuming that V is radial, the existence of radially symmetric solutions for (1.2) in $\mathbb{R}^N$ concentrating on a sphere centered in zero is established, and it is shown that the location of the concentration arises as a balance between the effect of the potential energy due to V and the volume energy of the solutions. In [5] and [9] concentration on circles in $\mathbb{R}^2$ and $\mathbb{R}^3$, respectively, is produced for a special class of solutions of (1.2). Concerning (1.2) in a ball with Dirichlet or Neumann boundary conditions and V radial, in [3] the effect of the boundary of the domain is analyzed; more precisely, it is proved that there exists a family of radial solutions concentrating on the boundary. Finally, we recall the recent papers [27] and [30], in which other phenomena of concentration on manifolds are studied.

Motivated by this rich literature, in this paper we want to study which kind of phenomena occur for the system of Maxwell–Schrödinger equations (1.1) in the semiclassical limit.

More precisely, for small values of the parameter $\hbar$, we prove the existence of a positive radially symmetric wave $v_\hbar$ and a potential $\phi_\hbar$ satisfying (1.1). Furthermore, in the limit when $\hbar \to 0^+$, $v_\hbar$ exhibits a concentration behavior around a sphere $\{|x| = s_0\}$: the analysis reveals that most of the support of $v_\hbar$ is concentrated in a neighborhood of such a sphere whose size depends on $\hbar$ and shrinks to zero as $\hbar \to 0^+$. The radius s_0 is determined by the exponent p and the charge ω. In order to state the exact result, denote by $w(y)$ the unique solution for the following ODE:

$$\begin{cases} w'' - w + w^p = 0 \text{ in } \mathbb{R}, \\ w > 0, \quad w(0) = \max_{y\in\mathbb{R}} w(y), \\ w(y) \to 0 \text{ as } |y| \to +\infty. \end{cases}$$

It is well known that $w(y)$ can be written explicitly as

$$w(y) = \left(\frac{p+1}{2}\right)^{\frac{1}{p-1}} \left(\cosh\left(\frac{p-1}{2}y\right)\right)^{-\frac{2}{p-1}}. \tag{1.3}$$

Then, roughly speaking, the limit profile of our solutions $v_\hbar$ resembles, after a suitable rescaling in the coordinates, the function w.

More precisely, the main result of this paper is the following.

THEOREM 1.1. *Assume that $1 < p < \frac{11}{7}$. Then there exists $\hbar_0 > 0$ such that for every $\hbar \in (0, \hbar_0)$ the system* (1.1) *has a solution $(v_\hbar, \phi_\hbar)$ with $\gamma = \gamma_\hbar = \hbar^{(p-1)/2}$ and*

(1) *$v_\hbar$, $\phi_\hbar$ are radially symmetric, $v_\hbar$, $|\nabla v_\hbar|$, $|\nabla \phi_\hbar| \in L^2(\mathbb{R}^3)$,*

(2) *$v_\hbar(r) = \hbar^{-(1/2)}\alpha_\hbar^{1/(p-1)} w\big(\sqrt{\alpha_\hbar}\frac{r-s_\hbar}{\hbar}\big) + o(1)\exp\big(-\mu\frac{|r-s_\hbar|}{\hbar}\big)$ uniformly in $\mathbb{R}^3$ for a suitable $\mu \in (0,1)$,*

(3) *$\phi_\hbar(r) = 4\pi\omega(1+o(1))\alpha_\hbar^{(5-p)/2(p-1)}(\int_{\mathbb{R}} w^2 dy)G(r; s_\hbar)$, where $G(r;\rho)$ is defined as follows:*

$$G(r;\rho) = \frac{\rho}{r}\min(r,\rho). \tag{1.4}$$

Furthermore, as $\hbar \to 0^+$,

$$\alpha_\hbar \to \alpha_0 := \frac{p+3}{11-7p} \tag{1.5}$$

and

$$s_\hbar \to s_0 := (\alpha_0 - 1)\alpha_0^{\frac{p-5}{2(p-1)}} \left(4\pi\omega^2 \int_{\mathbb{R}} w^2 dy\right)^{-1}. \tag{1.6}$$

REMARK 1.2. *The case $1 < p < \frac{3}{2}$ has been proved in* [10] *by variational method. Here we use localized energy method to cover the whole range for $p \in (1, \frac{11}{7})$. Notice that for $p \geq \frac{11}{7}$ it results in $\alpha_0 - 1 \leq 0$, hence formula* (1.6) *does not make sense; this suggests that the range $(1, \frac{11}{7})$ is also necessary to get concentration around a sphere. Furthermore, unlike* [10]*, the different method used in this paper allows us to provide the more precise description of the profile of the solutions $(v_\hbar, \phi_\hbar)$ given by* (2)–(3) *of Theorem* 1.1.

Remark 1.3. *Notice that by* (2) *of Theorem* 1.1 *it immediately follows that*

$$\int_{\mathbb{R}^3} |v_\hbar|^2 dx = 4\pi\alpha_\hbar^{\frac{5-p}{2(p-1)}} \int_{-\sqrt{\alpha_\hbar}\frac{s_\hbar}{\hbar}}^{+\infty} \left(\frac{\hbar y}{\sqrt{\alpha_\hbar}} + s_\hbar\right)^2 |w|^2 dy + o(1)$$
$$= 4\pi\alpha_0^{\frac{5-p}{2(p-1)}} s_0^2 \int_{\mathbb{R}} |w|^2 dy + o(1);$$

i.e., the L^2-norm of the waves $v_\hbar$ becomes bounded from below and from above. Furthermore, for every $\delta > 0$ and for small $\hbar$,

$$v_\hbar(r) \le C \exp\left(-\mu\frac{|r - s_\hbar|}{\hbar}\right) \ \text{for}\ |r - s_\hbar| \ge \delta,$$

where C is independent on $\hbar$ and δ. Hence the solutions $v_\hbar$ can be viewed as small perturbations of the function w rescaled in such a way to be very concentrated *near a suitable neighborhood of s_0; then their profile has the form of a solitary elevation which becomes a very sharp peak when $\hbar$ is small.*

The proof of Theorem 1.1 relies on a local approach and is related to the arguments employed in [34], in which a phenomenon of concentration around a sphere has been obtained for the stationary Gierer–Meinhardt system in $\mathbb{R}^N$. This approach is based on a finite dimensional reduction by using the classical Lyapunov–Schmidt reduction method. We first construct some approximate solution (obtained as a small perturbation of rescalation of w) and we solve (1.1) in its normal direction, and then we study the remaining finite dimensional equation. After this reduction process, by using the implicit function theorem we solve (1.1) in a suitable neighborhood of the approximate solution.

The paper consists of four more sections. In section 2 we introduce some notations and prove some preliminary estimates that play a key role in the rest of the arguments. In sections 3 and 4 we construct the approximate solutions and we carry out the Lyapunov–Schmidt procedure that allows us to reduce the problem to the study of a finite dimensional functional. Finally, the proof of Theorem 1.1 is given in section 5.

Notations.

- If $u : \mathbb{R}^N \to \mathbb{R}$ is a radially symmetric function, we will continue to denote by u the real function $r > 0 \mapsto u(r)$ with $|x| = r$.
- If u is a real valued function, then u_+ is its positive part and u_- its negative part.
- We will often use the symbol C for denoting a positive constant independent on ε. The value of C is allowed to vary from line to line (and also in the same formula).
- Given $A_\varepsilon \subset \mathbb{R}^N$, $f_\varepsilon, g_\varepsilon : A_\varepsilon \to \mathbb{R}$ two families of functions on A_ε and $k \in \mathbb{R}$, we write $f_\varepsilon = o(\varepsilon^k)g_\varepsilon$ *on* A_ε (resp. $f_\varepsilon = O(\epsilon^k)g_\varepsilon$ *on* A_ε) to mean that $\frac{f_\varepsilon}{g_\varepsilon} \to 0$ (resp. $|f_\varepsilon| \le C\varepsilon^k|g|$) uniformly on A_ε as $\varepsilon \to 0$.

2. Preliminaries. In order to obtain solutions of (1.1) we choose a suitable functional setting. Let $H_r^1(\mathbb{R}^3)$ and $L_r^2(\mathbb{R}^3)$ denote the subspace of the Sobolev space $H^1(\mathbb{R}^3)$ and $L^2(\mathbb{R}^3)$, respectively, formed by the radially symmetric functions endowed with the norms

$$\|u\|_{H^1(\mathbb{R}^3)}^2 = \int_{\mathbb{R}^3} \left(|\nabla u|^2 + |u|^2\right) dx, \quad \|u\|_{L^2(\mathbb{R}^3)}^2 = \int_{\mathbb{R}^3} |u|^2\, dx.$$

It is well known that $H_r^1(\mathbb{R}^3)$ is compactly imbedded in $L^p(\mathbb{R}^3)$ for $2 < p < 6$ (see [7, Theorem A.I', p. 341] or [39]). Furthermore, if $u \in H_r^1(\mathbb{R}^3)$, then

$$|u(x)| \leq (4\pi)^{-\frac{1}{2}} |x|^{-1} \|u\|_{H^1} \text{ a.e. in } \mathbb{R}^3$$

(see [39]). Define $D^1(\mathbb{R}^3)$ as the completion of $C_0^\infty(\mathbb{R}^3)$ with respect to the norm

$$\|u\|^2_{D^1(\mathbb{R}^3)} \equiv \int_{\mathbb{R}^3} |\nabla u|^2 \, dx,$$

and let $D_r^1(\mathbb{R}^3)$ denote the closed subspace consisting of the radially symmetric functions.

The following proposition holds (see [10] for the proof).

PROPOSITION 2.1. *For every $u \in H_r^1(\mathbb{R}^3)$ denote by $(-\Delta)^{-1}[u^2]$ the unique solution in $D_r^1(\mathbb{R}^3)$ of*

$$-\Delta v = u^2.$$

Then the following representation formula holds:

$$(-\Delta)^{-1}[u^2](x) = \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{1}{|x-y|} u^2(y) dy = \int_0^{+\infty} G(|x|, \rho) u^2(\rho) d\rho,$$

where $G(s, \rho)$ has been defined in (1.4). *Furthermore, the functional $F : u \in H_r^1(\mathbb{R}^3) \mapsto \int_{\mathbb{R}^3} u^2 (-\Delta)^{-1}[u^2] dx$ is compact and C^1, and*

$$F'(u)[w] = 4 \int_{\mathbb{R}^3} uw(-\Delta)^{-1}[u^2] dx \quad \forall u, w \in H_r^1(\mathbb{R}^3).$$

By suitably scaling, (1.1) can be reduced to the following system of equations:

$$\begin{cases} -\Delta u + u + \Gamma u\psi - u^p = 0 \text{ in } \mathbb{R}^3, \\ -\Delta \psi = \varepsilon u^2 \text{ in } \mathbb{R}^3, \\ u, \, \psi > 0, \quad u, \, \psi \to 0 \text{ as } |x| \to +\infty, \end{cases} \tag{2.1}$$

where

$$\varepsilon = \hbar = \gamma^{\frac{2}{p-1}}, \quad \Gamma = 4\pi\omega^2, \quad u(x) = \sqrt{\varepsilon} v(\varepsilon x), \quad \psi(x) = \frac{1}{4\pi\omega} \phi(\varepsilon x). \tag{2.2}$$

Associated with (2.1) is the following energy functional $E_\varepsilon \in C^1(H_r^1(\mathbb{R}^3), \mathbb{R})$:

$$E_\varepsilon[u] := \frac{1}{2} \int_{\mathbb{R}^3} \left(|\nabla u|^2 + u^2 \right) dx - \frac{1}{p+1} \int_{\mathbb{R}^3} u_+^{p+1} dx + \frac{\Gamma}{4} \varepsilon \int_{\mathbb{R}^3} u^2 (-\Delta)^{-1}[u^2] dx.$$

By using Proposition 2.1 the energy functional can be rewritten as

$$\begin{aligned} E_\varepsilon[u] = 2\pi \int_0^{+\infty} r^2 \left(|u'|^2 + |u|^2 \right) dr - \frac{4\pi}{p+1} \int_0^{+\infty} r^2 |u_+|^{p+1} dr \\ + \Gamma \pi \varepsilon \int_0^{+\infty} \int_0^{+\infty} r^2 G(r, \rho) u^2(r) u^2(\rho) dr d\rho. \end{aligned}$$

Before going on we define in the following lemma two suitable functions which will be useful in order to locate the asymptotic peak of the solutions.

Lemma 2.2. *Assume $p, N \in \mathbb{R}$ such that $N \geq 3$ and $1 < p < \frac{7N-10}{3N-2}$, and consider the two functions*

$$\rho(t) = (1+t)^{\frac{p-5}{2(p-1)}} t^2[(3p-7)t + 4(p-1)], \quad t > \hat{t} := \frac{2(p-1)}{7-3p},$$

and $t = t(s)$ the inverse function of

$$s(t) = \left(\Gamma \int_{\mathbb{R}} w^2 dy\right)^{-1} (1+t)^{\frac{p-5}{2(p-1)}} t, \quad t > \hat{t}.$$

Then the function $s^{N-3}\rho(t(s))$ has a unique (and nondegenerate) minimum point s_0 for $s \in (0, s(\hat{t}))$. In particular, if $N = 3$, then s_0 is given by (1.6).

Proof. Since $\frac{7N-10}{3N-2} < \frac{7}{3}$, it is easy to see that s is a strictly decreasing function for $t > \hat{t}$. Thus the inverse function $t = t(s)$ exists for $s \in (0, s(\hat{t}))$. Now compute

$$\begin{aligned}\left(s^{N-3}\rho\right)'(t) = &\left(\Gamma \int_{\mathbb{R}} w^2 dy\right)^{3-N} t^{N-2}(1+t)^{\frac{(p-5)N+4(3-p)}{2(p-1)}} \\ &\times\left(\frac{(7-3p)N + 2(p-5)}{2(p-1)} t - 2(N-1)\right)\big((7-3p)t - 2(p-1)\big).\end{aligned}$$

For $t > \hat{t}$, $\rho(t)$ has the following unique critical point:

$$t_0 = \frac{4(N-1)(p-1)}{(7-3p)N + 2(p-5)}. \tag{2.3}$$

Moreover, $\rho''(t_0) > 0$, which means that t_0 is also nondegenerate. Set $s_0 = s(t_0)$. Then the function $\rho(t(s))$ has a unique (and nondegenerate) minimum point s_0 for $s \in (0, s(\hat{t}))$. $\square$

Now we are in position to introduce our approximate solutions. As we have already stated in the introduction, we construct the approximations by suitably rescaling the function w and then using suitable truncations. Fix $s_1, s_2 \in (0, s(\hat{t}))$ such that $s_1 < s_2$ and, according to Lemma 2.2,

$$s_0 \in (s_1, s_2), \quad \frac{2t(s)}{1+t(s)} \neq \frac{4(p-1)}{5-p} \quad \forall s \in [s_1, s_2]. \tag{2.4}$$

Let $\eta(r)$ be a cut-off function satisfying

$$\eta(r) = 1 \text{ for } r \in (-\delta, \delta), \quad \eta(r) = 0 \text{ for } r \notin (-2\delta, +2\delta), \tag{2.5}$$

where $\delta > 0$ is a fixed number such that $s_1 - 2\delta > 0$. For every $s \in [s_1, s_2]$ define

$$w_s(y) := (1+t(s))^{\frac{1}{p-1}} w(\sqrt{1+t(s)}y)$$

and

$$U_{\varepsilon,s}(y) := w_s(y)\eta(\varepsilon y). \tag{2.6}$$

We immediately see that w_s satisfies

$$w_s'' - (1+t(s))w_s + w_s^p = 0, \quad w_s > 0, \quad w_s(y) = w_s(-y). \tag{2.7}$$

The definitions of $U_{\varepsilon,s}$ in (2.6) and w in (1.3) imply that for every $i \in \mathbb{N}$, denoting by $U^i_{\varepsilon,s}$ and $w^i_{\varepsilon,s}$ the derivatives of order i of the functions $U_{\varepsilon,s}$ and w_s, respectively, we have

$$U^i_{\varepsilon,s}(y) = O(1)e^{-|y|}, \quad U^i_{\varepsilon,s}(y) = w^i_s(y) + o(\varepsilon)e^{-|y|} \tag{2.8}$$

uniformly for $s \in [s_1, s_2]$. We need to recall the following well-known facts on the functions w and w_s which will be frequently used later on.

LEMMA 2.3. (1) *For every $s \in [s_1, s_2]$ the following identities hold:*

$$\int_{\mathbb{R}} (w_s')^2 dy = \frac{p-1}{2(p+1)} \int_{\mathbb{R}} w_s^{p+1} dy, \quad \int_{\mathbb{R}} w_s^2 dy = \frac{p+3}{2(1+t(s))(p+1)} \int_{\mathbb{R}} w_s^{p+1} dy. \tag{2.9}$$

(2) *Let $L_0 : H^2(\mathbb{R}) \to L^2(\mathbb{R})$ be the linear operator defined by*

$$L_0[\phi] = \phi'' - \phi + pw^{p-1}\phi.$$

Then

$$L_0\left[\frac{1}{p-1}w + \frac{1}{2}yw'\right] = w \tag{2.10}$$

and $\phi \in H^2(\mathbb{R})$ satisfies

$$L_0[\phi] = 0$$

if and only if $\phi = cw'$ for some constant c. As a consequence, if we consider the kernel $\mathcal{K}_0$ and the cokernel $\mathcal{C}_0$ of the operator L_0,

$$\mathcal{K}_0 = \{cw' \mid c \in \mathbb{R}\} \subset H^2(\mathbb{R}), \ \ \mathcal{C}_0 = \{cw' \mid c \in \mathbb{R}\} \subset L^2(\mathbb{R}),$$

then L_0 is an invertible operator from $\mathcal{K}_0^\perp$ to $\mathcal{C}_0^\perp$.

Proof. Since each w_s satisfies (2.7), we have $\left((w_s')^2\right)' = (1+t(s))(w_s^2)' - \frac{2}{p+1}(w_s^{p+1})'$, by which $(w_s')^2 = (1+t(s))w_s^2 - \frac{2}{p+1}w_s^{p+1}$. Combining this with the identity

$$\int_{\mathbb{R}} (w_s')^2 dy + (1+t(s)) \int_{\mathbb{R}} w_s^2 dy = \int_{\mathbb{R}} w_s^{p+1} dy$$

we deduce (2.9). To prove (2.10), it is sufficient to observe that $L_0[w] = (p-1)w^p$, $L_0[yw'] = 2(w - w^p)$.

The last part of the lemma follows from the uniqueness of w. ∎

Fix $\lambda > 0$. We conclude this section by analyzing the following nonlocal operator:

$$L : H^2(\mathbb{R}) \to L^2(\mathbb{R}), \ \ L[\phi] := L_0[\phi] - \lambda(p-1)\frac{\int_{\mathbb{R}} w\phi dy}{\int_{\mathbb{R}} w^2 dy} w.$$

The following lemma characterizes the kernel of L.

LEMMA 2.4. *If $\lambda \neq \frac{4}{5-p}$, then $\phi \in H^2(\mathbb{R})$ satisfies*

$$L[\phi] = 0$$

if and only if $\phi = cw'$ for some constant c.

Proof. Assume $L[\phi]=0$. From (2.10) we deduce that

$$L_0\left[\phi-\lambda\frac{\int_{\mathbb{R}} w\phi dy}{\int_{\mathbb{R}} w^2 dy}\left(w+\frac{p-1}{2}yw^{'}\right)\right]=L[\phi]=0.$$

By (2) of Lemma 2.3, we have

$$\phi-\lambda\frac{\int_{\mathbb{R}} w\phi dy}{\int_{\mathbb{R}} w^2 dy}\left(w+\frac{p-1}{2}yw^{'}\right)=cw^{'} \tag{2.11}$$

for some c. Multiplying (2.11) by w and integrating over $\mathbb{R}$, we obtain

$$(1-\lambda)\int_{\mathbb{R}} w\phi dy-\lambda\frac{p-1}{2}\frac{\int_{\mathbb{R}} w\phi dy}{\int_{\mathbb{R}} w^2 dy}\int_{\mathbb{R}} yww' dy=0.$$

After an integration by parts we deduce $\int_{\mathbb{R}} yww'dy=-\frac{1}{2}\int_{\mathbb{R}} w^2dy$, by which

$$\left(1-\lambda\left(1-\frac{p-1}{4}\right)\right)\int_{\mathbb{R}} w\phi dy=0,$$

and hence $\int_{\mathbb{R}} w\phi dy=0$, which implies $L_0[\phi]=L[\phi]=0$; then Lemma 2.3 leads to $\phi=cw^{'}$. □

3. The linearized problem. For every $s\in[s_1,s_2]$, where s_1, s_2 have be chosen in (2.4), we set

$$I_{\varepsilon,s}:=\left(-\frac{s}{\varepsilon},+\infty\right).$$

Denote by $L^p_\varepsilon(I_{\varepsilon,s})$ the space of the functions $\phi: I_{\varepsilon,s}\to\mathbb{R}$ such that, setting $u_\phi(x)=\phi\big(|x|-\frac{s}{\varepsilon}\big)$, it results in $u_\phi\in L^p(\mathbb{R}^3)$. $L^2_\varepsilon(I_{\varepsilon,s})$ can be equipped with the following scalar product:

$$\langle\phi,\psi\rangle_\varepsilon=\int_{I_{\varepsilon,s}}\phi\psi(s+\varepsilon y)^2dy\quad\forall\phi,\psi\in L^2_\varepsilon(I_{\varepsilon,s})$$

(which is equivalent to the norm $\|u_\phi\|_{L^2(\mathbb{R}^3)}$).

We define the following operator:

$$T_{\varepsilon,s}[\phi](y)=\int_{I_{\varepsilon,s}}G(s+\varepsilon y,s+\varepsilon z)\phi(z)dz,\ \ y\in I_{\varepsilon,s},\ \ \phi\in L^1_\varepsilon(I_{\varepsilon,s}).$$

Notice that

$$s-\varepsilon(2|z|+|y|)\le G(s+\varepsilon y,s+\varepsilon z)\le s+\varepsilon z\quad\forall y,z\in I_{\varepsilon,s}; \tag{3.1}$$

hence, by using (2.8) and Lemma 2.2, we get

$$\begin{aligned}T_{\varepsilon,s}[U^2_{\varepsilon,s}](y)&=s\int_{\mathbb{R}}U^2_{\varepsilon,s}dz+O(\varepsilon)(|y|+1)=s\int_{\mathbb{R}}w_s^2dz+O(\varepsilon)(|y|+1)\\&=s(1+t(s))^{\frac{5-p}{2(p-1)}}\int_{\mathbb{R}}w^2dz+O(\varepsilon)(|y|+1)=\frac{1}{\Gamma}t(s)+O(\varepsilon)(|y|+1)\end{aligned} \tag{3.2}$$

uniformly for $s \in [s_1, s_2]$. In the same way, we define $H^1_\varepsilon(I_{\varepsilon,s})$, the space of the functions $\phi : I_{\varepsilon,s} \to \mathbb{R}$, such that $u_\phi \in H^1_r(\mathbb{R}^3)$. We can equip $H^1_\varepsilon(I_{\varepsilon,s})$ with the following scalar product:

$$(\phi,\psi)_\varepsilon = \int_{I_{\varepsilon,s}} [\phi'\psi' + (1+\Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])\phi\psi](s+\varepsilon y)^2 dy \quad \forall \phi,\psi \in H^1_\varepsilon(I_{\varepsilon,s})$$

(which is equivalent to the norm $\|u_\phi\|_{H^1(\mathbb{R}^3)}$ since $0 \le T_{\varepsilon,s}[U^2_{\varepsilon,s}](y) \le \int_{I_{\varepsilon,s}}(s+\varepsilon z)U^2_{\varepsilon,s}dz$).

We introduce the following functions:

$$Z_{\varepsilon,s}(y) = U'''_{\varepsilon,s}(y) + \frac{2\varepsilon}{s+\varepsilon y}U''_{\varepsilon,s}(y) - (1+\Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])U'_{\varepsilon,s}(y), \quad y \in I_{\varepsilon,s}. \tag{3.3}$$

By integration by parts we immediately prove that

$$(\phi, U'_{\varepsilon,s})_\varepsilon = -\langle \phi, Z_{\varepsilon,s}\rangle_\varepsilon \quad \forall \phi \in H^1_\varepsilon(I_{\varepsilon,s}); \tag{3.4}$$

then orthogonality of the functions $U'_{\varepsilon,s}$ in $H^1_\varepsilon(I_{\varepsilon,s})$ is equivalent to orthogonality of $Z_{\varepsilon,s}$ in $L^2_\varepsilon(I_{\varepsilon,s})$.

First we study a linear problem: given $h \in C(\bar I_{\varepsilon,s}) \cap L^2_\varepsilon(I_{\varepsilon,s})$, find a function $\phi \in H^1_\varepsilon(I_{\varepsilon,s})$ and $c \in \mathbb{R}$ satisfying

$$\begin{cases} L_{\varepsilon,s}[\phi] = h + cZ_{\varepsilon,s}, \\ \phi'\left(-\frac{s}{\varepsilon}\right) = 0, \quad \langle \phi, Z_{\varepsilon,s}\rangle_\varepsilon = 0, \end{cases} \tag{3.5}$$

where

$$L_{\varepsilon,s}[\phi] := \phi'' + \frac{2\varepsilon}{s+\varepsilon y}\phi' - (1+\Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])\phi + pU^{p-1}_{\varepsilon,s}\phi - 2\Gamma T_{\varepsilon,s}[U_{\varepsilon,s}\phi]U_{\varepsilon,s}.$$

In order to solve the system (3.5) we need the following result based on ODE estimates.

Lemma 3.1. *Let $\phi \in C^2(\bar I_{\varepsilon,s})$ satisfy*

$$\left|\phi''(y) + \frac{2\varepsilon}{s+\varepsilon y}\phi'(y) - (1+\Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])\phi(y)\right| \le c_0 e^{-\mu|y|}, \quad \phi'\left(-\frac{s}{\varepsilon}\right) = 0,$$

for some $c_0 > 0$, and

$$\phi(y) \to 0 \quad \text{as } y \to +\infty.$$

Then, provided that $\mu > 0$ is sufficiently small,

$$|\phi(y)| \le 2e^2(|\phi(0)| + c_0)e^{-\mu|y|} \quad \forall y \in I_{\varepsilon,s}.$$

Proof. We use a comparison principle. Take $\chi(t)$, a smooth cut-off function such that

$$\chi(t) = 1 \text{ for } |t| \le 1, \quad \chi(t) = 0 \text{ for } |t| \ge 2, \quad 0 \le \chi \le 1.$$

Now consider the following auxiliary function:

$$\Phi(y) = A\Big[e^{\mu y} + (e^{\mu y_0} - e^{\mu y})\chi\Big(\mu\Big(y + \frac{s}{\varepsilon}\Big)\Big)\Big],$$

where

$$y_0 = -\frac{s}{\varepsilon} + \frac{1}{\mu}, \quad A = 2e(|\phi(0)| + c_0).$$

If $y \in (-\frac{s}{\varepsilon}, y_0)$, $\Phi(y) = Ae^{\mu y_0} \le Ae^{\mu y+1}$ and hence

$$\Phi'' + \frac{2\varepsilon}{s+\varepsilon y}\Phi' - (1 + \Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])\Phi = -A(1 + T_{\varepsilon,s}[U^2_{\varepsilon,s}])e^{\mu y_0} \le -c_0 e^{\mu y}.$$

If $y \in (-\frac{s}{\varepsilon} + \frac{2}{\mu}, 0)$, $\Phi(y) = Ae^{\mu y}$, and $\frac{\varepsilon}{s+\varepsilon y} \le \frac{\mu}{2}$, then

$$\Phi'' + \frac{2\varepsilon}{s+\varepsilon y}\Phi' - (1 + \Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])\Phi = A[2\mu^2 - (1 + T_{\varepsilon,s}[U^2_{\varepsilon,s}])]e^{\mu y} \le -c_0 e^{\mu y}$$

provided that μ is sufficiently small. Finally, it is easy to see that for $y \in (y_0, y_0 + \frac{1}{\mu})$,

$$e^{\mu y_0} \le e^{\mu y} \le e e^{\mu y_0}, \quad \Phi(y) \ge Ae^{\mu y - 1}, \quad \frac{\mu}{2} \le \frac{\varepsilon}{s+\varepsilon y} \le \mu;$$

hence

$$\Phi'' + \frac{2\varepsilon}{s+\varepsilon y}\Phi' - (1 + \Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])\Phi \le AO(\mu^2)e^{\mu y} - A(1 + \Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])e^{\mu y-1} \le -c_0 e^{\mu y}$$

provided that μ is sufficiently small.

In any case, we have that for $y \in (-\frac{s}{\varepsilon}, 0)$, $\Phi(y)$ satisfies

$$\begin{gathered} \Phi'' + \frac{2\varepsilon}{s+\varepsilon y}\Phi' - (1 + \Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])\Phi \le -c_0 e^{\mu y}, \\ \Phi'\left(-\frac{s}{\varepsilon}\right) = 0, \quad \Phi(0) \ge |\phi(0)|. \end{gathered} \tag{3.6}$$

Combining (3.6) with the hypothesis we obtain

$$(\Phi - \phi)''(y) + \frac{2\varepsilon}{s+\varepsilon y}(\Phi - \phi)'(y) - (1 + \Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])(\Phi - \phi)(y) \le 0 \tag{3.7}$$

for every $y \in \left[-\frac{s}{\varepsilon}, 0\right]$ and

$$(\Phi - \phi)(0) > 0, \quad (\Phi - \phi)'\left(-\frac{s}{\varepsilon}\right) = 0;$$

we claim that $(\Phi - \phi)(y) \ge 0$ for $y \in [-\frac{s}{\varepsilon}, 0)$. Otherwise, if we call $\bar{y}$ the minimum point of $\Phi - \phi$ in $[-\frac{s}{\varepsilon}, 0)$, then it would be $(\Phi - \phi)(\bar{y}) < 0$ and $(\Phi - \phi)'(\bar{y}) = 0$, $(\Phi - \phi)''(\bar{y}) > 0$, in contradiction to (3.7). Hence we have proved that $\phi \le \Phi$ in $[-\frac{s}{\varepsilon}, 0]$. On the other hand, by (3.6) and the hypothesis we also get

$$(\Phi + \phi)''(y) + \frac{2\varepsilon}{s+\varepsilon y}(\Phi + \phi)'(y) - (1 + \Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])(\Phi + \phi)(y) \le 0 \ \ \forall y \in \left[-\frac{s}{\varepsilon}, 0\right]$$

and $(\Phi + \phi)(0) > 0$, $(\Phi + \phi)'(-\frac{s}{\varepsilon}) = 0$. Proceeding as before we conclude $\phi \ge -\Phi$ in $[-\frac{s}{\varepsilon}, 0]$.

For $y \in [0, +\infty)$, we use $\hat{\Phi}(y) = Ae^{-\mu y}$ and hence

$$\hat{\Phi}'' + \frac{2\varepsilon}{s+\varepsilon y}\hat{\Phi}' - (1 + \Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])\hat{\Phi} \le A(\mu^2 - (1 + \Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}]))e^{-\mu y} \le -c_0 e^{-\mu y}$$

provided that μ is sufficiently small. By repeating the previous argument we obtain $|\phi| \leq \hat{\Phi}$ in $[0, +\infty)$ and the conclusion follows. $\square$

Let $\mu \in (0,1)$ be a number sufficiently small such that Lemma 3.1 holds and for every function $\phi : I_{\varepsilon,s} \to \mathbb{R}$ define

$$\|\phi\|_* = \|e^{\mu\langle y\rangle}\phi(y)\|_{L^\infty(I_{\varepsilon,s})},$$

where $\langle y\rangle = (1+y^2)^{1/2}$. Since $\frac{2\varepsilon}{s+\varepsilon y}U''_{\varepsilon,s} = O(\varepsilon)e^{-|y|}$, by using (2.7), (2.8), and (3.2) we obtain

$$\begin{aligned} Z_{\varepsilon,s}(y) &= w_s'''(y) - (1+t(s))w_s'(y) + O(\varepsilon)e^{-\mu\langle y\rangle} \\ &= -pw_s^{p-1}(y)w_s'(y) + O(\varepsilon)e^{-\mu\langle y\rangle} \end{aligned} \tag{3.8}$$

uniformly for $s \in [s_1, s_2]$.

Next we prove the following a priori estimate for (3.5).

LEMMA 3.2. *There exists a constant $C > 0$ such that, provided that ε is sufficiently small, if $s \in [s_1, s_2]$ and (ϕ, c, h) satisfy* (3.5)*, the following holds:*

$$\|\phi\|_* \leq C\|h\|_*.$$

Proof. We argue by contradiction. Assume the existence of a sequence $\varepsilon_k \to 0$, $(s_k) \subset [s_1, s_2]$ and $(\tilde{\phi}_k, \tilde{c}_k) \in H^1_\varepsilon(I_{\varepsilon_k,s_k}) \times \mathbb{R}$, $h_k \in L^2_\varepsilon(I_{\varepsilon_k,s_k}) \cap C(\bar{I}_{\varepsilon_k,s_k})$ satisfying (3.5) such that

$$\|\tilde{\phi}_k\|_* > k\|\tilde{h}_k\|_*.$$

In particular, $\|\tilde{h}_k\|_* < +\infty$ for every k. For the sake of simplicity, in the remaining part of the proof we replace the subscript (ε_k, s_k) with k. By standard regularity results $\tilde{\phi}_k \in C^2(\bar{I}_k)$; furthermore, for every k $\|T_k(U_k\tilde{\phi}_k)\|_\infty \leq \int_{I_k}(s_k+\varepsilon_k z)U_k\tilde{\phi}_k dz < +\infty$ and, consequently, $\| -pU_k^{p-1}\tilde{\phi}_k + 2\Gamma T_k[U_k\tilde{\phi}_k]U_k + \tilde{h}_k + \tilde{c}_k Z_k\|_* < +\infty$. Then Lemma 3.1 implies $\|\tilde{\phi}_k\|_* < +\infty$ for every k. Hence it makes sense to set $\phi_k = \frac{\tilde{\phi}_k}{\|\tilde{\phi}_k\|_*}$, $c_k = \frac{\tilde{c}_k}{\|\tilde{\phi}_k\|_*}$, $h_k = \frac{\tilde{h}_k}{\|\tilde{\phi}_k\|_*}$. We obtain that (ϕ_k, c_k, h_k) satisfies (3.5) and

$$\|\phi_k\|_* = 1, \quad \|h_k\|_* = o(1).$$

Assume without loss of generality that $s_k \to s$ as $k \to +\infty$. By multiplying the equation $L_k[\phi_k] = h_k + c_k Z_k$ by U_k' and integrating over I_k, we get

$$c_k \int_{I_k} Z_k U_k' dy = -\int_{I_k} h_k U_k' dy + \int_{I_k} L_k[\phi_k]U_k' dy. \tag{3.9}$$

First examine the left-hand side of (3.9): by using (2.8) and (3.8) we deduce

$$\int_{I_k} Z_k U_k' dy = -p\int_{\mathbb{R}} w_{s_k}^{p-1}(w_{s_k}')^2 dy + o(1) = -p\int_{\mathbb{R}} w_s^{p-1}(w_s')^2 dy + o(1). \tag{3.10}$$

The first term on the right-hand side of (3.9) can be estimated as

$$\int_{I_k} |h_k U_k'| dy \leq \|h_k\|_* \int_{\mathbb{R}} |U_k'| e^{\mu\langle y\rangle} dy = O(1)\|h_k\|_* = o(1). \tag{3.11}$$

The last term in (3.9) equals

$$
\begin{aligned}
&\int_{I_{\varepsilon_k}} (L_k[\phi_k])U_k' dy \\
&= \int_{I_k} \phi_k \left[U_k''' + \frac{2\varepsilon_k^2}{(s_k+\varepsilon_k y)^2} U_k' - \frac{2\varepsilon_k}{s_k+\varepsilon_k y} U_k'' - (1+\Gamma T_k[U_k^2])U_k' + pU_k^{p-1}U_k' \right] dy \\
&\quad -2\Gamma \int_{I_k} T_k[U_k\phi_k]U_k U_k' dy \\
&= \int_{I_k} \phi_k \Big[U_k'''(y) - (1+\Gamma T_k[U_k^2])U_k'(y) + pU_k^{p-1}U_k' \Big] dy + O(\varepsilon_k)\|\phi_k\|_* \\
&\quad -2\Gamma \int_{I_k} T_k[U_k\phi_k]U_k U_k' dy
\end{aligned}
$$

since, according to the definition of U_k in (2.6) and to (2.8), $\frac{2\varepsilon_k^2}{(s_k+\varepsilon_k y)^2}U_k' = O(\varepsilon_k)e^{-|y|}$ and $\frac{2\varepsilon_k}{s_k+\varepsilon_k y}U_k' = O(\varepsilon_k)e^{-|y|}$. By (2.7), (2.8), and (3.2) we obtain

$$
\int_{I_k} \phi_k \big[U_k'''(y) - (1+\Gamma T_k[U_k^2])U_k' + pU_k^{p-1}U_k' \big] dy = O(\varepsilon_k)\|\phi_k\|_*. \tag{3.12}
$$

On the other hand, by (2.8) and (3.1) we deduce

$$
T_k[U_k\phi_k](y) = s\int_{I_k} U_k\phi_k dy + O(\varepsilon_k)(|y|+1)\|\phi_k\|_*, \tag{3.13}
$$

by which

$$
\int_{I_k} T_k[U_k\phi_k]U_k U_k' dy = O(\varepsilon_k)\|\phi_k\|_* \int_{I_k} (|y|+1)|U_k U_k'| dy = O(\varepsilon_k)\|\phi_k\|_*.
$$

Hence we have proved that

$$
\int_{I_k} L_k[\phi_k]U_k' dz = O(\varepsilon_k)\|\phi_k\|_* = O(\varepsilon_k). \tag{3.14}
$$

Combining (3.9), (3.10), (3.11), and (3.14) we achieve $|c_k| = o(1)$, by which $\|h_k + c_k Z_k\|_* = o(1)$.

Next we claim that $\phi_k \to 0$ uniformly in any compact interval of $\mathbb{R}$.

By multiplying (3.5) by $(s_k+\varepsilon_k y)^2\phi_k$ and integrating by parts we immediately get that the sequence $(\phi_k,\phi_k)_\varepsilon = \int_{I_k}(s_k+\varepsilon_k y)^2(|\phi_k'|^2+|\phi_k|^2)dy$ is bounded. Now we consider $\bar\phi_k(y) = \phi_k(y)\eta(\varepsilon_k y) \in H^1(\mathbb{R})$, where η has been defined in (2.5). Then it is easy to see that $\bar\phi_k$ is bounded in $H^1(\mathbb{R})$, and hence $\bar\phi_k \to \bar\phi_0$ weakly in $H^1(\mathbb{R})$. Taking into account of (3.2) and (3.13), $\bar\phi_0$ satisfies

$$
\Delta\bar\phi_0 - (1+t(s))\bar\phi_0 + pw_s^{p-1}\bar\phi_0 - 2s\Gamma\left(\int_{\mathbb{R}} w_s\bar\phi_0 dy\right) w_s = 0 \text{ in } \mathbb{R}, \ \ |\bar\phi_0| \le e^{-\mu\langle y\rangle}.
$$

By rescaling, setting $\phi_0(y) = \bar\phi_0(\frac{y}{\sqrt{1+t(s)}})$, and recalling the definition of w_s in (2.6), we obtain

$$
\phi_0'' - \phi_0 + pw^{p-1}\phi_0 - 2\frac{t(s)}{1+t(s)}\frac{\int_{\mathbb{R}} w\phi_0 dy}{\int_{\mathbb{R}} w^2 dy} w = 0.
$$

By (2.4), because of the choice of s,

$$2\frac{t(s)}{1+t(s)} \neq \frac{4(p-1)}{5-p};$$

Lemma 2.4 implies $\phi_0 = cw'$. On the other hand, $\int_{I_k} \phi_k Z_k(s_k+\varepsilon_k y)^2 dy = 0$ and hence, since by (3.8) $Z_k(s_k+\varepsilon_k y)^2 \to -s^2 p w_s^{p-1} w_s'$ for every $y \in \mathbb{R}$, by Lebesgue's dominated convergence theorem $\int_{\mathbb{R}} \bar{\phi}_0 w_s^{p-1} w_s' = 0$, that is, $\int_{\mathbb{R}} \phi_0 w^{p-1} w' = 0$, by which $c = 0$; we have proved that $\phi_k \to 0$ in any compact interval of $\mathbb{R}$. Taking into account (2.8), this shows that

$$\|U_k^{p-1}\phi_k\|_* \leq \sup_{y\in I_k} |e^{\mu\langle y\rangle} w_{s_k}^{p-1}(y)\phi_k(y)| = o(1).$$

By using again Lebesgue's dominated convergence theorem, we have that

$$\int_{I_k} U_k \phi_k dy \to 0,$$

which implies, using (3.13),

$$\big\|T_k[U_k\phi_k]U_k\big\|_* = o(1).$$

Thus we have arrived at the following situation: ϕ_k satisfies

$$\phi_k^{''} + \frac{2\varepsilon_k}{s_k+\varepsilon_k y}\phi_k^{'} - (1+\Gamma T_k[U_k^2])\phi_k = o(1)e^{-\mu\langle y\rangle},$$

$$\phi_k^{'}\Big(-\frac{s_k}{\varepsilon_k}\Big) = 0, \quad \phi_k \to 0 \text{ as } y \to \infty.$$

Furthermore, $\phi_k(0) = o(1)$. Hence we can apply Lemma 3.1 and obtain that $\phi_k(y) = o(1)e^{-\mu\langle y\rangle}$, which is a contradiction since $\|\phi_k\|_* = 1$. This proves the lemma. ∎

Now we are in position to provide the existence of a solution for the system (3.5).

LEMMA 3.3. *For $\varepsilon > 0$ sufficiently small, for every $s \in [s_1, s_2]$ and $h \in L_\varepsilon^2(I_{\varepsilon,s}) \cap C(\bar{I}_{\varepsilon,s})$, there exists a unique pair (ϕ, c) solving* (3.5). *Moreover, by Lemma* 3.2,

$$\|\phi\|_* \leq C\|h\|_*. \tag{3.15}$$

Proof. The existence follows from Fredholm alternative. To this aim, let us consider $\mathcal{H}_\varepsilon$ the closed subset of $H_\varepsilon^1(I_{\varepsilon,s})$ defined by

$$\mathcal{H}_\varepsilon = \{u \in H_\varepsilon^1(I_{\varepsilon,s}) \,|\, (u, U_{\varepsilon,s}^{'})_\varepsilon = 0\}.$$

Notice that, by (3.4), ϕ solves the system (3.5) if and only if $\phi \in \mathcal{H}_\varepsilon$ and

$$(\phi,\psi)_\varepsilon - p\langle U_{\varepsilon,s}^{p-1}\phi,\psi\rangle_\varepsilon + 2\Gamma\langle T_{\varepsilon,s}[U_{\varepsilon,s}\phi]U_{\varepsilon,s},\psi\rangle_\varepsilon = -\langle h,\psi\rangle_\varepsilon \ \ \forall\psi \in \mathcal{H}_\varepsilon. \tag{3.16}$$

Once we know ϕ we can determine a unique c from the equation

$$-p\langle U_{\varepsilon,s}^{p-1}\phi, U_{\varepsilon,s}'\rangle_\varepsilon + 2\Gamma\langle T_{\varepsilon,s}[U_{\varepsilon,s}\phi]U_{\varepsilon,s}, U_{\varepsilon,s}'\rangle_\varepsilon = -\langle h, U_{\varepsilon,s}'\rangle_\varepsilon + c(U_{\varepsilon,s}', U_{\varepsilon,s}')_\varepsilon.$$

Thus it remains to solve (3.16). According to Riesz's representation theorem, take $\mathcal{K}_\varepsilon(\phi)$, $\bar{h} \in \mathcal{H}_\varepsilon$ such that

$$(\mathcal{K}_\varepsilon(\phi),\psi)_\varepsilon = -p\langle U_{\varepsilon,s}^{p-1}\phi,\psi\rangle_\varepsilon + 2\Gamma\langle T_{\varepsilon,s}[U_{\varepsilon,s}\phi]U_{\varepsilon,s},\psi\rangle_\varepsilon \ \ \forall\psi \in \mathcal{H}_\varepsilon,$$

$$(\bar{h}, \psi)_\varepsilon = -\langle h, \psi \rangle_\varepsilon \ \ \forall \psi \in \mathcal{H}_\varepsilon.$$

Then problem (3.16) consists in finding $\phi \in \mathcal{H}_\varepsilon$ such that

$$\phi + \mathcal{K}_\varepsilon(\phi) = \bar{h}. \tag{3.17}$$

It is easy to prove that $\mathcal{K}_\varepsilon$ is a linear compact operator from $\mathcal{H}_\varepsilon$ to $\mathcal{H}_\varepsilon$.

Using Fredholm's alternatives, (3.17) has a unique solution for each $\bar{h}$ if and only if (3.17) has a unique solution for $\bar{h} = 0$. Let $\phi \in \mathcal{H}_\varepsilon$ be a solution of $\phi + \mathcal{K}_\varepsilon(\phi) = 0$; then ϕ solves the system (3.5) with $h = 0$ for some $c \in \mathbb{R}$. Lemma 3.2 implies $\phi \equiv 0$. ◻

4. Finite dimensional reduction. This section is devoted to solving the following nonlinear system with the unknowns $(\phi, \beta) \in H^1_\varepsilon(I_{\varepsilon,s}) \times \mathbb{R}$:

$$\begin{cases} S_{\varepsilon,s}[U_{\varepsilon,s} + \phi] = \beta Z_{\varepsilon,s}, & y \in I_{\varepsilon,s}, \\ \phi'\left(-\frac{s}{\varepsilon}\right) = 0, & \langle \phi, Z_{\varepsilon,s} \rangle_\varepsilon = 0, \end{cases} \tag{4.1}$$

where

$$S_{\varepsilon,s}[\psi] = \psi'' + \frac{2\varepsilon}{s + \varepsilon y}\psi' - \psi + \psi_+^p - \Gamma \psi T_{\varepsilon,s}[\psi^2].$$

LEMMA 4.1. *Fix $\sigma \in (\frac{1}{2}, 1)$. Provided that $\varepsilon > 0$ is sufficiently small, for every $s \in [s_1, s_2]$ there is a unique pair $(\phi_{\varepsilon,s}, \beta_{\varepsilon,s}) \in H^1_\varepsilon(I_{\varepsilon,s}) \times \mathbb{R}$ satisfying* (4.1) *and*

$$\|\phi_{\varepsilon,s}\|_* \le \varepsilon^\sigma, \ \ (\phi_{\varepsilon,s}, \phi_{\varepsilon,s})_\varepsilon \le \varepsilon^\sigma. \tag{4.2}$$

Furthermore, setting $\Phi_{\varepsilon,s}(x) = \phi_{\varepsilon,s}\big(|x| - \frac{s}{\varepsilon}\big)$, the maps $\Phi_\varepsilon : s \in [s_1, s_2] \mapsto \Phi_{\varepsilon,s} \in H^1_r(\mathbb{R}^3)$ and $\beta_\varepsilon : s \in [s_1, s_2] \mapsto \beta_{\varepsilon,s} \in \mathbb{R}$ are continuous.

Proof. We write the equation in (4.1) in the form

$$L_{\varepsilon,s}[\phi] = E_{\varepsilon,s} + M_{\varepsilon,s}[\phi] + \beta Z_{\varepsilon,s} \tag{4.3}$$

and use contraction mapping theorem. Here

$$-E_{\varepsilon,s} = U''_{\varepsilon,s} + \frac{2\varepsilon}{s + \varepsilon y} U'_{\varepsilon,s} - (1 + \Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])U_{\varepsilon,s} + U^p_{\varepsilon,s} \tag{4.4}$$

and

$$M_{\varepsilon,s}[\phi] = -(U_{\varepsilon,s} + \phi)^p_+ + U^p_{\varepsilon,s} + pU^{p-1}_{\varepsilon,s}\phi + \Gamma\big[(U_{\varepsilon,s} + \phi)T_{\varepsilon,s}[\phi^2] + 2\phi T_{\varepsilon,s}[U_{\varepsilon,s}\phi]\big].$$

By (2.7), (2.8), and (3.2) we immediately obtain

$$\|E_{\varepsilon,s}\|_* = \|U''_{\varepsilon,s} - (1 + t(s))U_{\varepsilon,s} + U^p_{\varepsilon,s}\|_* + C\varepsilon \le C\varepsilon \ \ \forall s \in [s_1, s_2], \tag{4.5}$$

where we have also used the estimate $\frac{2\varepsilon}{s+\varepsilon y}U'_{\varepsilon,s} = O(\varepsilon)e^{-|y|}$ uniformly for $s \in [s_1, s_2]$. Now we have to estimate the term $M_{\varepsilon,s}[\phi]$. Set $\mathcal{B}_{\varepsilon,s} = \{\phi \in C(\bar{I}_{\varepsilon,s}) \,|\, \|\phi\|_* \le \varepsilon^\sigma\}$. Given ϕ_1, $\phi_2 \in \mathcal{B}_{\varepsilon,s}$, we compute

$$\begin{aligned} &\|-(U_{\varepsilon,s} + \phi_1)^p_+ + pU^{p-1}_{\varepsilon,s}\phi_1 + (U_{\varepsilon,s} + \phi_2)^p_+ - pU^{p-1}_{\varepsilon,s}\phi_2\|_* \\ &\qquad \le \|\phi_1 - \phi_2\|_* \sup_{\xi \in \mathcal{B}_{\varepsilon,s}} \|p(U_{\varepsilon,s} + \xi)^{p-1}_+ - pU^{p-1}_{\varepsilon,s}\|_\infty \\ &\qquad \le C\varepsilon^{\sigma(p-1)}\|\phi_1 - \phi_2\|_* \ \ \forall s \in [s_1, s_2]. \end{aligned}$$

For every $\phi \in C(\bar{I}_{\varepsilon,s})$: $|T_{\varepsilon,s}[\phi](y)| \le \|\phi\|_* \int_{I_{\varepsilon,s}} (s+\varepsilon z)e^{-\mu\langle z\rangle} dz \le C\|\phi\|_*$,

$$\begin{aligned}\|(U_{\varepsilon,s}+\phi_1)T_{\varepsilon,s}[\phi_1^2] - (U_{\varepsilon,s}+\phi_2)T_{\varepsilon,s}[\phi_2^2]\|_* \\ \le \|(U_{\varepsilon,s}+\phi_1)(T_{\varepsilon,s}[\phi_1^2-\phi_2^2])\|_* + \|(\phi_1-\phi_2)T_{\varepsilon,s}[\phi_2^2]\|_* \\ \le C\|\phi_1^2-\phi_2^2\|_* + C\|\phi_1-\phi_2\|_*\varepsilon^\sigma \le C\|\phi_1-\phi_2\|_*\varepsilon^\sigma \quad \forall s \in [s_1,s_2].\end{aligned}$$

In a similar way

$$\|\phi_1 T_{\varepsilon,s}[U_{\varepsilon,s}\phi_1] - \phi_2 T_{\varepsilon,s}[U_{\varepsilon,s}\phi_2]\|_* \le C\|\phi_1-\phi_2\|_*\varepsilon^\sigma \quad \forall s \in [s_1,s_2],$$

by which

$$\|M_{\varepsilon,s}(\phi_1) - M_{\varepsilon,s}(\phi_2)\|_* \le C(\varepsilon^\sigma + \varepsilon^{\sigma(p-1)})\|\phi_1-\phi_2\|_* \quad \forall \phi_1,\phi_2 \in \mathcal{B}_{\varepsilon,s}, \ \forall s \in [s_1,s_2]. \tag{4.6}$$

According to Lemma 3.3, for every $\phi \in \mathcal{B}_{\varepsilon,s}$ we define $\mathcal{A}_{\varepsilon,s}[\phi] \in H^1_\varepsilon(I_{\varepsilon,s})$ to be the unique solution to the system (3.5) with $h = E_{\varepsilon,s} + M_{\varepsilon,s}[\phi]$. Then by Lemma 3.3 we have

$$\|\mathcal{A}_{\varepsilon,s}[\phi]\|_* \le C\|E_{\varepsilon,s} + M_{\varepsilon,s}[\phi]\|_* \le C(\varepsilon + \varepsilon^{2\sigma} + \varepsilon^{\sigma p}) \le \varepsilon^\sigma \quad \forall s \in [s_1,s_2]$$

and hence $\mathcal{A}_{\varepsilon,s}[\phi] \in \mathcal{B}_{\varepsilon,s}$. Moreover, since $\mathcal{A}_{\varepsilon,s}[\phi_1] - \mathcal{A}_{\varepsilon,s}[\phi_2]$ solves the system (3.5) with $h = M_{\varepsilon,s}[\phi_1] - M_{\varepsilon,s}[\phi_2]$, by (4.6) we also have that

$$\|\mathcal{A}_\varepsilon[\phi_1] - \mathcal{A}_\varepsilon[\phi_2]\|_* \le C\|M_{\varepsilon,s}[\phi_1] - M_{\varepsilon,s}[\phi_2]\|_* < \|\phi_1-\phi_2\|_* \quad \forall s \in [s_1,s_2]; \tag{4.7}$$

i.e., the map $\mathcal{A}_{\varepsilon,s}$ is a contraction map from $\mathcal{B}_{\varepsilon,s}$ to $\mathcal{B}_{\varepsilon,s}$. By the contraction mapping theorem, (4.1) has a unique solution $(\phi_{\varepsilon,s}, \beta_{\varepsilon,s}) \in \mathcal{B}_{\varepsilon,s} \times \mathbb{R}$. By multiplying both members of (4.3) by $(s+\varepsilon z)^2\phi_{\varepsilon,s}$ and integrating by parts we obtain

$$(\phi_{\varepsilon,s}, \phi_{\varepsilon,s})_\varepsilon = \langle pU_{\varepsilon,s}^{p-1}\phi_{\varepsilon,s} - 2\Gamma T_{\varepsilon,s}[U_{\varepsilon,s}\phi_{\varepsilon,s}]U_{\varepsilon,s} - E_{\varepsilon,s} - M_{\varepsilon,s}[\phi_{\varepsilon,s}], \phi_{\varepsilon,s}\rangle_\varepsilon \le \varepsilon^\sigma, \tag{4.8}$$

where we have used (4.5) and (4.6); in particular, since $\|\Phi_{\varepsilon,s}\|^2_{H^1(\mathbb{R}^3)} \le \varepsilon^{-2}4\pi(\phi_{\varepsilon,s}, \phi_{\varepsilon,s})_\varepsilon$, fixed $\varepsilon > 0$, the family $\Phi_{\varepsilon,s}$ is bounded in $H^1_r(\mathbb{R}^3)$. On the other hand, by using (3.4) we have

$$\begin{aligned}(\phi_{\varepsilon,s}, U'_{\varepsilon,s})_\varepsilon = &\langle pU_{\varepsilon,s}^{p-1}\phi_{\varepsilon,s} - 2\Gamma T_{\varepsilon,s}[U_{\varepsilon,s}\phi_{\varepsilon,s}]U_{\varepsilon,s} - E_{\varepsilon,s} - M_{\varepsilon,s}[\phi_{\varepsilon,s}], U'_{\varepsilon,s}\rangle_\varepsilon \\ &+\beta_{\varepsilon,s}(U'_{\varepsilon,s}, U'_{\varepsilon,s})_\varepsilon,\end{aligned}$$

by which we deduce that the family $\beta_{\varepsilon,s}$ is also bounded for fixed $\varepsilon > 0$.

Now consider $\{s_n\} \subset [s_1,s_2]$ such that $s_n \to \bar{s}$. Up to a subsequence, $\Phi_{\varepsilon,s_n} \rightharpoonup \bar{\Phi}$ weakly in $H^1_r(\mathbb{R}^3)$ and $\beta_{\varepsilon,s_n} \to \bar{\beta}$. Then, setting $\bar{\phi}(y) = \bar{\Phi}(y + \frac{\bar{s}}{\varepsilon})$ for $y \in I_{\varepsilon,\bar{s}}$, $(\bar{\phi}, \bar{\beta})$ solves the equation

$$L_{\varepsilon,\bar{s}}(\bar{\phi}) = E_{\varepsilon,\bar{s}} + M_{\varepsilon,\bar{s}}[\bar{\phi}] + \bar{\beta}Z_{\varepsilon,\bar{s}}, \quad \bar{\phi}'\left(\frac{\bar{s}}{\varepsilon}\right) = 0, \quad \langle\bar{\phi}, Z_{\varepsilon,\bar{s}}\rangle = 0, \quad \|\bar{\phi}\|_* \le \varepsilon^\sigma.$$

From uniqueness it follows that $\bar{\phi} = \phi_{\varepsilon,\bar{s}}$ and $\bar{\beta} = \beta_{\varepsilon,\bar{s}}$. By (4.8) we get

$$(\phi_{\varepsilon,s_n}, \phi_{\varepsilon,s_n})_\varepsilon \to \langle pU_{\varepsilon,\bar{s}}^{p-1}\bar{\phi} - 2\Gamma T_{\varepsilon,\bar{s}}[U_{\varepsilon,\bar{s}}\bar{\phi}]U_{\varepsilon,\bar{s}} - E_{\varepsilon,\bar{s}} - M_{\varepsilon,\bar{s}}[\bar{\phi}], \bar{\phi}\rangle_\varepsilon = (\bar{\phi}, \bar{\phi})_\varepsilon,$$

hence (since the norm $(\phi,\phi)_\varepsilon$ is equivalent to $\|\Phi\|_{H^1(\mathbb{R}^3)}$), we deduce $\Phi_{\varepsilon,s_n}\to\Phi_{\varepsilon,\bar s}$ in $H^1(\mathbb{R}^3)$. ∎

LEMMA 4.2. *For $\varepsilon>0$ sufficiently small, the map $\Phi_\varepsilon : s\in[s_1,s_2]\mapsto\Phi_{\varepsilon,s}\in H^1_r(\mathbb{R}^3)$ constructed in Lemma 4.1 is C^1.*

Proof. Consider the following map $H_\varepsilon:[s_1,s_2]\times H^1_r(\mathbb{R}^3)\times\mathbb{R}\to H^1_r(\mathbb{R}^3)\times\mathbb{R}$ of class C^1:

$$H_\varepsilon(s,\Phi,\beta)=\begin{pmatrix}\Big(\Delta-(1+\Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])\Big)^{-1}\Big(S_{\varepsilon,s}[U_{\varepsilon,s}+\phi]-\beta Z_{\varepsilon,s}\Big)(|x|-\frac{s}{\varepsilon})\\ (\phi,U'_{\varepsilon,s})_\varepsilon\end{pmatrix},$$

where $\Phi(x)=\phi(|x|-\frac{s}{\varepsilon})$ and $v=\big(\Delta-(1+\Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])\big)^{-1}(h)$ is defined as the unique solution in $H^1_\varepsilon(I_{\varepsilon,s})$ of

$$v''+\frac{2\varepsilon}{s+\varepsilon y}v'-\big(1+\Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}]\big)v=h,\ \ v'\Big(-\frac{s}{\varepsilon}\Big)=0,\ \ y\in I_{\varepsilon,s}. \tag{4.9}$$

It is immediate that (ϕ,β) solves the system (4.1) if and only if $H_\varepsilon(s,\Phi,\beta)=0$. We are going to prove that, provided that ε is sufficiently small, for every $s\in[s_1,s_2]$ the linear operator

$$\frac{\partial H_\varepsilon(s,\Phi,\beta)}{\partial(\Phi,\beta)}|_{(s,\Phi_{\varepsilon,s},\beta_{\varepsilon,s})}:H^1_r(\mathbb{R}^3)\times\mathbb{R}\to H^1_r(\mathbb{R}^3)\times\mathbb{R}$$

is invertible. But first notice how, assuming this, the thesis easily follows; indeed, the continuity of the map Φ_ε proved in Lemma 4.1 implies that Φ_ε actually coincides with the implicit function associated to H_ε, hence the C^1-regularity of Φ_ε will follow from the implicit function theorem.

Now we compute

$$\frac{\partial H_\varepsilon(s,\Phi,\beta)}{\partial(\Phi,\beta)}|_{(s,\Phi_{\varepsilon,s},\beta_{\varepsilon,s})}[\Phi,c]$$
$$=\begin{pmatrix}\Big(\Delta-(1+\Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}])\Big)^{-1}\Big(S'_{\varepsilon,s}[U_{\varepsilon,s}+\phi_{\varepsilon,s}](\phi)-cZ_{\varepsilon,s}\Big)(|x|-\frac{s}{\varepsilon})\\ (\phi,U'_{\varepsilon,s})_\varepsilon\end{pmatrix}.$$

Proceeding as in the proof of Lemma 3.3, (Φ,c) solves the system

$$\frac{\partial H_\varepsilon(s,\Phi,\beta)}{\partial(\Phi,\beta)}|_{(s,\Phi_{\varepsilon,s},\beta_{\varepsilon,s})}[\Phi,c]=(\Theta,d)$$

if and only if $\phi=\bar\phi+\frac{d}{(U'_{\varepsilon,s},U'_{\varepsilon,s})_\varepsilon}U'_{\varepsilon,s}$, where $\bar\phi\in\mathcal{H}_\varepsilon$ ($\mathcal{H}_\varepsilon$ is the closed subset of $H^1_\varepsilon(I_{\varepsilon,s})$ defined in Lemma 3.3) and

$$\Big(\big(\Delta-\big(1+\Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}]\big)\big)^{-1}\big(S'_{\varepsilon,s}[U_{\varepsilon,s}+\phi_{\varepsilon,s}](\phi)-cZ_{\varepsilon,s}\big),\psi\Big)_\varepsilon=(\theta,\psi)_\varepsilon\ \ \forall\psi\in\mathcal{H}_\varepsilon, \tag{4.10}$$

where $\Theta(x)=\theta(|x|-\frac{s}{\varepsilon})$. By (4.9) we have $((\Delta-(1+\Gamma T_{\varepsilon,s}[U^2_{\varepsilon,s}]))^{-1}(h),\psi)_\varepsilon=-\langle h,\psi\rangle_\varepsilon$; hence (4.10) may be rewritten as

$$\begin{aligned}(\bar\phi,\psi)_\varepsilon-&p\langle(U_{\varepsilon,s}+\phi_{\varepsilon,s})^{p-1}_+\phi,\psi\rangle_\varepsilon+\Gamma\langle T_{\varepsilon,s}[\phi^2_{\varepsilon,s}+2U_{\varepsilon,s}\phi_{\varepsilon,s}]\phi,\psi\rangle_\varepsilon\\ &+2\Gamma\langle(U_{\varepsilon,s}+\phi_{\varepsilon,s})T_{\varepsilon,s}[(U_{\varepsilon,s}+\phi_{\varepsilon,s})\phi],\psi\rangle_\varepsilon=(\theta,\psi)_\varepsilon\ \ \forall\psi\in\mathcal{H}_\varepsilon.\end{aligned}$$

Once we know $\bar\phi$, the related c is given by the identity

$$\begin{aligned} d-&p\langle (U_{\varepsilon,s}+\phi_{\varepsilon,s})_+^{p-1}\phi, U'_{\varepsilon,s}\rangle_\varepsilon + \Gamma d'\langle T_{\varepsilon,s}[\phi_{\varepsilon,s}^2+2U_{\varepsilon,s}\phi_{\varepsilon,s}]\phi, U'_{\varepsilon,s}\rangle_\varepsilon \\ &+2\Gamma d'\langle (U_{\varepsilon,s}+\phi_{\varepsilon,s})T_{\varepsilon,s}[(U_{\varepsilon,s}+\phi_{\varepsilon,s})\phi], U'_{\varepsilon,s}\rangle_\varepsilon = c(U'_{\varepsilon,s},U'_{\varepsilon,s})_\varepsilon + (\theta, U'_{\varepsilon,s})_\varepsilon, \end{aligned}$$

where we have set $d' = \frac{d}{(U'_{\varepsilon,s},U'_{\varepsilon,s})_\varepsilon}$. According to Riesz's representation theorem, take $\mathcal{W}_{\varepsilon,t}(\bar\phi)$, $\bar\theta \in \mathcal{H}_\varepsilon$ such that

$$\begin{aligned} (\mathcal{W}_{\varepsilon,t}(\bar\phi),\psi)_\varepsilon =& -p\langle (U_{\varepsilon,t}+\phi_{\varepsilon,t})_+^{p-1}\bar\phi,\psi\rangle_\varepsilon + \Gamma\langle T_\varepsilon[\phi_{\varepsilon,t}^2+2U_{\varepsilon,t}\phi_{\varepsilon,t}]\bar\phi,\psi\rangle_\varepsilon \\ &+2\Gamma\langle (U_{\varepsilon,t}+\phi_{\varepsilon,t})T_\varepsilon[(U_{\varepsilon,t}+\phi_{\varepsilon,t})\bar\phi],\psi\rangle_\varepsilon \quad \forall\psi\in\mathcal{H}_\varepsilon, \end{aligned}$$

$$\begin{aligned} (\bar\theta,\psi)_\varepsilon =& -pd'\langle (U_{\varepsilon,t}+\phi_{\varepsilon,t})_+^{p-1}U'_{\varepsilon,t},\psi\rangle + \Gamma d'\langle T_\varepsilon[\phi_{\varepsilon,t}^2+2U_{\varepsilon,t}\phi_{\varepsilon,t}]U'_{\varepsilon,t},\psi\rangle_\varepsilon \\ &+2\Gamma d'\langle (U_{\varepsilon,t}+\phi_{\varepsilon,t})T_\varepsilon[(U_{\varepsilon,t}+\phi_{\varepsilon,t})U'_{\varepsilon,t}],\psi\rangle_\varepsilon \quad \forall\psi\in\mathcal{H}_\varepsilon. \end{aligned}$$

Then problem (4.10) consists in finding $\bar\phi \in \mathcal{H}_\varepsilon$ such that

$$\bar\phi + \mathcal{W}_\varepsilon(\bar\phi) = \theta - \bar\theta. \tag{4.11}$$

Since $\|\phi_{\varepsilon,s}\|_* \le \varepsilon^\sigma$, combining the definition of $\mathcal{W}_\varepsilon$ with that of $\mathcal{K}_\varepsilon$ (see Lemma 3.3), when $\varepsilon \to 0$ we have

$$\mathcal{W}_\varepsilon - \mathcal{K}_\varepsilon \to 0.$$

Since we have proved that $I+\mathcal{K}_\varepsilon$ is invertible, then the theory of the linear operator assures the invertibility of $I+\mathcal{W}_\varepsilon$ for small ε. This concludes the proof of the lemma. □

5. The reduced energy functional: Proof of Theorem 1.1. For every $s \in [s_1,s_2]$ set

$$v_{\varepsilon,s}(x) = U_{\varepsilon,s}\Big(|x|-\frac{s}{\varepsilon}\Big) + \phi_{\varepsilon,s}\Big(|x|-\frac{s}{\varepsilon}\Big),$$

where $\phi_{\varepsilon,s}$ has been constructed in Lemma 4.1, and consider the function

$$K_\varepsilon : [s_1,s_2] \to \mathbb{R}, \quad K_\varepsilon(s) := \frac{\varepsilon^2}{4\pi}E_\varepsilon[v_{\varepsilon,s}].$$

First we provide the following estimate.

Lemma 5.1. *For $s \in [s_1,s_2]$, we have*

$$K_\varepsilon(s) = \frac{1}{8(p+1)}\frac{\int_{\mathbb{R}} w^{p+1}dy}{(\int_{\mathbb{R}} w^2dy)^2}\rho(t(s)) + o(1), \tag{5.1}$$

where $\rho(t)$ and $t(s)$ have been defined in Lemma 2.2.

Proof. We compute

$$\begin{aligned} &K_\varepsilon(s) \\ &= \int_{-\frac{s}{\varepsilon}}^{+\infty}\left[\frac12|(U_{\varepsilon,s}+\phi_{\varepsilon,s})'|^2 + \frac12|U_{\varepsilon,s}+\phi_{\varepsilon,s}|^2 - \frac{1}{p+1}(U_{\varepsilon,s}+\phi_{\varepsilon,s})_+^{p+1}\right](s+\varepsilon y)^2dy \\ &\quad + \frac{\Gamma}{4}\int_{-\frac{s}{\varepsilon}}^{\infty}(U_{\varepsilon,s}+\phi_{\varepsilon,s})^2T_{\varepsilon,s}[(U_{\varepsilon,s}+\phi_{\varepsilon,s})^2](s+\varepsilon y)^2dy \\ &= \frac{s^2}{2}\int_{\mathbb{R}}|w_s'|^2dy + \frac{s^2}{2}\int_{\mathbb{R}}w_s^2dy - \frac{s^2}{p+1}\int_{\mathbb{R}}w_s^{p+1}dy + \frac{s^2}{4}t(s)\int_{\mathbb{R}}w_s^2dy + o(1), \end{aligned}$$

where we have used (2.8), (3.2), and the estimates $\|\phi_{\varepsilon,s}\|_*$, $(\phi_{\varepsilon,s},\phi_{\varepsilon,s})_\varepsilon \le \varepsilon^\sigma$ given by Lemma 4.1. Now we use (2.9) to obtain

$$\begin{aligned}
K_\varepsilon(s) &= s^2\frac{(3p-7)t(s)+4(p-1)}{8(p+1)(1+t(s))}\int_{\mathbb{R}} w_s^{p+1}dy + o(1)\\
&= s^2\frac{(3p-7)t(s)+4(p-1)}{8(p+1)}(1+t(s))^{\frac{5-p}{2(p-1)}}\int_{\mathbb{R}} w^{p+1}dy + o(1)\\
&= s^2\frac{\rho(t(s))}{8t(s)^2(p+1)}(1+t(s))^{\frac{5-p}{p-1}}\int_{\mathbb{R}} w^{p+1}dy + o(1)\\
&= \frac{\rho(t(s))}{8(p+1)}\frac{\int_{\mathbb{R}} w^{p+1}dy}{\left(\Gamma\int_{\mathbb{R}} w^2dy\right)^2} + o(1). \qquad \square
\end{aligned}$$

COROLLARY 5.2. *For ε sufficiently small, the function $s\in[s_1,s_2]\mapsto K_\varepsilon(s)$ has a minimum $s_\varepsilon = s_0 + o(1)$.*

Proof. This follows immediately from the previous lemma. By Lemma 2.2 the function $\rho(t(s))$ has a unique local minimum point $s_0\in[s_1,s_2]$ and, moreover, s_0 is nondegenerate; thus, by (5.1) and by the continuity of $K_\varepsilon(s)$, for ε small enough the minimum point of K_ε is attained at some s_ε which tends to s_0 as $\varepsilon\to 0$. $\square$

Proof of Theorem 1.1. Fix $\varepsilon>0$ sufficiently small such that Lemma 4.1 and Lemma 4.2 hold for $\varepsilon\in(0,\varepsilon_0)$. According to Lemma 4.1, for every $\varepsilon\in(0,\varepsilon_0)$ and $s\in[s_1,s_2]$ the function $v_{\varepsilon,s}$ solves the equation in $\mathbb{R}^3$ as follows:

$$\Delta v_{\varepsilon,s}(x) - v_{\varepsilon,s}(x) + (v_{\varepsilon,s})_+^p(x) - \varepsilon\Gamma(-\Delta)^{-1}[v_{\varepsilon,s}^2](x) = \beta_{\varepsilon,s}Z_{\varepsilon,s}\Big(|x|-\frac{s}{\varepsilon}\Big). \tag{5.2}$$

Set $u_\varepsilon = v_{\varepsilon,s_\varepsilon}$, where $s_\varepsilon\in(s_1,s_2)$ is the minimum point of K_ε, according to Corollary 5.2. Then we have

$$\frac{d}{ds}E_\varepsilon(v_{\varepsilon,s})_{|_{s=s^\varepsilon}} = 0; \tag{5.3}$$

by using the C^1 regularity of the map $s\in[s_1,s_2]\mapsto\Phi_{\varepsilon,s}\in H_r^1(\mathbb{R}^3)$, (5.3) can be rewritten as

$$\int_{\mathbb{R}^3}\left[\nabla u_\varepsilon\nabla\frac{d}{ds}v_{\varepsilon,s} + u_\varepsilon\frac{d}{ds}v_{\varepsilon,s} - (u_\varepsilon)_+^p\frac{d}{ds}v_{\varepsilon,s} + \varepsilon\Gamma u_\varepsilon(-\Delta)^{-1}[u_\varepsilon^2]\frac{d}{ds}v_{\varepsilon,s}\right]dx\bigg|_{s=s^\varepsilon} = 0,$$

which is equivalent, by (5.2), to

$$\beta_{\varepsilon,s_\varepsilon}\int_{\mathbb{R}^3} Z_{\varepsilon,s_\varepsilon}\Big(|x|-\frac{s_\varepsilon}{\varepsilon}\Big)\frac{d}{ds}v_{\varepsilon,s}dx\bigg|_{s=s_\varepsilon} = 0,$$

that is,

$$\beta_{\varepsilon,s_\varepsilon}\int_{I_{\varepsilon,s}} Z_{\varepsilon,s_\varepsilon}(y)\frac{d}{ds}(v_{\varepsilon,s})\bigg|_{s=s_\varepsilon}\Big(y+\frac{s_\varepsilon}{\varepsilon}\Big)(s_\varepsilon+\varepsilon y)^2dy = 0. \tag{5.4}$$

Notice that, by (2.8), we easily compute

$$\begin{aligned}
&\frac{d}{ds}\Big[U_{\varepsilon,s}\Big(\cdot-\frac{s}{\varepsilon}\Big)\Big]\Big(y+\frac{s}{\varepsilon}\Big)\\
&\quad= \frac{t'(s)}{(p-1)(1+t(s))}w_s(y) + w_s'(y)\left(\frac{t'(s)y}{2\sqrt{1+t(s)}} - \frac{\sqrt{1+t(s)}}{\varepsilon}\right) + o(\varepsilon)e^{-\mu\langle y\rangle}\\
&\quad= -\frac{\sqrt{1+t(s)}}{\varepsilon}w_s'(y) + O(1)e^{-\mu\langle y\rangle}.
\end{aligned} \tag{5.5}$$

Differentiating the equation $\int_{\mathbb{R}^3} \Phi_{\varepsilon,s} Z_{\varepsilon,s}(|x|-\frac{s}{\varepsilon})dx = \varepsilon^{-2}\langle \phi_{\varepsilon,s}, Z_{\varepsilon,s}\rangle_\varepsilon = 0$ with respect to s, we get

$$\int_{I_{\varepsilon,s}} Z_{\varepsilon,s}\frac{d}{ds}(\Phi_{\varepsilon,s})\Big(y+\frac{s}{\varepsilon}\Big)(s+\varepsilon y)^2 dy = -\int_{I_{\varepsilon,s}} \phi_{\varepsilon,s}\frac{d}{ds}\Big[Z_{\varepsilon,s}\Big(\cdot-\frac{s}{\varepsilon}\Big)\Big]\Big(y+\frac{s}{\varepsilon}\Big)(s+\varepsilon y)^2 dy.$$

By computing similarly as in (5.5) we get

$$\frac{d}{ds}\Big[Z_{\varepsilon,s}\Big(\cdot-\frac{s}{\varepsilon}\Big)\Big]\Big(y+\frac{s}{\varepsilon}\Big) = O(\varepsilon^{-1})e^{-\mu\langle y\rangle},$$

by which

$$\int_{I_{\varepsilon,s}} Z_{\varepsilon,s}\frac{d}{ds}(\Phi_{\varepsilon,s})\Big(y+\frac{s}{\varepsilon}\Big)(s+\varepsilon y)^2 dy = O(\varepsilon^{\sigma-1}). \tag{5.6}$$

Combining (3.8), (5.5), and (5.6) we obtain

$$\int_{I_{\varepsilon,s}} Z_{\varepsilon,s}\frac{d}{ds}(v_{\varepsilon,s})\Big(y+\frac{s}{\varepsilon}\Big)(s+\varepsilon y)^2 dy = -\frac{s^2 p\sqrt{1+t(s)}}{\varepsilon}\int_{\mathbb{R}} w_s^{p-1}(w_s')^2 dy + O(\varepsilon^{\sigma-1}).$$

In particular, $\int_{I_{\varepsilon,s_\varepsilon}} Z_{\varepsilon,s_\varepsilon}\frac{d}{ds}(v_{\varepsilon,s})\big|_{s=s_\varepsilon}\big(y+\frac{s_\varepsilon}{\varepsilon}\big)(s_\varepsilon+\varepsilon y)^2 dy \neq 0$ for ε sufficiently small, then (5.4) implies $\beta_{\varepsilon,s_\varepsilon}=0$. Hence u_ε solves the equation

$$\Delta u_\varepsilon - u_\varepsilon + (u_\varepsilon)_+^p - \varepsilon\Gamma u_\varepsilon(-\Delta)^{-1}[u_\varepsilon^2] = 0 \text{ in } \mathbb{R}^3. \tag{5.7}$$

By multiplying both members of (5.7) by $(u_\varepsilon)_-$ and integrating by parts we get

$$\int_{\mathbb{R}^3}|\nabla(u_\varepsilon)_-|^2 dx + \int_{\mathbb{R}^3}|(u_\varepsilon)_-|^2 + \varepsilon\Gamma\int_{\mathbb{R}^3}(u_\varepsilon)_-^2(-\Delta)^{-1}[u_\varepsilon^2]dx = 0,$$

by which $(u_\varepsilon)_-=0$. From the strong maximum principle it follows that $u_\varepsilon>0$. Hence u_ε and $\psi_\varepsilon=\varepsilon(-\Delta)^{-1}[u_\varepsilon^2]$ actually solve the system (2.1). In order to obtain (2) of Theorem 1.1, using (2.8) and the estimate $\|\phi_{\varepsilon,s}\|_*\leq\varepsilon^\sigma$ we obtain

$$\begin{aligned}\frac{1}{\sqrt{\varepsilon}}u_\varepsilon\Big(\frac{r}{\varepsilon}\Big) &= \frac{1}{\sqrt{\varepsilon}}U_{\varepsilon,s_\varepsilon}\Big(\frac{r-s_\varepsilon}{\varepsilon}\Big)+\frac{1}{\sqrt{\varepsilon}}\phi_{\varepsilon,s_\varepsilon}\Big(\frac{r-s_\varepsilon}{\varepsilon}\Big)\\ &= \frac{1}{\sqrt{\varepsilon}}(1+t(s_\varepsilon))^{\frac{1}{p-1}}w\Big(\sqrt{1+t(s_\varepsilon)}\frac{r-s_\varepsilon}{\varepsilon}\Big)+o(1)\exp\Big(-\mu\frac{|r-s_\varepsilon|}{\varepsilon}\Big).\end{aligned}$$

Finally, notice that for every $r,\rho\in(0,+\infty)$ it results in $|G\big(r,\rho\big)-G(r,s_\varepsilon)|\leq|\rho-s_\varepsilon|\frac{1}{s_\varepsilon}\Big(2+\frac{\rho}{s_\varepsilon}\Big)G(r,s_\varepsilon)$, by which

$$\begin{aligned}\psi_\varepsilon\Big(\frac{r}{\varepsilon}\Big) &= \frac{1}{\varepsilon}\int_0^{+\infty}G\big(r,\rho\big)u_\varepsilon^2\Big(\frac{\rho}{\varepsilon}\Big)d\rho\\ &= \frac{1}{\varepsilon}G(r,s_\varepsilon)\int_0^{+\infty}u_\varepsilon^2\Big(\frac{\rho}{\varepsilon}\Big)d\rho+O(1)G(r,s_\varepsilon)\frac{1}{\varepsilon s_\varepsilon}\int_0^{+\infty}|\rho-s_\varepsilon|\Big(2+\frac{\rho}{s_\varepsilon}\Big)u_\varepsilon^2\Big(\frac{\rho}{\varepsilon}\Big)d\rho\\ &= G(r,s_\varepsilon)\left((1+t(s_\varepsilon))^{\frac{5-p}{2(p-1)}}\int_{\mathbb{R}}w^2 d\rho+o(1)\right).\end{aligned}$$

Taking into account (2.2), Theorem 1.1 is proved. ∎

REMARK 5.3. *In order to better understand the presence of the critical power* $\frac{11}{7}$*, it could be useful to look at the system* (1.1) *in higher dimension* $N \geq 3$*, apart from the physical model under consideration. Proceeding similarly as in the case* $N = 3$*, we can construct the family of functions* $v^N_{\varepsilon,s}$ *lying in the normal direction of the approximated solution surface, and we can prove that the following asymptotic expansion of the energy holds:*

$$\frac{\varepsilon^2}{\omega_N} E_\varepsilon[v^N_{\varepsilon,s}] = \frac{1}{8(p+1)} \frac{\int_{\mathbb{R}} w^{p+1} dy}{(\int_{\mathbb{R}} w^2 dy)^2} s^{N-3} \rho(t(s)) + o(1),$$

where ω_N *denotes the surface measure of the unit sphere of* $\mathbb{R}^N$*. Then, by using Lemma* 2.2*, we deduce that the dimension-dependent critical number appearing in higher dimension is given by* $\frac{7N-10}{3N-2}$ *and, repeating the same computations as in* $N = 3$*, for every* $1 < p < \frac{7N-10}{3N-2}$ *we can prove the existence of a family of solutions concentrating on the sphere whose radius is given by the unique minimum point* s_0 *of the function* $s^{N-3}\rho(t(s))$ *in* $(0, s(\hat{t}))$*.*

REMARK 5.4. *It is possible to prove the existence of radial solutions to the system* (1.1) *for every* $1 < p < 5$ *by a simple rescaling argument; indeed, it is sufficient to minimize the functional*

$$T(u) = \frac{1}{2} \int_{\mathbb{R}^3} \left(|\nabla u|^2 + u^2\right) dx + \frac{\omega}{4} \int_{\mathbb{R}^3} u^2 (-\Delta)^{-1} [4\pi\omega u^2]\, dx$$

over the manifold

$$\mathcal{M} := \left\{ u \in H^1_r \;\Big|\; \int_{\mathbb{R}^3} |u|^{p+1} dx = 1 \right\}.$$

According to the compact injection $H^1_r \hookrightarrow L^p$ *for* $1 < p < 5$*,* $\mathcal{M}$ *is a compact manifold, while* T *is weakly lower semicontinuous on* H^1_r*. Hence we easily get the existence of a minimizing function* u*. The constraint causes a Lagrange multiplier to appear, and one obtains that* u *solves* $-\Delta u + u + \omega u(-\Delta)^{-1}[4\pi\omega u^2] = \gamma |u|^{p-1} u$*. We notice that since* $T(u) = T(|u|)$*, we may assume that* $u \geq 0$*; then, from the strong maximum principle,* $u > 0$*. Setting* $v_\hbar(x) = \hbar^{-1} u(\frac{x}{\hbar})$ *and* $\phi_\hbar(x) = (-\Delta)^{-1}[4\pi\omega v_\hbar^2](x) = (-\Delta)^{-1}[4\pi\omega u^2](\frac{x}{\hbar})$*, we obtain that* $(v_\hbar, \phi_\hbar)$ *solves* (1.1) *with* $\gamma_\hbar = \gamma \hbar^{p-1}$*.*

REFERENCES

[1] A. AMBROSETTI, M. BADIALE, AND S. CINGOLANI, *Semiclassical states of nonlinear Schrödinger equations*, Arch. Ration. Mech. Anal., 140 (1997), pp. 285–300.

[2] A. AMBROSETTI, A. MALCHIODI, AND W. M. NI, *Singularly perturbed elliptic equations with symmetry: Existence of solutions concentrating on spheres, Part* I, Comm. Math. Phys., 235 (2003), pp. 427–466.

[3] A. AMBROSETTI, A. MALCHIODI, AND W. M. NI, *Singularly perturbed elliptic equations with symmetry: Existence of solutions concentrating on spheres, Part* II, Indiana Univ. Math. J., 53 (2004), pp. 297–329.

[4] M. BADIALE AND T. D'APRILE, *Concentration around a sphere for a singularly perturbed Schrödinger equation*, Nonlinear Anal., 49 (2002), pp. 947–985.

[5] V. BENCI AND T. D'APRILE, *The semiclassical limit of the nonlinear Schrödinger equation in a radial potential*, J. Differential Equations, 184 (2002), pp. 109–138.

[6] V. BENCI AND D. FORTUNATO, *An eigenvalue problem for the Schrödinger-Maxwell equations*, Top. Meth. Nonlinear Anal., 11 (1998), pp. 283–293.

[7] H. BERESTYCKI AND P. L. LIONS, *Nonlinear scalar field equations,* I: *Existence of a ground state*, Arch. Ration. Mech. Anal., 82 (1983), pp. 313–345.

[8] G. M. Coclite and V. Georgiev, *Solitary waves for Maxwell-Schrödinger equations*, Electron. J. Differential Equations, 94 (2004).
[9] T. D'Aprile, *On a class of solutions with non-vanishing angular momentum for nonlinear Schrödinger equations*, Differential Integral Equations, 16 (2003), pp. 349–384.
[10] T. D'Aprile, *Semiclassical states for the nonlinear Schrödinger equation with the electromagnetic field*, Nonlinear Differential Equations Appl., to appear.
[11] T. D'Aprile and D. Mugnai, *Existence of solitary waves for the nonlinear Klein-Gordon Maxwell and Schrödinger-Maxwell equations*, Proc. Roy. Soc. Edinburgh Sect. A, 134 (2004), pp. 893–906.
[12] M. Del Pino and P. Felmer, *Local mountain passes for semilinear elliptic problems in unbounded domains*, Calc. Var. Partial Differential Equations, 4 (1996), pp. 121–137.
[13] M. Del Pino and P. Felmer, *Multi-peak bound states for nonlinear Schrödinger equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 15 (1998), pp. 127–149.
[14] M. Del Pino and P. Felmer, *Semi-classical states for nonlinear Schrödinger equations*, J. Funct. Anal., 149 (1997), pp. 245–265.
[15] M. Del Pino and P. Felmer, *Semi-classical states of nonlinear Schrödinger equations: A variational reduction method*, Math. Ann., 324 (2002), pp. 1–32.
[16] M. Del Pino, P. L. Felmer, and J. Wei, *On the role of mean curvature in some singularly perturbed Neumann problems*, SIAM J. Math. Anal., 31 (1999), pp. 63–79.
[17] A. Floer and A. Weinstein, *Nonspreading wave packets for the cubic Schrödinger equation with a bounded potential*, J. Funct. Anal., 69 (1986), pp. 397–408.
[18] M. Grossi, *Some results on a class of nonlinear Schrödinger equations*, Math. Z., 235 (2000), pp. 687–705.
[19] M. Grossi, A. Pistoia, and J. Wei, *Existence of multipeak solutions for a semilinear Neumann problem via nonsmooth critical point theory*, Calc. Var. Partial Differential Equations, 11 (2000), pp. 143–175.
[20] C. Gui, *Multipeak solutions for a semilinear Neumann problem*, Duke Math. J., 84 (1996), pp. 739–769.
[21] C. Gui and J. Wei, *Multiple interior peak solutions for some singularly perturbed Neumann problems*, J. Differential Equations, 158 (1999), pp. 1–27.
[22] C. Gui and J. Wei, *On multiple mixed interior and boundary peak solutions for some singularly perturbed Neumann problems*, Canad. J. Math., 52 (2000), pp. 522–538.
[23] C. Gui, J. Wei, and M. Winter, *Multiple boundary peak solutions for some singularly perturbed Neumann problems*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 17 (2000), pp. 47–82.
[24] X. Kang and J. Wei, *On interacting bumps of semiclassical states of nonlinear Schrödinger equations*, Adv. Differential Equations, 5 (2000), pp. 899–928.
[25] Y. Y. Li, *On a singularly perturbed elliptic equation*, Adv. Differential Equations, 6 (1997), pp. 955–980.
[26] Y. Y. Li, *On a singularly perturbed equation with Neumann boundary conditions*, Comm. Partial Differential Equations, 23 (1998), pp. 487–545.
[27] A. Malchiodi, *Concentration at Curves for a Singularly Perturbed Neumann Problem in Three-Dimensional Domains*, preprint, 2003.
[28] A. Malchiodi and M. Montenegro, *Multidimensional boundary layers for a singularly perturbed Neumann problem*, Duke Math. J., 124 (2004), pp. 105–143.
[29] A. Malchiodi and M. Montenegro, *Boundary concentration phenomena for a singularly perturbed elliptic problem*, Comm. Pure Appl. Math., 55 (2002), pp. 1507–1568.
[30] A. Malchiodi, W. M. Ni, and J. Wei, *Multiple clustered layer solutions for semilinear Neumann problems on a ball*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 22 (2005), pp. 143–163.
[31] W. M. Ni and I. Takagi, *On the shape of least-energy solutions to a semi-linear Neumann problem*, Comm. Pure Appl. Math., 44 (1991), pp. 819–851.
[32] W. M. Ni and I. Takagi, *Locating the peaks of least-energy solutions to a semilinear Neumann problem*, Duke Math. J., 70 (1993), pp. 247–281.
[33] W. M. Ni, I. Takagi, and J. Wei, *On the location and profile of spike-layer solutions to a singularly perturbed semilinear Dirichlet problem: Intermediate solutions*, Duke Math. J., 94 (1998), pp. 597–618.
[34] W. M. Ni and J. Wei, *On Ground State Solutions Concentrating on Spheres for the Gierer-Meinhardt System*, preprint, 2003.
[35] W. M. Ni and J. Wei, *On the location and profile of spike-layer solutions to singularly perturbed semilinear Dirichlet problems*, Comm. Pure Appl. Math., 48 (1995), pp. 723–761.
[36] Y. J. Oh, *Existence of semi-classical bound states of nonlinear Schrödinger equation with potential on the class* $(V)_a$, Comm. Partial Differential Equations, 13 (1998), pp. 1499–1519.

[37] Y. J. Oh, *On positive multi-lump bound states of nonlinear Schrödinger equation under multiple well potential*, Comm. Math. Phys., 131 (1990), pp. 223–253.
[38] P. Rabinowitz, *On a class of nonlinear Schrödinger equations*, Z. Angew. Math. Phys., 43 (1992), pp. 270–291.
[39] W. A. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149–162.
[40] X. Wang, *On concentration of positive bound states of nonlinear Schrödinger equations*, Comm. Math. Phys., 153 (1993), pp. 229–244.
[41] J. Wei, *On the boundary spike layer solutions of a singularly perturbed Neumann problem*, J. Differential Equations, 134 (1997), pp. 104–133.
[42] J. Wei, *On the construction of single-peaked solutions to a singularly perturbed semilinear Dirichlet problem*, J. Differential Equations, 129 (1996), pp. 315–333.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 343–346

LACK OF COMPACTNESS IN TWO-SCALE CONVERGENCE*

MARC BRIANE† AND JUAN CASADO-DÍAZ‡

Abstract. This article deals with the links between compensated compactness and two-scale convergence. More precisely, we ask the following question: Is the div-curl compactness assumption sufficient to pass to the limit in a product of two sequences which two-scale converge with respect to the pair of variables $(x, x/\varepsilon)$? We reply in the negative. Indeed, the div-curl assumption allows us to control oscillations which are faster than $1/\varepsilon$ but not the slower ones.

Key words. two-scale convergence, compensated compactness, counterexample

AMS subject classifications. 35B27, 35B40

DOI. 10.1137/040621351

1. Introduction. In order to study the asymptotic behavior of periodic problems arising in homogenization theory, Nguetseng introduced in [7] (see also Allaire [1]) the notion of two-scale convergence:

Let Ω be a bounded open subset of $\mathbb{R}^d$, $Y := (-\frac{1}{2}, \frac{1}{2})^d$, and let M be a positive integer. A bounded sequence u_ε in $L^1_{\text{loc}}(\Omega)^M$ two-scale converges to a function $\hat{u}$ in $L^1_{\text{loc}}(\Omega \times \mathbb{R}^d)^M$ and Y-periodic with respect to the last variable if, for any $\psi \in C_c^\infty(\Omega, C_\#^\infty(Y))^M$, we have

$$\lim_{\varepsilon \to 0} \int_\Omega u_\varepsilon(x)\, \psi\left(x, \frac{x}{\varepsilon}\right) dx = \int_\Omega \int_Y \hat{u}(x,y)\, \psi(x,y)\, dx\, dy. \tag{1.1}$$

A compactness theorem due to Nguetseng [7] establishes that if u_ε is bounded in $L^p(\Omega)^M$, then there exists a subsequence of u_ε which two-scale converges to $\hat{u} \in L^p(\Omega; L^p_\#(Y))^M$.

Taking in (1.1) $\psi(x,y)$ independent of y, we deduce that if u_ε two-scale converges to $\hat{u}$, then it converges weakly in $L^p(\Omega)^M$ to $u := \int_Y \hat{u}(x,y)\, dy$. On the other hand, if u_ε strongly converges to u in $L^1(\Omega)^M$, then it also two-scale converges to u. Therefore two-scale convergence is stronger than weak convergence and weaker than the strong one. Moreover, it provides an expression of the limit of the product $u_\varepsilon\, \psi(x, \frac{x}{\varepsilon})$ of (1.1) in which each term only weakly converges.

In the periodic homogenization we usually deal with a sequence u_ε which is not only bounded in $L^p(\Omega)^M$ but whose some combinations of its derivatives are also bounded. In this context, let us recall that if u_ε converges weakly in $W^{1,p}(\Omega)^M$, for $1 \le p < +\infty$, to a function u, then it converges strongly in $L^p_{\text{loc}}(\Omega)^M$ ($L^p(\Omega)^M$ if Ω smooth) and so u_ε two-scale converges to u. Then we can conjecture that the classical results of the compensated compactness theory due to Murat and Tartar (see, e.g., [6] and [8]), and in particular the div-curl theorem, still hold true when we replace the weak convergence in $L^p(\Omega)^M$ with two-scale convergence. In fact we have the following result:

*Received by the editors December 13, 2004; accepted for publication (in revised form) December 28, 2004; published electronically October 14, 2005.
http://www.siam.org/journals/sima/37-2/62135.html

†Centre de Mathématiques, I.N.S.A. de Rennes & I.R.M.A.R., 35043 Rennes cedex, France (mbriane@insa-rennes.fr).

‡Dpto. de Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla, 41012 Sevilla, Spain (jcasadod@us.es).

PROPOSITION 1.1. *Let $(Y, Y_1, \dots, Y_n)$ be $(n+1)$ parallelotops of $\mathbb{R}^d$ of Lebesgue measure equal to 1, and let U, V be two vector-valued functions in $L^2(\Omega; C_\#(Y \times Y_1 \times \cdots \times Y_n))^d$, where $C_\#(Y \times Y_1 \times \cdots \times Y_n)$ denotes the set of the continuous functions on $(\mathbb{R}^d)^{n+1}$ which are Y-periodic with respect to the variable y and Y_k-periodic with respect to the variable y_k for any $k = 1, \dots, n$. Let $\varepsilon_k = \varepsilon_k(\varepsilon)$ for $k = 1, \dots, n$ be n well-ordered scales such that*

$$\lim_{\varepsilon \to 0} \frac{\varepsilon_1}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{\varepsilon_{k+1}}{\varepsilon_k} = 0 \quad \textit{for any } k = 1, \dots, n-1. \tag{1.2}$$

Consider the vector-valued sequences u_ε and v_ε defined by

$$u_\varepsilon(x) := U\left(x, \frac{x}{\varepsilon}, \frac{x}{\varepsilon_1}, \dots, \frac{x}{\varepsilon_n}\right) \quad \textit{and} \quad v_\varepsilon(x) := V\left(x, \frac{x}{\varepsilon}, \frac{x}{\varepsilon_1}, \dots, \frac{x}{\varepsilon_n}\right), \tag{1.3}$$

and assume that

$$\operatorname{div} u_\varepsilon \textit{ is compact in } H^{-1}(\Omega) \quad \textit{and} \quad \operatorname{curl} v_\varepsilon \textit{ is compact in } H^{-1}(\Omega)^{d\times d}. \tag{1.4}$$

Then the two-scale limits $\hat{u}$ of u_ε, $\hat{v}$ of v_ε, and $\hat{w}$ of $u_\varepsilon \cdot ? v_\varepsilon$ exist and satisfy

$$\hat{w} = \hat{u} \cdot \hat{v}. \tag{1.5}$$

Proposition 1.1 shows that the div-curl condition (1.4) implies some compactness in the two-scale convergence process (as in the classical case) when the oscillations of the sequences are faster than $\frac{1}{\varepsilon}$. Unfortunately, this is not the case for general sequences, particularly when the oscillations are slower than $\frac{1}{\varepsilon}$. This assertion follows from the following theorem, which is the main result of the present paper:

THEOREM 1.2. *Assume that $d \geq 2$. Then there exist two functions $U, V \in C^\infty_\#(2Y)^d$ such that the sequence $u_\varepsilon(x) := U(\frac{x}{\varepsilon})$ is divergence-free, the sequence $v_\varepsilon(x) := V(\frac{x}{\varepsilon})$ is curl-free, but the two-scale limits of u_ε, v_ε, and $u_\varepsilon \cdot v_\varepsilon$ do not satisfy* (1.5).

The key ingredient of this counterexample is that 2-periodic functions are considered although the test functions are 1-periodic.

In order to understand the lack of compactness in two-scale convergence, let us recall the equivalence between the two-scale convergence theory and the method introduced by Arbogast, Douglas, and Hornung [3] to study the oscillations of a sequence u_ε in $L^1_{\text{loc}}(\mathbb{R}^d)^M$. Their method consists in introducing the function $\hat{u}_\varepsilon : \mathbb{R}^d \times Y \to \mathbb{R}^M$ defined by

$$\hat{u}_\varepsilon(x, y) = \sum_{k \in \mathbb{Z}^d} 1_{\varepsilon k + \varepsilon Y}(x)\, u_\varepsilon(\varepsilon k + \varepsilon y). \tag{1.6}$$

The equivalence between the two approaches is then given by the following result (see, e.g., [5] and [4]):

THEOREM 1.3. *Assume that u_ε is bounded in $L^p(\Omega)^M$, with $1 < p < +\infty$. Then $\hat{u}_\varepsilon$ converges weakly to $\hat{u}$ in $L^p(\Omega; L^p(Y))^M$ if and only if u_ε two-scale converges to $\hat{u}$.*

The functions $\hat{u}_\varepsilon(x, y)$ are not continuous with respect to the variable x. If a combination of derivatives of u_ε is bounded, we also get a bound for the same combination of derivatives with respect to the variable y of $\hat{u}_\varepsilon$ but not with respect to the variable x. This explains the lack of compactness in two-scale convergence.

2. Proof of the results. In this section we prove Proposition 1.1 and Theorem 1.2.

Proof of Proposition 1.1. We follow the multiscale procedure of [2]. Thanks to the separation of scales (1.2) the sequences u_ε, v_ε, and $u_\varepsilon \cdot v_\varepsilon$, respectively, two-scale converge to $\hat{u} := \int_{Y_1} \cdots \int_{Y_n} U$, $\hat{v} := \int_{Y_1} \cdots \int_{Y_n} V$, and $\hat{w} := \int_{Y_1} \cdots \int_{Y_n} U \cdot V$. Putting test functions of type $\varepsilon_k\, \Phi(x, \frac{x}{\varepsilon}, \frac{x}{\varepsilon_1}, \dots, \frac{x}{\varepsilon_k})$ from $k = n$ to 1 in the div-curl assumption (1.4) implies that

$$\operatorname{div}_{y_k} \left(\int_{Y_{k+1}} \cdots \int_{Y_n} U \right) = 0 \text{ and } \operatorname{curl}_{y_k} \left(\int_{Y_{k+1}} \cdots \int_{Y_n} V \right) = 0 \quad \text{for } k = 1, \dots, n,$$

whence, integrating by parts the product of $\int_{Y_{k+1}} \cdots \int_{Y_n} U$ and $\int_{Y_{k+1}} \cdots \int_{Y_n} V$ (which is equal to the gradient in y_k of a periodic function plus a function depending only on the other variables $y_1, \dots, y_{k-1}$) successively from $k = n$ to 1, yields

$$\hat{w} = \int_{Y_1} \cdots \int_{Y_n} U \cdot V = \left(\int_{Y_1} \cdots \int_{Y_n} U \right) \cdot \left(\int_{Y_1} \cdots \int_{Y_n} V \right) = \hat{u} \cdot \hat{v},$$

which implies the desired equality (1.5).

Proof of Theorem 1.2. Let us consider two vector-valued functions $\Phi, \Psi \in C_c^\infty(Y)^d$ such that $\operatorname{div} \Phi = 0$, $\operatorname{curl} \Psi = 0$, and $\Phi \cdot \Psi \neq 0$ (this is possible since $d > 1$), which we extend to $\mathbb{R}^d$ by Y-periodicity. Let $\eta : \mathbb{R} \to \mathbb{R}$ be the 1-periodic function $\eta := \sum_{i \in \mathbb{Z}} 1_{(i - \frac{1}{4}, i + \frac{1}{4})}$ and let us define the following sequences

$$u_\varepsilon(x) := \eta \left(\frac{x_1}{2\,\varepsilon} \right) \Phi \left(\frac{x}{\varepsilon} \right) \quad \text{and} \quad v_\varepsilon(x) := \eta \left(\frac{x_1}{2\,\varepsilon} \right) \Psi \left(\frac{x}{\varepsilon} \right).$$

Since in each cube $\varepsilon k + \varepsilon Y$, for $k \in \mathbb{Z}^d$, $\eta(\frac{x_1}{2\,\varepsilon})$ is constant, and $\Phi(\frac{x}{\varepsilon})$, $\Psi(\frac{x}{\varepsilon})$ vanish on the boundary of $\varepsilon k + \varepsilon Y$, we have $u_\varepsilon, v_\varepsilon \in C^\infty(\mathbb{R}^N)$, $\operatorname{div} u_\varepsilon = 0$, and $\operatorname{curl} v_\varepsilon = 0$ in $\mathbb{R}^d$. Moreover, since $\eta(\frac{x_1}{2\,\varepsilon})$ is constant in $\varepsilon k + \varepsilon Y$ for any $k \in \mathbb{Z}^d$, it is invariant by the transformation (1.6). So we get

$$\hat{u}_\varepsilon(x, y) = \eta \left(\frac{x_1}{2\,\varepsilon} \right) \Phi(y), \quad \hat{v}_\varepsilon(x, y) = \eta \left(\frac{x_1}{2\,\varepsilon} \right) \Psi(y), \quad \widehat{u_\varepsilon \cdot v_\varepsilon}(x, y) = \eta^2 \left(\frac{x_1}{2\,\varepsilon} \right) \Phi(y) \cdot \Psi(y).$$

By Theorem 1.3 the two-scale limits $\hat{u}$ of u_ε, $\hat{v}$ of v_ε, and $\hat{w}$ of $u_\varepsilon \cdot v_\varepsilon$ are thus given by

$$\hat{u}(x, y) = \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \eta(s)\, ds \right) \Phi(y) = \frac{1}{2}\, \Phi(y), \quad \hat{v}(x, y) = \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \eta(s)\, ds \right) \Psi(y) = \frac{1}{2}\, \Psi(y),$$
$$\text{and} \quad \hat{w}(x, y) = \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \eta^2(s)\, ds \right) \Phi(y) \cdot \Psi(y) = \frac{1}{2}\, \Phi(y) \cdot \Psi(y),$$

whence $\hat{w} \neq \hat{u} \cdot \hat{v}$.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[2] G. Allaire and M. Briane, *Multiscale convergence and reiterated homogenization*, Proc. Roy. Soc. Edinburgh A, 126 (1996), pp. 297–342.

[3] T. Arbogast, J. Douglas, and U. Hornung, *Derivation of the double porosity model of single phase flow via homogenization theory*, SIAM J. Math. Anal., 21 (1990), pp. 823–836.

[4] D. Cioranescu, A. Damlamian, and G. Griso, *Periodic unfolding and homogenization*, C. R. Acad. Sci. Paris, 335 (2002), pp. 99–104.

[5] M. Lenczner, *Homogénéisation d'un circuit électrique*, C. R. Acad. Sci. Paris, 324 (1997), pp. 537–542.

[6] F. Murat, *Compacité par compensation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 5 (1978), pp. 489–507.

[7] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.

[8] L. Tartar, *Compensated compactness and applications to partial differential equations, nonlinear analysis and mechanics*, in Heriot-Watt Symposium IV, Pitman, San Francisco, 1979, pp. 136–212.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 347–371

UNIVERSAL BOUNDS ON COARSENING RATES FOR MEAN-FIELD MODELS OF PHASE TRANSITIONS*

SHIBIN DAI† AND ROBERT L. PEGO‡

Abstract. We prove one-sided universal bounds on coarsening rates for two kinds of mean-field models of phase transitions, one with a coarsening rate $l \sim t^{1/3}$ and the other with $l \sim t^{1/2}$. Here l is a characteristic length scale. These bounds are both proved by following a strategy developed by Kohn and Otto [*Comm. Math. Phys.*, 229 (2002), pp. 375–395]. The $l \sim t^{1/2}$ rate is proved using a new dissipation relation which extends the Kohn–Otto method. In both cases, the dissipation relations are subtle and their proofs are based on a residual lemma (Lagrange identity) for the Cauchy–Schwarz inequality.

Key words. Ostwald ripening, Lifshitz–Slyozov–Wagner equations, scaling exponents

AMS subject classifications. 82C26, 35B40, 35Q80, 74N20

DOI. 10.1137/040618047

1. Introduction.

In the late stages of heterogeneously nucleated phase transitions, a two-phase mixture is created, composed of particles of one phase dispersed in a matrix of the other. Initially the particles are small and their total surface area is large. According to thermodynamics, the system evolves in order to decrease the total surface area and conserve the total mass or volume of the particles. Smaller particles shrink and disappear and larger ones grow. As a result, the typical length scale that characterizes the particle size increases. It is widely observed that the length scale behaves as a temporal power law.

In this paper, we will try to give this power-law behavior a rigorous mathematical explanation in the context of mean-field models. In mean-field models, particles exchange mass by some interaction through a mean field $\theta(t)$ which is determined as a function of time t by the conservation of mass. There are many mechanisms that can dominate the mass transfer process [15]. We will consider two of them in this paper that correspond to two kinds of mean-field models with different power-law behaviors.

In the first model, particle growth is controlled by bulk or volume diffusion, with or without kinetic drag at the interface. Each particle radius R obeys the growth law

$$\dot{R} = \frac{1}{R+\beta}\left(\theta(t) - \frac{1}{R}\right), \tag{1.1}$$

where $\beta \geq 0$ is a constant. The particle size distribution $f(t, R)$ satisfies the transport equation

$$\partial_t f + \partial_R \left(\frac{1}{R+\beta}\left(\theta - \frac{1}{R}\right) f\right) = 0. \tag{1.2}$$

*Received by the editors November 1, 2004; accepted for publication January 11, 2005; published electronically October 14, 2005. This material is based upon work supported by the National Science Foundation under grant DMS 03-05985.
http://www.siam.org/journals/sima/37-2/61804.html

†Department of Mathematics, University of Maryland, College Park, MD 20742 (sdai@math.umd.edu).
‡Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (rpego@cmu.edu).

To conserve the total mass, the mean field θ satisfies

$$\theta(t) = \frac{\int_0^\infty (R+\beta)^{-1} R^{n-2} f(t,R)\, dR}{\int_0^\infty (R+\beta)^{-1} R^{n-1} f(t,R)\, dR}, \tag{1.3}$$

where n is the dimension of space. When $\beta = 0$, (1.1)–(1.3) is the classical model by Lifshitz and Slyozov [9] and Wagner [16]. Equation (1.1) is an approximation to the Mullins–Sekerka sharp-interface model with a modified Gibbs–Thomson law in the situation in which the particles are sparsely located in a domain Ω:

$$-\Delta u = 0 \quad \text{outside the particles,} \tag{1.4}$$

$$n \cdot \nabla u = V \quad \text{on } \Gamma, \tag{1.5}$$

$$u = \kappa + \beta V \quad \text{on } \Gamma. \tag{1.6}$$

Here Γ is the boundary of the particles, u is a chemical potential, n is the outer normal to Γ, κ is the mean curvature, and V is the normal velocity of Γ. Note that (1.6) is the Gibbs–Thomson law modified by a kinetic drag term βV. In [10, 11], Niethammer rigorously derived the model (1.1)–(1.3) in ^{3} for $\beta = 0$ and $\beta > 0$, respectively, from a model similar to (1.4)–(1.6) under the condition that the total capacity of the particles was small.

The second model arises formally by taking $\beta \to \infty$ after rescaling time by β. In this model, particle growth is controlled by the attachment reaction at the interface [1]. Now each particle radius R obeys the law

$$\dot{R} = \theta(t) - \frac{1}{R}. \tag{1.7}$$

The corresponding transport equation of the particle size distribution becomes

$$\partial_t f + \partial_R\left(\left(\theta - \frac{1}{R}\right) f\right) = 0. \tag{1.8}$$

In this case, the mean field θ satisfies

$$\theta(t) = \frac{\int_0^\infty R^{n-2} f(t,R)\, dR}{\int_0^\infty R^{n-1} f(t,R)\, dR}. \tag{1.9}$$

Equation (1.7) is the normalized mean curvature flow for a collection of spheres; i.e., it is a special case of the following sharp-interface model:

$$V = -\kappa + \frac{1}{|\Gamma|} \int_\Gamma \kappa\, dS, \tag{1.10}$$

where $|\Gamma|$ is the total area of the particle surface, κ is the mean curvature of the particle surface, and V is the normal velocity of the particle surface.

As in many systems, the coarsening rates for these mean-field models can be predicted by heuristic reasoning based on scaling invariance. The models we are considering are invariant under the scalings

$$R = \eta \hat{R}, \quad t = \eta^3 \hat{t}, \quad \theta = \eta^{-1} \hat{\theta} \quad \text{for (1.1) if } \beta = 0; \tag{1.11}$$

$$R = \eta \hat{R}, \quad t = \eta^2 \hat{t}, \quad \theta = \eta^{-1} \hat{\theta} \quad \text{for (1.7).} \tag{1.12}$$

If one expects that over long times the behavior of the coarsening system will appear scale invariant in some rough or statistical sense, then this kind of scaling invariance suggests that a characteristic length scale $l(t)$ ought to satisfy $l(t) = \eta l(t/\eta^p)$ with $p = 3$ or 2, respectively, so that $l(t)$ will be given by a temporal power law

$$l(t) \sim t^{1/3} \quad \text{for (1.1)}, \tag{1.13}$$

$$l(t) \sim t^{1/2} \quad \text{for (1.7)}. \tag{1.14}$$

When $\beta \neq 0$, under the scaling (1.11) equation (1.1) keeps its form if β is replaced by $\widehat{\beta} = \beta/\eta$. Then (1.1) is not invariant since we assume β to be a constant. However, this suggests that as the length scale becomes large, the influence of kinetic drag can be neglected and should not influence the ultimate coarsening rate for the volume-diffusion–controlled growth model.

In three dimensions ($n = 3$), the classical Lifshitz–Slyozov–Wagner (LSW) theory suggests that the size distribution function approaches a universal self-similar solution where the critical radius $R_c = \theta^{-1}$ follows the temporal power law $R_c \sim t^{1/3}$. Such a power law is observed in experiments. However, Niethammer and Pego [12] proved that, mathematically, for solutions of (1.2) the size distribution function does not necessarily converge to the predicted universal similarity solution, and the long-time behavior need not be self-similar.

Thus, the question is whether anything can be said universally about the coarsening rate of solutions in classical LSW theory. We cannot expect all solutions to coarsen at the same rate. For example, if initially all the particles are of the same size, then the system does not coarsen at all. One would like to be able to show that the expected power-law behavior is typical in some sense. What is "typical" is not clear, but a related question is whether it is possible that some solutions coarsen faster than expected. We know of no heuristic reason that would prevent such behavior.

Recently, Kohn and Otto [6] introduced a powerful method to answer this question. They obtain rigorous, universally valid time-averaged upper bounds on coarsening rates, in the setting of Cahn–Hilliard equations, which are diffuse-interface models for phase transitions (see also [3, 7, 8] for subsequently related results). Kohn and Otto consider the standard Cahn–Hilliard equation, whose sharp-interface limit is the Mullins–Sekerka model (1.4)–(1.6) with $\beta = 0$, and the Cahn–Hilliard equation with degenerate mobility, whose sharp interface limit is the surface diffusion model. Scaling invariance suggests that these two models have coarsening rates $l \sim t^{1/3}$ and $l \sim t^{1/4}$, respectively. Define $⨍_0^T := \frac{1}{T}\int_0^T$ to indicate the time-averaged integral. The results of Kohn and Otto, in their simplest form, are estimates of the following forms:

(i) $⨍_0^T E^2(t)\, dt \geq C_2 ⨍_0^T (t^{-1/3})^2\, dt$ for $T \geq C_3 L(0)^3$ (standard Cahn–Hilliard);

(ii) $⨍_0^T E^3(t)\, dt \geq C_2 ⨍_0^T (t^{-1/4})^3\, dt$ for $T \geq C_3 L(0)^4$ (Cahn–Hilliard with degenerate mobility).

Here E is the volume-averaged free energy, which is a decreasing function of time and scales as inverse to length, L is a "length scale" that is dual to E, and C_2 and C_3 are positive constants that depend only on the dimension of space n. Thus, these estimates are time-averaged versions of $E \geq C_2 t^{-1/3}$ and $E \geq C_2 t^{-1/4}$, respectively, which correspond to upper bounds on the length scale E^{-1}. These results show that, in a time-averaged sense, it is impossible for solutions to coarsen at a rate faster than the expected power law.

Our goal in this paper is to prove universal time-averaged upper bounds on corresponding coarsening rates for the mean-field models—see (2.4) and (2.5) below.

Again, we find that no solution can coarsen at a rate faster than that expected from scaling. The mean-field models that we study have three aspects that distinguish them from those models considered in [3, 6, 7, 8]:

(i) Mean-field models concern the evolution of dilute systems; i.e., the second phase consists of only a small fraction of the whole mixture. Kohn and Otto's analysis for the Cahn–Hilliard equations breaks down in this extreme case.

(ii) There is no spatial information and hence no pattern scale in mean-field models. This requires a different definition and interpretation of the dual length scale L.

(iii) For the normalized mean curvature flow (1.10), there is no result available for the corresponding diffuse-interface model—the conserved Allen–Cahn equation (see [14] for an asymptotic analysis of this correspondence).

To handle these differences, we will need to define all relevant quantities in terms of the distribution of particle radii. For the interface-reaction–controlled model, we will establish a new dissipation relation that extends the Kohn–Otto method and enables us to prove bounds that correspond to a coarsening rate of the form $l \sim t^{1/2}$. The proof of the dissipation relations in both mean-field models requires a different technique from previous works. A key ingredient in our proofs is the use of residual lemma (Lagrange identity) for the Cauchy–Schwarz inequality to compare the dissipation rates of E and L.

2. Strategy and main results. Let us describe our strategy for obtaining bounds on coarsening rates for the mean-field models (1.2) and (1.8) and state our main results. We work at first with a collection of finitely many particles undergoing coarsening with growth laws (1.1) and (1.7), respectively, for each particle. Such a system of particles has a discrete size distribution. We will apply a strategy similar to that of Kohn and Otto [6] to get time-averaged bounds for such discrete systems, and then pass to limits in section 5 to establish the bounds for arbitrary size distributions that have finite $(n+1)$st moment.

Kohn and Otto's strategy involves two quantities that measure length scales and three key steps. The first quantity is a volume-averaged free energy or negative entropy that decreases with time and scales as inverse to length. The second quantity scales like length, but its physical interpretation is not as clear. What is important is that, in a sense to be made precise, the second quantity is dual to the first one while at the same time being controlled by it.

In our situation, thermodynamics suggests that a natural quantity that is decreasing is the surface energy, which is proportional to the total surface area S of all the particles. Analogous to the cases considered in [3, 6, 7, 8], we will consider a kind of volume average of the surface area, which gives a quantity scaling as inverse to length. Because the total volume V of the particles is conserved, it is reasonable to consider the ratio S/V. For a finite particle system, we therefore define

$$E := \frac{\sum R_i^{n-1}}{\sum R_i^n}, \tag{2.1}$$

where n is the dimension of space and the sum goes over all surviving particles. E can also be considered as the volume-weighted average of curvatures $\{1/R_i\}$. In sections 3 and 4, we will prove that E is indeed decreasing in both models considered.

We need a length scale L that is dual to E. Since radius is dual to curvature, we define L to be the volume-weighted average of the radii $\{R_i\}$, i.e.,

$$L := \frac{\sum R_i^{n+1}}{\sum R_i^n}. \tag{2.2}$$

The first step of the Kohn–Otto method is to establish an *interpolation inequality* that expresses the duality of E and L. With the definitions (2.1) and (2.2) this is easy. By the Cauchy–Schwarz inequality,

$$\sum R_i^n = \sum R_i^{(n-1)/2} R_i^{(n+1)/2} \leq \left(\sum R_i^{n-1} \sum R_i^{n+1}\right)^{1/2},$$

and this immediately yields the required interpolation inequality,

$$EL \geq 1. \tag{2.3}$$

The second step is to obtain a *dissipation inequality* that controls $\dot{L}$ in terms of $\dot{E}$. In sections 3 and 4 below, we will prove that

$$\begin{aligned} |\dot{L}|^2 &\leq C_1\,(-\dot{E}) \quad \text{for volume-diffusion–controlled growth (1.1)},\\ |\dot{L}|^2 &\leq D_1\,(-\dot{E})\,L \quad \text{for interface-reaction–controlled growth (1.7)}, \end{aligned}$$

where C_1 and D_1 are positive constants depending only on the dimension of space n. We remark that in the cases considered in [3, 6, 7, 8], the difficult part is proving the interpolation inequalities; the dissipation relations are rather easy to prove. By contrast, in the situation of the mean-field models considered here, under definitions (2.1) and (2.2) the interpolation inequality is a simple consequence of the Cauchy–Schwarz inequality and it is the dissipation relations that need careful treatment.

The third step is an ODE argument. For the case of volume-diffusion–controlled growth, Lemma 3 in [6] and the two inequalities $EL \geq 1$ and $|\dot{L}|^2 \leq C_1(-\dot{E})$ directly give us appropriate time-averaged bounds on coarsening rates. Those that involve only E, the volume-averaged surface area, take a simple form, saying that for any $1 < p < 3$ there exist positive constants C_2 and C_3, depending only on n, p, and nothing else, such that

$$\fint_0^T E(t)^p\,dt \geq C_2 \fint_0^T \left(t^{-1/3}\right)^p\,dt \quad \text{for} \quad T \geq C_3 L(0)^3. \tag{2.4}$$

This is exactly a time-averaged version of $E \geq t^{-1/3}$, which corresponds to an upper bound on the "length scale" E^{-1}.

For the case of interface-reaction–controlled growth, we will establish an ODE lemma in section 4 to show that the inequalities $EL \geq 1$ and $|\dot{L}|^2 \leq D_1 L(-\dot{E})$ give us appropriate time-averaged estimates. In particular, for any $1 < p < 2$ there exist positive constants D_2 and D_3, depending only on n, p, and nothing else, such that

$$\fint_0^T E(t)^p\,dt \geq D_2 \fint_0^T \left(t^{-1/2}\right)^p\,dt \quad \text{for} \quad T \geq D_3 L(0)^2. \tag{2.5}$$

This is a time-averaged version of $E \geq t^{-1/2}$.

Once these results for discrete systems are established, we will pass to the case of general size distributions in section 5 by applying the well-posedness and compactness results for a family of mean-field models established by Niethammer and Pego in [13].

All of our models under consideration are included in that work except for the two-dimensional (2D) volume-diffusion–controlled growth model with $\beta = 0$. Thus this case is not included in our main theorems on coarsening rates for general size distributions.

The results in [13] enable us to approximate a general distribution by a sequence of discrete ones. These results, together with an extended moment compactness result proved here in an appendix, enable us to take limits in the estimates for the discrete sequence. This leads to our main results on coarsening rates for general size distributions.

We consider such size distributions to belong to $\mathcal{P}_n$, the set of Borel probability measures on $[0,\infty)$ with finite nth moment. Topologically we regard $\mathcal{P}_n$ as a subset of the Banach space of finite Radon measures on $[0,\infty)$, which is dual to $C_0([0,\infty))$, the space of continuous functions on $[0,\infty)$ that vanish at infinity. A *measure-valued solution* of the transport equation (1.2) or (1.8) is a weak-star continuous map $t \mapsto \nu_t$ taking $[0,\infty) \to \mathcal{P}_n$ that is a solution in the sense of distributions on $(0,\infty)\times(0,\infty)$. Based on the results in [13], we will see that for each initial size distribution $\mu \in \mathcal{P}_n$, there is a unique measure-valued solution with initial value $\nu_0 = \mu$ that preserves the nth moment (total volume). The corresponding mean field is given for a.e. $t > 0$ by

$$\theta(t) = \int_0^\infty \frac{R^{n-2}}{R+\beta} d\nu_t(R) \Big/ \int_0^\infty \frac{R^{n-1}}{R+\beta} d\nu_t(R) \tag{2.6}$$

in the case of volume-diffusion–controlled growth and

$$\theta(t) = \int_0^\infty R^{n-2} d\nu_t(R) \Big/ \int_0^\infty R^{n-1} d\nu_t(R) \tag{2.7}$$

in the case of interface-reaction–controlled growth. The quantities corresponding to (2.1) and (2.2) are defined by

$$E(t) := \int_0^\infty R^{n-1}\, d\nu_t(R) \Big/ \int_0^\infty R^n\, d\nu_t(R), \tag{2.8}$$

$$L(t) := \int_0^\infty R^{n+1}\, d\nu_t(R) \Big/ \int_0^\infty R^n\, d\nu_t(R). \tag{2.9}$$

Our main results take the following form.

THEOREM 2.1 (volume-diffusion–controlled growth). *Let $n \geq 2$ be an integer and $\beta \geq 0$, with $\beta > 0$ if $n = 2$, and let p be real with $1 < p < 3$. Then there exist positive constants C_2 and C_3, depending on p, n, and nothing else, such that whenever ν is a measure-valued solution of the transport equation* (1.2) *and ν_0 has finite nth and $(n+1)$st moments, we have*

$$⨍_0^T E(t)^p\, dt \geq C_2 ⨍_0^T \left(t^{-1/3}\right)^p dt \quad \textit{for } T \geq C_3\, L(0)^3. \tag{2.10}$$

THEOREM 2.2 (interface-reaction–controlled growth). *Let $n \geq 2$ be an integer and let p be real with $1 < p < 2$. Then there exist positive constants D_2 and D_3, depending on p, n, and nothing else, such that whenever ν is a measure-valued solution of the transport equation* (1.8) *and ν_0 has finite nth and $(n+1)$st moments, we have*

$$⨍_0^T E(t)^p\, dt \geq D_2 ⨍_0^T \left(t^{-1/2}\right)^p dt \quad \textit{for } T \geq D_3\, L(0)^2. \tag{2.11}$$

3. Discrete systems I: Volume-diffusion–controlled growth. In this section, our aim is to prove the coarsening estimate (2.4) for any collection of finitely many spherical particles in ^{n} that undergoes coarsening controlled by volume diffusion with or without kinetic drag. The following growth law holds for each particle:

$$\dot{R}_i = \frac{1}{R_i + \beta}\left(\theta - \frac{1}{R_i}\right), \quad (1 \le i \le N(t)), \tag{3.1}$$

where R_i is the radius of the ith particle, $N(t)$ is the number of surviving particles at time t, θ is the mean field, and the dot denotes the time derivative.

By the conservation of total mass,

$$0 = \frac{d}{dt}\sum R_i^n = n\sum R_i^{n-1}\dot{R}_i = n\sum \frac{R_i^{n-1}}{R_i + \beta}\left(\theta - \frac{1}{R_i}\right). \tag{3.2}$$

Here the sum goes over all surviving particles. Thus

$$\theta = \frac{\sum (R_i + \beta)^{-1} R_i^{n-2}}{\sum (R_i + \beta)^{-1} R_i^{n-1}}. \tag{3.3}$$

The right-hand side of (3.1) is smooth as long as there is no particle disappearing. The conservation of total mass guarantees that the solution for (3.1) and (3.3) cannot blow up in finite time. So the solution is smooth and unique from time $t_0 = 0$ up to t_1 when some particles disappear. Restarting from t_1 with the remaining particles, we again get a smooth solution until a next time t_2 when some other particles disappear. In this way, we can find finitely many times $\{t_i\}$ such that the solution for (3.1) and (3.3) globally exists, is unique, and is smooth in each time interval $(t_i, t_{i+1}), i = 0, 1, \dots$.

By definition (2.1),

$$E = \frac{\sum R_i^{n-1}}{\sum R_i^n}. \tag{3.4}$$

Notice that E is nonincreasing in time—we have

$$\begin{aligned} \dot{E} &= \frac{n-1}{\sum R_i^n}\sum R_i^{n-2}\dot{R}_i = \frac{n-1}{\sum R_i^n}\sum \frac{R_i^{n-2}}{R_i + \beta}\left(\theta - \frac{1}{R_i}\right) \\ &= \frac{n-1}{\sum R_i^n}\left[\frac{\left(\sum (R_i + \beta)^{-1} R_i^{n-2}\right)^2}{\sum (R_i + \beta)^{-1} R_i^{n-1}} - \sum \frac{R_i^{n-3}}{R_i + \beta}\right] \le 0 \end{aligned} \tag{3.5}$$

since, by the Cauchy–Schwarz inequality,

$$\sum \frac{R_i^{n-2}}{R_i + \beta} = \sum \left[\frac{R_i^{(n-1)/2}}{(R_i + \beta)^{1/2}} \frac{R_i^{(n-3)/2}}{(R_i + \beta)^{1/2}}\right] \le \left(\sum \frac{R_i^{n-1}}{R_i + \beta}\sum \frac{R_i^{n-3}}{R_i + \beta}\right)^{1/2}.$$

By definition (2.2),

$$L = \frac{\sum R_i^{n+1}}{\sum R_i^n}. \tag{3.6}$$

Inequality (2.3) gives us the required interpolation inequality $EL \ge 1$. Next we establish a dissipation relation that controls $\dot{L}$ in terms of $\dot{E}$. Taking the time derivative of L, we get

$$\dot{L} = \frac{n+1}{\sum R_i^n} \sum R_i^n \dot{R}_i \;=\; \frac{n+1}{\sum R_i^n} \sum \frac{R_i^n}{R_i+\beta}\left(\theta - \frac{1}{R_i}\right) \tag{3.7}$$
$$= \frac{n+1}{\sum R_i^n}\left[\frac{\sum (R_i+\beta)^{-1}R_i^n \cdot \sum (R_i+\beta)^{-1}R_i^{n-2}}{\sum (R_i+\beta)^{-1}R_i^{n-1}} - \sum \frac{R_i^{n-1}}{R_i+\beta}\right].$$

We can infer $\dot{L} \ge 0$ using again the Cauchy–Schwarz inequality, but we will not need this fact. We want to prove a dissipation inequality

$$|\dot{L}|^2 \le C_1(-\dot{E}) \tag{3.8}$$

for some constant C_1 depending only on n. Choosing $C_1(n) = (n+1)^2/(n-1)$, and plugging in the expressions (3.5) and (3.7), (3.8) becomes

$$\left[\sum \frac{R_i^n}{R_i+\beta} \sum \frac{R_i^{n-2}}{R_i+\beta} - \left(\sum \frac{R_i^{n-1}}{R_i+\beta}\right)^2\right]^2 \tag{3.9}$$
$$\le \sum R_i^n \sum \frac{R_i^{n-1}}{R_i+\beta} \cdot \left[\sum \frac{R_i^{n-1}}{R_i+\beta} \sum \frac{R_i^{n-3}}{R_i+\beta} - \left(\sum \frac{R_i^{n-2}}{R_i+\beta}\right)^2\right].$$

LEMMA 3.1. *Inequality* (3.9) *holds for any sequence of positive numbers* $\{R_i\}_{i=1}^N$.

To prove Lemma 3.1, we need the following lemma from [2].

LEMMA 3.2 (Lagrange identity).

$$\left(\sum_{i=1}^N x_i^2\right)\left(\sum_{i=1}^N y_i^2\right) - \left(\sum_{i=1}^N x_i y_i\right)^2 = \sum_{\substack{i,j=1\\ i<j}}^N (x_i y_j - x_j y_i)^2 \tag{3.10}$$

for any sequences of real numbers $\{x_i\}_{i=1}^N$ *and* $\{y_i\}_{i=1}^N$.

Proof of Lemma 3.1. The proof consists of several careful applications of the Lagrange identity and the Cauchy–Schwarz inequality. Taking $x_i = (R_i^n/(R_i+\beta))^{1/2}$ and $y_i = (R_i^{n-2}/(R_i+\beta))^{1/2}$ in (3.10), we get

$$I := \sum \frac{R_i^n}{R_i+\beta} \; \sum \frac{R_i^{n-2}}{R_i+\beta} - \left(\sum \frac{R_i^{n-1}}{R_i+\beta}\right)^2 \tag{3.11}$$
$$= \sum_{\substack{i,j=1\\ i<j}}^N \left[\left(\frac{R_i^n}{R_i+\beta}\right)^{1/2}\left(\frac{R_j^{n-2}}{R_j+\beta}\right)^{1/2} - \left(\frac{R_j^n}{R_j+\beta}\right)^{1/2}\left(\frac{R_i^{n-2}}{R_i+\beta}\right)^{1/2}\right]^2$$
$$= \sum_{\substack{i,j=1\\ i<j}}^N \frac{R_i^{n-2}R_j^{n-2}}{(R_i+\beta)(R_j+\beta)}(R_i-R_j)^2$$
$$\le \left\{\sum_{\substack{i,j=1\\ i<j}}^N \frac{R_i^{n-1}R_j^{n-1}(R_i-R_j)^2}{(R_i+\beta)(R_j+\beta)}\right\}^{1/2} \cdot \left\{\sum_{\substack{i,j=1\\ i<j}}^N \frac{R_i^{n-3}R_j^{n-3}(R_i-R_j)^2}{(R_i+\beta)(R_j+\beta)}\right\}^{1/2}.$$

Taking $x_i = R_i^{n/2}$ and $y_i = (R_i^{n-1}/(R_i+\beta))^{1/2}$ in (3.10), we get

$$\sum R_i^n \sum \frac{R_i^{n-1}}{R_i+\beta} \tag{3.12}$$

$$= \left[\sum\left(\frac{R_i^{2n-1}}{R_i+\beta}\right)^{1/2}\right]^2 + \sum_{\substack{i,j=1\\ i<j}}^N \left[R_i^{n/2}\left(\frac{R_j^{n-1}}{R_j+\beta}\right)^{1/2} - R_j^{n/2}\left(\frac{R_i^{n-1}}{R_i+\beta}\right)^{1/2}\right]^2$$
$$\geq \sum_{\substack{i,j=1\\ i<j}}^N \frac{R_i^{n-1}R_j^{n-1}}{(R_i+\beta)(R_j+\beta)}\left[R_i^{1/2}(R_i+\beta)^{1/2} - R_j^{1/2}(R_j+\beta)^{1/2}\right]^2.$$

Taking $x_i = (R_i^{n-1}/(R_i+\beta))^{1/2}$ and $y_i = (R_i^{n-3}/(R_i+\beta))^{1/2}$ in (3.10), we get

$$\sum \frac{R_i^{n-1}}{R_i+\beta}\sum\frac{R_i^{n-3}}{R_i+\beta} - \left(\sum\frac{R_i^{n-2}}{R_i+\beta}\right)^2 \tag{3.13}$$
$$= \sum_{\substack{i,j=1\\ i<j}}^N \left[\left(\frac{R_i^{n-1}}{R_i+\beta}\right)^{1/2}\left(\frac{R_j^{n-3}}{R_j+\beta}\right)^{1/2} - \left(\frac{R_j^{n-1}}{R_j+\beta}\right)^{1/2}\left(\frac{R_i^{n-3}}{R_i+\beta}\right)^{1/2}\right]^2$$
$$= \sum_{\substack{i,j=1\\ i<j}}^N \frac{R_i^{n-3}R_j^{n-3}}{(R_i+\beta)(R_j+\beta)}(R_i-R_j)^2.$$

Thus

$$II := \sum R_i^n \sum\frac{R_i^{n-1}}{R_i+\beta}\left[\sum\frac{R_i^{n-1}}{R_i+\beta}\sum\frac{R_i^{n-3}}{R_i+\beta} - \left(\sum\frac{R_i^{n-2}}{R_i+\beta}\right)^2\right] \tag{3.14}$$
$$\geq \sum_{\substack{i,j=1\\ i<j}}^N \frac{R_i^{n-1}R_j^{n-1}}{(R_i+\beta)(R_j+\beta)}\left[R_i^{1/2}(R_i+\beta)^{1/2} - R_j^{1/2}(R_j+\beta)^{1/2}\right]^2$$
$$\cdot \sum_{\substack{i,j=1\\ i<j}}^N \frac{R_i^{n-3}R_j^{n-3}}{(R_i+\beta)(R_j+\beta)}(R_i-R_j)^2.$$

Comparing (3.11) and (3.14), $I^2 \leq II$ is an immediate consequence of the inequality

$$(R_i-R_j)^2 \leq \left[R_i^{1/2}(R_i+\beta)^{1/2} - R_j^{1/2}(R_j+\beta)^{1/2}\right]^2 \quad \text{for all } i,j. \tag{3.15}$$

Inequality (3.15) holds since

$$\left[R_i^{1/2}(R_i+\beta)^{1/2} - R_j^{1/2}(R_j+\beta)^{1/2}\right]^2 - (R_i-R_j)^2$$
$$= \beta(R_i+R_j) + 2R_iR_j - 2R_i^{1/2}R_j^{1/2}(R_i+\beta)^{1/2}(R_j+\beta)^{1/2},$$

and

$$\left[\beta(R_i+R_j)+2R_iR_j\right]^2 - \left[2R_i^{1/2}R_j^{1/2}(R_i+\beta)^{1/2}(R_j+\beta)^{1/2}\right]^2$$
$$= \beta^2(R_i-R_j)^2 \geq 0. \qquad \square$$

The dissipation inequality (3.8) follows from Lemma 3.1. Applying Lemma 3 in [6], we directly get the following estimates.

THEOREM 3.3. *For any $0 \leq \lambda \leq 1$ and $0 < r < 3$ satisfying $\lambda r > 1$ and $(1-\lambda)r < 2$, there exist positive constants C_2 and C_3, depending only on λ, r, and the dimension of space n, such that for any solution $\{R_i\}$ of equations* (3.1) *and* (3.3), *we have*

$$⨍_0^T E^{\lambda r}L^{-(1-\lambda)r}\,dt \geq C_2 ⨍_0^T \left(t^{-1/3}\right)^r dt \quad for \quad T \geq C_3\, L(0)^3, \tag{3.16}$$

where E and L are defined in terms of (2.1) *and* (2.2), *respectively.*

Proof. Lemma 3.1 guarantees that the dissipation relation (3.8) holds. Together with the interpolation inequality (2.3), we have

$$EL \geq 1 \quad \text{and} \quad |\dot{L}|^2 \leq C_1(-\dot{E}).$$

Lemma 3 in [6] leads directly to (3.3). ∎

Taking $\lambda = 1$ and $r = p$ for $1 < p < 3$ in Theorem 3.3 yields (2.4).

4. Discrete systems II: Interface-reaction–controlled growth. Our aim in this section is to prove the coarsening estimate (2.5) for any collection of finitely many spherical particles in $\mathbb{R}^n$ that undergoes coarsening controlled by interface reactions. Each particle obeys the growth law

$$\dot{R}_i = \theta - \frac{1}{R_i}, \quad (1 \leq i \leq N(t)), \tag{4.1}$$

where R_i is the radius of the ith particle and θ is the mean field.

By the conservation of total mass,

$$0 = \frac{d}{dt}\sum R_i^n = n\sum R_i^{n-1}\dot{R}_i = n\sum R_i^{n-1}\left(\theta - \frac{1}{R_i}\right), \tag{4.2}$$

and thus

$$\theta = \frac{\sum R_i^{n-2}}{\sum R_i^{n-1}}. \tag{4.3}$$

Solutions of the system (4.1) and (4.3) have the same global existence and piecewise smooth properties as that of (3.1) and (3.3). Taking the time derivative of $E = \sum R_i^{n-1}/\sum R_i^n$, we have

$$\begin{aligned} \dot{E} &= \frac{n-1}{\sum R_i^n}\sum R_i^{n-2}\dot{R}_i = \frac{n-1}{\sum R_i^n}\sum R_i^{n-2}\left(\theta - \frac{1}{R_i}\right) \\ &= \frac{n-1}{\sum R_i^n}\left[\frac{\left(\sum R_i^{n-2}\right)^2}{\sum R_i^{n-1}} - \sum R_i^{n-3}\right] \leq 0, \end{aligned} \tag{4.4}$$

since

$$\sum R_i^{n-2} = \sum\left[R_i^{(n-1)/2}R_i^{(n-3)/2}\right] \leq \left(\sum R_i^{n-1}\right)^{1/2}\left(\sum R_i^{n-3}\right)^{1/2}. \tag{4.5}$$

Taking the time derivative of $L = \sum R_i^{n+1}/\sum R_i^n$, we have

$$\dot{L} = \frac{n+1}{\sum R_i^n}\sum R_i^n\dot{R}_i = \frac{n+1}{\sum R_i^n}\left[\frac{\sum R_i^n\sum R_i^{n-2}}{\sum R_i^{n-1}} - \sum R_i^{n-1}\right]. \tag{4.6}$$

Again, by the Cauchy–Schwarz inequality, we can infer $\dot{L} \geq 0$.

As described in section 2, we will need a dissipation inequality that relates $\dot{L}$ and $\dot{E}$. We claim that

$$|\dot{L}|^2 \leq D_1 L(-\dot{E}) \tag{4.7}$$

for some positive constant D_1 depending only on n. Choosing $D_1(n) = (n+1)^2/(n-1)$, and plugging in the expressions (4.4) and (4.6), inequality (4.7) becomes

$$\left[\sum R_i^n \sum R_i^{n-2} - \left(\sum R_i^{n-1}\right)^2\right]^2 \\ \leq \sum R_i^{n-1} \sum R_i^{n+1} \cdot \left[\sum R_i^{n-1} \sum R_i^{n-3} - \left(\sum R_i^{n-2}\right)^2\right]. \tag{4.8}$$

LEMMA 4.1. *Inequality* (4.8) *holds for any sequence of positive numbers* $\{R_i\}_{i=1}^N$.

Proof. Similar to the proof of Lemma 3.1, we will apply the Lagrange identity (3.10) and the Cauchy–Schwarz inequality. Taking $x_i = R_i^{n/2}$ and $y_i = R_i^{(n-2)/2}$ in (3.10), we have

$$\begin{aligned} I &:= \sum R_i^n \sum R_i^{n-2} - \left(\sum R_i^{n-1}\right)^2 \\ &= \sum_{\substack{i,j=1 \\ i<j}}^N \left[R_i^{n/2} R_j^{(n-2)/2} - R_j^{n/2} R_i^{(n-2)/2}\right]^2 \\ &= \sum_{\substack{i,j=1 \\ i<j}}^N R_i^{n-2} R_j^{n-2} (R_i - R_j)^2 \\ &\leq \left[\sum_{\substack{i,j=1 \\ i<j}}^N R_i^{n-1} R_j^{n-1} (R_i - R_j)^2\right]^{1/2} \left[\sum_{\substack{i,j=1 \\ i<j}}^N R_i^{n-3} R_j^{n-3} (R_i - R_j)^2\right]^{1/2}. \end{aligned} \tag{4.9}$$

Taking $x_i = R_i^{(n-1)/2}$ and $y_i = R_i^{(n+1)/2}$ in (3.10), we have

$$\begin{aligned} \sum R_i^{n-1} \sum R_i^{n+1} &= \left(\sum R_i^n\right)^2 + \sum_{\substack{i,j=1 \\ i<j}}^N \left(R_i^{(n-1)/2} R_j^{(n+1)/2} - R_j^{(n-1)/2} R_i^{(n+1)/2}\right)^2 \\ &= \left(\sum R_i^n\right)^2 + \sum_{\substack{i,j=1 \\ i<j}}^N R_i^{n-1} R_j^{n-1} (R_j - R_i)^2. \end{aligned} \tag{4.10}$$

Taking $x_i = R_i^{(n-1)/2}$ and $y_i = R_i^{(n-3)/2}$ in (3.10), we have

$$\begin{aligned} \sum R_i^{n-1} \sum R_i^{n-3} - \left(\sum R_i^{n-2}\right)^2 &= \sum_{\substack{i,j=1 \\ i<j}}^N \left[R_i^{(n-1)/2} R_j^{(n-3)/2} - R_j^{(n-1)/2} R_i^{(n-3)/2}\right]^2 \\ &= \sum_{\substack{i,j=1 \\ i<j}}^N R_i^{n-3} R_j^{n-3} (R_i - R_j)^2. \end{aligned} \tag{4.11}$$

Thus

$$\begin{aligned} II &:= \sum R_i^{n-1} \sum R_i^{n+1} \left[\sum R_i^{n-1} \sum R_i^{n-3} - \left(\sum R_i^{n-2}\right)^2\right] \\ &\geq \sum_{\substack{i,j=1 \\ i<j}}^N R_i^{n-1} R_j^{n-1} (R_j - R_i)^2 \sum_{\substack{i,j=1 \\ i<j}}^N R_i^{n-3} R_j^{n-3} (R_j - R_i)^2 \\ &\geq I^2. \qquad \square \end{aligned} \tag{4.12}$$

At this point we have established the desired interpolation and dissipation inequalities. The third step toward our coarsening estimates is an ODE lemma.

LEMMA 4.2 (ODE lemma). *Let $E(t)$ and $L(t)$ be two continuous and piecewise smooth positive functions. Assume that for some T_1, $0 \le T_1 \le \infty$, $E(t)$ satisfies*

$$\dot{E} < 0 \text{ a.e. on } (0, T_1), \quad \dot{E} = 0 \text{ on } (T_1, \infty). \tag{4.13}$$

If $E(t)$ and $L(t)$ satisfy

$$EL \ge 1 \quad \text{and} \quad |\dot{L}|^2 \le D_1 L(-\dot{E}), \tag{4.14}$$

then for any $0 \le \lambda \le 1$ and $r > 0$ satisfying

$$r < 3, \quad \lambda r > 1 \quad \text{and} \quad (1-\lambda) r < 2, \tag{4.15}$$

we have

$$\fint_0^T E(t)^{\lambda r} L(t)^{1-(1-\lambda)r}\, dt \ge D_2 \fint_0^T (t^{-1/2})^{r-1}\, dt \ \text{ for } T \ge D_3 L(0)^2, \tag{4.16}$$

where D_2 and D_3 are positive constants depending only on λ, r, and D_1.

We remark that this lemma is key for obtaining bounds on coarsening rates for the $t^{1/2}$ growth law. We will extend the ideas in the proof of Lemma 3 in [6] to establish this result. A special case of Lemma 4.2 is to take $r = p + 1$ and $\lambda = p/(p + 1)$ for $1 < p < 2$. In this case, we obtain (2.5)

$$\fint_0^T E(t)^p\, dt \ge D_2 \fint_0^T (t^{-1/2})^p\, dt \ \text{ for } T \ge D_3 L(0)^2, \tag{4.17}$$

where D_2 and D_3 are positive constants depending only on p and D_1.

Proof of Lemma 4.2. (1) If $T_1 = 0$, then $\dot{E} = 0$ on $(0, \infty)$. By assumption (4.14), we get $\dot{L} = 0$ on $(0, \infty)$. Hence $E(t) = E(0)$ and $L(t) = L(0)$ for all $t \in (0, \infty)$. By (4.15), $\lambda r > 1$ and $0 \le \lambda \le 1$ imply that $r > 1/\lambda \ge 1$. Hence we have $1 < r < 3$. Then

$$\begin{aligned}\fint_0^T E(t)^{\lambda r} L(t)^{1-(1-\lambda)r}\, dt &= E(0)^{\lambda r} L(0)^{1-(1-\lambda)r} \\ &\ge L(0)^{1-r} \\ &\ge \frac{2}{3-r} T^{(1-r)/2} \quad \text{ if } T \ge \left(\frac{2}{3-r}\right)^{2/(r-1)} L(0)^2 \\ &= D_2' \fint_0^T (t^{-1/2})^{r-1}\, dt \quad \text{ if } T \ge D_3' L(0)^2,\end{aligned} \tag{4.18}$$

where

$$D_2' = 1 \quad \text{and} \quad D_3' = \left(\frac{2}{3-r}\right)^{2/(r-1)}.$$

(2) Now we consider the case when $T_1 > 0$. $\dot{E}(t) < 0$ on $(0, T_1)$ implies that E is a strictly decreasing function of t on $(0, T_1)$. Hence $E(t)$ is invertible on $(0, T_1)$ and we regard $t \in (0, T_1)$ as a function of ε, with ε denoting the independent variable to distinguish it from $E = E(t)$ and avoid confusion. Note that ε ranges from $E(0)$ to

$E(\infty) := \lim_{t\to\infty} E(t)$, since $\dot{E}(t) = 0$ for $t \in (T_1, \infty)$ implies that $E(t) = E(T_1)$ for any $t > T_1$. Consequently, for $t \in (0, T_1)$, $L(t)$ can be viewed as a function of ε. Thus

$$\frac{dL}{dt} = \frac{dL}{d\varepsilon}\frac{dE}{dt} \quad \text{for } t \in (0, T_1) \tag{4.19}$$

and $|\dot{L}|^2 \le D_1 L(-\dot{E})$ implies that

$$\left|\frac{dL}{d\varepsilon}\right|^2 (-\dot{E}) \le D_1 L(\varepsilon). \tag{4.20}$$

Multiplying both sides by a positive function $f(\varepsilon)$ and integrating from 0 to T, we have

$$\int_0^T f(E(t))L(t)\,dt \ge \frac{1}{D_1}\int_{E_T}^{E_0} f(\varepsilon)\left(\frac{dL}{d\varepsilon}\right)^2 d\varepsilon \tag{4.21}$$

if $T < T_1$, and

$$\int_0^T f(E(t))L(t)\,dt \ge \int_0^{T_1} f(E(t))L(t)\,dt \ge \frac{1}{D_1}\int_{E_T}^{E_0} f(\varepsilon)\left(\frac{dL}{d\varepsilon}\right)^2 d\varepsilon \tag{4.22}$$

if $T \ge T_1$, where $E_0 = E(0)$ and $E_T = E(T)$. Taking $f(\varepsilon) = \varepsilon^{\lambda r} L(\varepsilon)^{-(1-\lambda)r}$, we get

$$\int_0^T E(t)^{\lambda r} L(t)^{1-(1-\lambda)r}\,dt \ge \frac{1}{D_1}\int_{E_T}^{E_0} \varepsilon^{\lambda r} L(\varepsilon)^{-(1-\lambda)r}\left(\frac{dL}{d\varepsilon}\right)^2 d\varepsilon. \tag{4.23}$$

We will change variables so that the right-hand side becomes an integral of a square of some gradient. Consider

$$\hat{\varepsilon} = \frac{1}{\lambda r - 1}\varepsilon^{-(\lambda r - 1)}, \quad \hat{L} = \frac{1}{1 - r(1-\lambda)/2} L^{1-r(1-\lambda)/2}. \tag{4.24}$$

Our requirements $\lambda r > 1$ and $(1-\lambda)r < 2$ guarantee that $\hat{\varepsilon} > 0, \hat{L} > 0$ and $\hat{\varepsilon} \to \infty, \hat{L} \to \infty$ as $\varepsilon \to 0$ and $L \to \infty$, respectively. Also, we have

$$\frac{d\hat{\varepsilon}}{d\varepsilon} = -\varepsilon^{-\lambda r} \tag{4.25}$$

and

$$\left(\frac{d\hat{L}}{d\hat{\varepsilon}}\right)^2 d\hat{\varepsilon} = \left(\frac{d\hat{L}}{dL}\right)^2 \left(\frac{dL}{d\varepsilon}\right)^2 \left(\frac{d\varepsilon}{d\hat{\varepsilon}}\right)^2 \left(\frac{d\hat{\varepsilon}}{d\varepsilon}\right) d\varepsilon = -\varepsilon^{\lambda r} L^{-(1-\lambda)r}\left(\frac{dL}{d\varepsilon}\right)^2 d\varepsilon. \tag{4.26}$$

Hence

$$\int_{E_T}^{E_0} \varepsilon^{\lambda r} L^{-(1-\lambda)r}\left(\frac{dL}{d\varepsilon}\right)^2 d\varepsilon = \int_{\hat{E}_0}^{\hat{E}_T} \left(\frac{d\hat{L}}{d\hat{\varepsilon}}\right)^2 d\hat{\varepsilon}. \tag{4.27}$$

The right-hand side is bounded below by its minimum over all functions $\hat{L}(\hat{\varepsilon})$ with the same endpoint values

$$\hat{L}_0 := \hat{L}(\hat{E}_0) = \frac{1}{1 - r(1-\lambda)/2} L(0)^{1-r(1-\lambda)/2}$$

and

$$\hat{L}_T := \hat{L}(\hat{E}_T) = \frac{1}{1-r(1-\lambda)/2} L(t)^{1-r(1-\lambda)/2},$$

and the minimum is achieved when $\hat{L}$ is a linear function of $\hat{\varepsilon}$. Thus

$$\int_0^T E^{\lambda r}(t) L^{1-(1-\lambda)r}(t)\, dt \geq \frac{1}{D_1} \frac{(\hat{L}_T - \hat{L}_0)^2}{\hat{E}_T - \hat{E}_0}. \tag{4.28}$$

(2a) If $L(T) \geq 2L(0)$, then

$$\hat{L}_0 \leq 2^{r(1-\lambda)/2-1} \hat{L}_T < \hat{L}_T.$$

Hence

$$\hat{L}_T - \hat{L}_0 \geq \left(1 - 2^{r(1-\lambda)/2-1}\right) \hat{L}_T$$

and, consequently,

$$\begin{aligned} \int_0^T E^{\lambda r}(t) L^{1-(1-\lambda)r}(t)\, dt &\geq \frac{1}{D_1} \frac{(\hat{L}_T - \hat{L}_0)^2}{\hat{E}_T - \hat{E}_0} \geq \frac{\left(1 - 2^{r(1-\lambda)/2-1}\right)^2}{D_1} \frac{\hat{L}_T^2}{\hat{E}_T} \\ &\geq \frac{\left(1 - 2^{r(1-\lambda)/2-1}\right)^2}{D_1} \frac{(\lambda r - 1)}{(1-r(1-\lambda)/2)^2}\, E^{\lambda r - 1} L^{2-(1-\lambda)r} \\ &= \hat{D}_2\, (EL)^{((2\lambda-1)r+1)/(r-1)} \cdot \left(E^{\lambda r} L^{1-(1-\lambda)r}\right)^{-(r-3)/(1-r)}, \end{aligned} \tag{4.29}$$

where

$$\hat{D}_2 := \frac{(\lambda r - 1)}{D_1} \left(\frac{1 - 2^{r(1-\lambda)/2-1}}{1 - r(1-\lambda)/2} \right)^2.$$

Since $1 < r < 3$ and $\lambda r > 1$,

$$(2\lambda - 1)r + 1 = 2\lambda r + 1 - r > 3 - r > 0.$$

Thus

$$\frac{(2\lambda-1)r+1}{r-1} > 0 \quad \text{and} \quad \frac{r-3}{1-r} > 0.$$

So $EL \geq 1$ implies $(EL)^{((2\lambda-1)r+1)/(r-1)} \geq 1$ and hence

$$\int_0^T E^{\lambda r}(t) L^{1-(1-\lambda)r}(t)\, dt \geq \hat{D}_2 \left(E^{\lambda r} L^{1-(1-\lambda)r}\right)^{-(r-3)/(1-r)} \quad \text{if } L_T \geq 2L_0. \tag{4.30}$$

Define

$$h(T) := \int_0^T E^{\lambda r}(t) L^{1-(1-\lambda)r}(t)\, dt. \tag{4.31}$$

Then $h'(T) = E^{\lambda r}(T) L^{1-(1-\lambda)r}(T)$ and (4.30) can be rewritten as

$$h(T) \geq \hat{D}_2 (h'(T))^{-(r-3)/(1-r)} \quad \text{if } L_T \geq 2L_0 \tag{4.32}$$

or, equivalently,

$$h'(T)\big(h(T)\big)^{(r-1)/(3-r)} \geq \hat{D}_2^{(r-1)/(3-r)} \quad \text{if } L_T \geq 2L_0. \tag{4.33}$$

(2b) If $L(T) < 2L(0)$, then by $EL \geq 1$,

$$E_T \geq L_T^{-1} \geq \frac{1}{2}L_0^{-1},$$

$$E_T^{\lambda r} L_T^{1-(1-\lambda)r} = \Big(E_T L_T\Big)^{\lambda r} L_T^{1-r} \geq L_T^{1-r} \geq L_0^{1-r}\, 2^{1-r}.$$

Thus

$$h'(T) \geq L_0^{1-r}\, 2^{1-r} \quad \text{if } \ L(T) \leq 2L(0). \tag{4.34}$$

Combining (4.33) and (4.34), we have

$$h'(T)\Big[h(T) + L_0^{3-r}\Big]^{(r-1)/(3-r)} \geq \min\{2^{1-r}, \hat{D}_2^{(r-1)/(3-r)}\} =: m \quad \text{for all } T.$$

Thus

$$\frac{d}{dt}\Big[h(t) + L_0^{3-r}\Big]^{2/(3-r)} = \frac{2}{3-r} h'(t)\Big[h(t) + L_0^{3-r}\Big]^{(r-1)/(3-r)} \geq \frac{2m}{3-r}. \tag{4.35}$$

By integration over time from 0 to T, we get

$$\begin{aligned} h(T) &\geq \left[\frac{2m}{3-r}T + L_0^2\right]^{(3-r)/2} - L_0^{3-r} \\ &\geq \left(\frac{2m}{3-r}\right)^{(3-r)/2} T^{(3-r)/2} - L_0^{3-r} \\ &\geq \frac{1}{2}\left(\frac{2m}{3-r}\right)^{(3-r)/2} T^{(3-r)/2} \quad \text{if } \ T > 2^{2/(3-r)}\frac{(3-r)}{2m}L_0^2. \end{aligned} \tag{4.36}$$

Equivalently,

$$⨍_0^T E(t)^{\lambda r} L(t)^{1-(1-\lambda)r}\, dt \geq D_2''⨍_0^T (t^{-1/2})^{r-1}\, dt \quad \text{for } T > D_3'' L_0^2, \tag{4.37}$$

where

$$D_2'' = \frac{3-r}{4}\left(\frac{2m}{3-r}\right)^{(3-r)/2} \quad \text{and} \quad D_3'' = 2^{2/(3-r)}\frac{(3-r)}{2m}.$$

(3) Combining (1) and (2), we conclude that

$$⨍_0^T E(t)^{\lambda r} L(t)^{1-(1-\lambda)r}\, dt \geq D_2⨍_0^T (t^{-1/2})^{r-1}\, dt \quad \text{for } T > D_3 L_0^2, \tag{4.38}$$

where

$$D_2 = \min\{D_2', D_2''\} \quad \text{and} \quad D_3 = \max\{D_3', D_3''\}. \qquad \square$$

We claim the following estimate for the collection of particles that undergoes coarsening determined by (4.1).

THEOREM 4.3. *For any* $0 \le \lambda \le 1$ *and* $0 < r < 3$ *satisfying* $\lambda r > 1$ *and* $(1-\lambda)r < 2$, *there exist positive constants* D_2 *and* D_3, *depending only on* λ, r, *and the dimension of space* n, *such that for any solution* $\{R_i\}$ *of equations* (4.1) *and* (4.3), *we have*

$$(4.39) \qquad \fint_0^T E(t)^{\lambda r} L(t)^{1-(1-\lambda)r}\, dt \ge D_2 \fint_0^T (t^{-1/2})^{r-1}\, dt \quad \text{for } T \ge D_3 L(0)^2,$$

where E *and* L *are defined in terms of* (2.1) *and* (2.2), *respectively.*

Proof. As we discussed at the beginning of this section, solutions $\{R_i\}$ of equations (4.1) and (4.3) are continuous and piecewise smooth. Hence E and L defined by (2.1) and (2.2) are continuous and piecewise smooth. Furthermore, by (4.4), $\dot{E} \le 0$ and $\dot{E} = 0$ if and only if all R_i are equal. Notice that if all R_i are equal, then the system (4.1) and (4.3) reaches an equilibrium point and the solution stops coarsening. Consequently, if $\dot{E} = 0$ at some time t_1, then $\dot{E}(t) = 0$ for all $t \ge t_1$. Hence, $\dot{E}$ satisfies the condition (4.13) of Lemma 4.2.

On the other hand, the interpolation inequality (2.3) and the dissipation relation (4.7) say

$$EL \ge 1 \quad \text{and} \quad |\dot{L}|^2 \le D_1 L(-\dot{E}).$$

The theorem is then an immediate consequence of Lemma 4.2. $\square$

5. Coarsening rates for particle systems with general size distributions. Now it is time to consider our mean-field models with more general size distributions. Definitions (2.1) and (2.2) imply that, in the more general case, E and L should be defined in terms of the $(n-1)$st, nth, and $(n+1)$st moments of the size distributions. Thus it is necessary to require the initial size distributions to be in $\mathcal{P}_{n+1}$, the set of Borel probability measures on $[0, \infty)$ with finite $(n+1)$st moments. By Hölder's inequality, it is immediate to see that $\mathcal{P}_{n+1}$ is a subset of $\mathcal{P}_n$.

In [13], Niethammer and Pego proved well-posedness and compactness results for a family of mean-field models. All of our models under consideration are included in that work except for the 2D volume-diffusion–controlled growth model with $\beta = 0$. Their results guarantee the existence and uniqueness of measure-valued solutions of equation (1.2) or (1.8). A measure-valued solution is a weak-star continuous map $t \mapsto \nu_t$ taking $[0, \infty) \to \mathcal{P}_n$ that is a solution in the sense of distributions on $(0, \infty) \times (0, \infty)$; i.e., for all $\phi \in C_c^\infty([0, \infty) \times (0, \infty))$ (smooth functions with compact support),

$$(5.1) \qquad \int_0^\infty \int_0^\infty \left(\partial_t \phi + \frac{1}{R+\beta}\left(\theta(t) - \frac{1}{R}\right) \partial_R \phi \right) d\nu_t\, dt + \int_0^\infty \phi(0, \cdot)\, d\nu_0 = 0$$

in the case of volume-diffusion–controlled growth (1.2), or

$$(5.2) \qquad \int_0^\infty \int_0^\infty \left(\partial_t \phi + \left(\theta(t) - \frac{1}{R}\right) \partial_R \phi \right) d\nu_t\, dt + \int_0^\infty \phi(0, \cdot)\, d\nu_0 = 0$$

in the case of interface-reaction–controlled growth (1.8).

Our main results are estimates in terms of these measure-valued solutions.

THEOREM 5.1 (volume-diffusion–controlled growth). *Let* $n \ge 2$ *be an integer and* $\beta \ge 0$, *with* $\beta > 0$ *if* $n = 2$. *For any* $0 \le \lambda \le 1$ *and* $0 < r < 3$ *satisfying* $\lambda r > 1$ *and* $(1-\lambda)r < 2$, *there exist positive constants* C_2 *and* C_3, *depending only on* λ, r, *and the* *dimension of space n, such that whenever ν is a measure-valued solution of the transport equation* (1.2) *and the initial value ν_0 has finite nth and $(n+1)$st moments, we have*

$$\fint_0^T E(t)^{\lambda r} L(t)^{-(1-\lambda)r}\, dt \geq C_2 \fint_0^T (t^{-1/3})^r\, dt \quad \text{for } T \geq C_3 L(0)^3, \tag{5.3}$$

where $E(t)$ and $L(t)$ are defined by (2.8) *and* (2.9)*, respectively, and the mean field $\theta(t)$ is defined by* (2.6)*.*

Taking $r = p$ and $\lambda = 1$ for $1 < p < 3$ in Theorem 5.1 gives Theorem 2.1.

THEOREM 5.2 (interface-reaction–controlled growth). *Let $n \geq 2$ be an integer. For any $0 \leq \lambda \leq 1$ and $0 < r < 3$ satisfying $\lambda r > 1$ and $(1-\lambda)r < 2$, there exist positive constants D_2 and D_3, depending only on λ, r, and the dimension of space n, such that whenever ν is a measure-valued solution of the transport equation* (1.8) *and the initial value ν_0 has finite nth and $(n+1)$st moments, we have*

$$\fint_0^T E(t)^{\lambda r} L(t)^{1-(1-\lambda)r}\, dt \geq D_2 \fint_0^T (t^{-1/2})^{r-1}\, dt \quad \text{for } T \geq D_3 L(0)^2, \tag{5.4}$$

where $E(t)$ and $L(t)$ are defined by (2.8) *and* (2.9)*, respectively, and the mean field $\theta(t)$ is defined by* (2.7)*.*

Taking $r = p+1$ and $\lambda = p/(p+1)$ for $1 < p < 2$ in Theorem 5.2 gives Theorem 2.2.

The remaining part of this section is devoted to proving the theorems above. To do this, we will need a change of variables as is done in [13]. In that paper, rather than directly working on distributions of particle radii R, Niethammer and Pego change the problems into equivalent ones expressed in terms of rescaled particle volumes $x(:= R^n)$ and work with a size-ranking function for particle volumes.

According to (1.1) and (1.7), the particle volume x satisfies the following growth law:

$$\dot{x} = a(x)\theta - b(x), \tag{5.5}$$

where

$$a(x) = \frac{n x^{1-1/n}}{x^{1/n} + \beta}, \quad b(x) = \frac{n x^{1-2/n}}{x^{1/n} + \beta} \quad \text{for volume-diffusion–controlled case,} \tag{5.6}$$

$$a(x) = n x^{1-1/n}, \quad b(x) = n x^{1-2/n} \quad \text{for interface-reaction–controlled case,} \tag{5.7}$$

and $\theta(t) = \int b(x)\, d\nu_t(x) / \int a(x)\, d\nu_t(x)$. Here ν is the measure-valued solution in the sense of distributions for the transport equation

$$\partial_t u + \partial_x\big((a(x)\theta - b(x))u\big) = 0. \tag{5.8}$$

The results of Niethammer and Pego are established by a further change of variables ([13]; see also [12]). For any size distribution of particles which is a probability measure μ on $[0, \infty)$, they define a *size-ranking function* $x = \hat{x}(\mu) : (0, 1] \to [0, \infty)$ by

$$x(\varphi) = \begin{cases} \sup\{y \mid \mu([y, \infty)) > \varphi\} & \text{for} \quad 0 < \varphi < 1, \\ 0 & \text{for} \quad \varphi = 1. \end{cases} \tag{5.9}$$

This is the right-continuous inverse of the tail distribution function $\varphi(x) = \mu([x, \infty))$. The map $\hat{x}$ gives a 1–1 correspondence between the set of Borel probability measures on $[0, \infty)$ and the set of right-continuous decreasing functions x on $(0, 1]$ with $x(1) = 0$.

The following space for size ranking is introduced in [13]:

$$L^1_d = \{x : (0,1] \to \quad | \; x \in L^1((0,1)),\ x(1) = 0,\ \text{and } x \text{ is decreasing} \\ \text{and right continuous on } (0,1]\}.$$

It is a closed subspace of $L^1((0,1))$. We will also perform our estimates in this space.

By statement 2.5.18(3) in [5], for any continuous function $f : (0, \infty) \to \mathbf{R}$ with compact support,

$$\int_0^1 f(x(\varphi))\, d\varphi = \int_0^\infty f(y)\, d\mu(y). \tag{5.10}$$

For any positive number $\alpha > 0$, $y \mapsto y^\alpha$ can be approximated by a monotonically increasing sequence of such functions, and thus by the monotone convergence theorem

$$\int_0^1 x(\varphi)^\alpha\, d\varphi = \int_0^\infty y^\alpha\, d\mu(y), \tag{5.11}$$

where both sides may be infinite. Hence $\mu \in \mathcal{P}_\alpha$ (Borel probability measures with finite αth moment) if and only if x is right-continuous decreasing on $(0, 1]$ with $x(1) = 0$ and $\int_0^1 x(\varphi)^\alpha\, d\varphi < \infty$.

The growth law (5.5) can be rewritten as an integral equation,

$$x(t,\varphi) = x(0,\varphi) + \int_0^t (a(x(s,\varphi))\theta(s) - b(x(s,\varphi)))\, ds \tag{5.12}$$

with

$$\theta(t) = \int_0^{\bar{\varphi}(t)} b(x(t,\varphi))\, d\varphi \Big/ \int_0^1 a(x(t,\varphi))\, d\varphi \quad \text{for a.e. } t > 0, \tag{5.13}$$

where $\bar{\varphi}(t) := \sup\{\varphi | x(t,\varphi) > 0\}$.

Theorem 2.3 of [13] established the existence and uniqueness of the initial value problem for (5.12) and (5.13) under some assumptions ((H1)–(H5) in [13]) which our problems satisfy except for the 2D volume-diffusion–controlled growth model with $\beta = 0$. This theorem claims that for any $x_0 \in L^1_d$, there exists a unique function $x \in C([0, \infty), L^1_d)$ such that (5.12) and (5.13) hold with $x(0, \varphi) = x_0(\varphi)$. This is equivalent to the existence and uniqueness (Theorem 2.1 of [13]) of a weak-star continuous solution $\nu : [0, \infty) \to \mathcal{P}_1$ for the transport equation (5.8) in the sense of distributions on $(0, \infty) \times (0, \infty)$ with initial value $\nu_0 = \hat{x}^{-1}(x_0)$.

Proposition 6.1 of [13] established an L^1 compactness result for (5.12) and (5.13); namely, given $T \in (0, \infty)$, for a compact sequence of initial values $\{x_{0k}\} \subset L^1_d$, the corresponding sequence of solutions x_k is compact in $C([0, T], L^1_d)$ and any limit x is again a solution of (5.12) and (5.13).

Based on this result, in the appendix we prove an L^p compactness result for (5.12) and (5.13) for any $1 < p < \infty$; namely, given $T \in (0, \infty)$, for a sequence of initial values $\{x_{0k}\} \subset L^1_d \cap L^p((0, 1))$ which is compact in $L^p((0, 1))$, the corresponding sequence of solutions x_k is compact in $C([0, T], L^p((0, 1)))$ and any limit x is again a solution of (5.12) and (5.13).

Given $x_0 \in L^1_d \cap L^{(n+1)/n}((0,1))$, for any positive integer N, we divide the interval $(0,1)$ uniformly into N subintervals and define a function $x_{0N}(\varphi)$ by

$$x_{0N}(\varphi) = N \int_{(i-1)/N}^{i/N} x_0(\psi)\, d\psi \ (=: x_{0N}^i), \qquad \frac{i-1}{N} \le \varphi < \frac{i}{N},\ (i = 1, \ldots, N). \tag{5.14}$$

Then $x_{0N} \in L^1_d \cap L^{(n+1)/n}((0,1))$ is piecewise constant, and $x_{0N} \to x_0$ in $L^{(n+1)/n}((0,1))$ as $N \to \infty$.

By the above compactness and uniqueness results, the solutions $\{x_N\}$ for (5.12) and (5.13) with initial values $\{x_{0N}\}$ converge in the space $C([0,T], L^{(n+1)/n}((0,1)))$ to the solution x for (5.12) and (5.13) with initial value x_0.

For any N, $\{x_{0N}^i\}_{i=1}^N$ gives a discrete collection of particles and the corresponding collection of radii $\{R_i := (x_{0N}^i)^{1/n}\}$ undergoes coarsening determined by (1.1) or (1.7). Hence the estimates (3.16) and (4.39) claimed in Theorems 3.3 and 4.3 hold for

$$E_N(t) = \frac{\sum R_i(t)^{n-1}}{\sum R_i(t)^n} = \int_0^1 x_N(t,\varphi)^{(n-1)/n}\, d\varphi \Big/ \int_0^1 x_N(t,\varphi)\, d\varphi$$

and

$$L_N(t) = \frac{\sum R_i(t)^{n+1}}{\sum R_i(t)^n} = \int_0^1 x_N(t,\varphi)^{(n+1)/n}\, d\varphi \Big/ \int_0^1 x_N(t,\varphi)\, d\varphi.$$

We will establish the convergence results for $E_N(t)$ and $L_N(t)$ in Lemma 5.4. To do this, let us first prove a general convergence result for L^p functions.

LEMMA 5.3. *For nonnegative functions $f_k, f \in L^p(\Omega)$ $(k = 1, 2, \ldots)$ with $1 < p < \infty$ and Ω a bounded open subset of* ^{n}, *if*

$$\int_\Omega |f_k(y)^p - f(y)^p|\, dy \to 0 \qquad \text{as } k \to \infty, \tag{5.15}$$

then

$$\int_\Omega |f_k(y) - f(y)|^p\, dy \to 0 \qquad \text{as } k \to \infty. \tag{5.16}$$

Proof. The convergence (5.15) implies that $\{f_k\}$ is bounded in $L^p(\Omega)$. Notice that $L^p(\Omega)$ is a reflexive Banach space since $1 < p < \infty$. Thus there exist a subsequence $\{f_{k_j}\}$ and $w \in L^p(\Omega)$ such that f_{k_j} converges weakly to w: $f_{k_j} \rightharpoonup w$ as $j \to \infty$. Hence

$$||w||_{L^p(\Omega)} \le \liminf_{j\to\infty} ||f_{k_j}||_{L^p(\Omega)} = ||f||_{L^p(\Omega)}. \tag{5.17}$$

By $f_{k_j}^p \to f^p$ in $L^1(\Omega)$, there exists a further subsequence, denoted the same, such that $f_{k_j}(y) \to f(y)$ for a.e. $y \in \Omega$. Hence, by Fatou's lemma and Hölder's inequality,

$$\int_\Omega f^p = \int_\Omega \liminf_{j\to\infty} f^{p-1} f_{k_j} \le \liminf_{j\to\infty} \int_\Omega f^{p-1} f_{k_j} = \int_\Omega f^{p-1} w \le \left(\int_\Omega f^p\right)^{1-\frac{1}{p}} \left(\int_\Omega w^p\right)^{1/p}.$$

Thus

$$||f||_{L^p(\Omega)} \le ||w||_{L^p(\Omega)}. \tag{5.18}$$

Comparing inequalities (5.17) and (5.18), we get $||f||_{L^p(\Omega)} = ||w||_{L^p(\Omega)}$. Thus

$$f_{k_j} \rightharpoonup w \quad \text{in} \quad L^p(\Omega), \tag{5.19}$$

$$||f_{k_j}||_{L^p(\Omega)} \to ||w||_{L^p(\Omega)}. \tag{5.20}$$

Thus (see, e.g., [4])

$$f_{k_j} \to w \quad \text{in} \quad L^p(\Omega), \tag{5.21}$$

and there exists a further subsequence of f_{k_j} that converges a.e. to w. Since $f_{k_j} \to f$ a.e. in Ω, we have $w = f$ and hence

$$f_{k_j} \to f \quad \text{in} \quad L^p(\Omega). \tag{5.22}$$

The above argument works for every weakly convergent subsequence and hence the whole sequence f_k converges strongly to f in $L^p(\Omega)$. ∎

LEMMA 5.4. *For any $t > 0$, we have*

$$E_N(t) \to E(t) := \int_0^1 x(t,\varphi)^{(n-1)/n}\, d\varphi \Big/ \int_0^1 x(t,\varphi)\, d\varphi \quad \textit{as} \;\; N \to \infty, \tag{5.23}$$

$$L_N(t) \to L(t) := \int_0^1 x(t,\varphi)^{(n+1)/n}\, d\varphi \Big/ \int_0^1 x(t,\varphi)\, d\varphi \quad \textit{as} \;\; N \to \infty. \tag{5.24}$$

Proof. Fix $t > 0$. By the conservation of total mass and the convergence of initial value $x_{0N} \to x_0$ in $L^1((0,1))$,

$$\int_0^1 x_N(t,\varphi)\, d\varphi = \int_0^1 x_{0N}(\varphi)\, d\varphi \to \int_0^1 x_0(\varphi)\, d\varphi = \int_0^1 x(t,\varphi)\, d\varphi \tag{5.25}$$

as $N \to \infty$. By the compactness of $\{x_N\}$ in $C\big([0,T], L^p((0,1))\big)$ for all $T > 0$ and all $p > 1$,

$$\int_0^1 x_N(t,\varphi)^{(n+1)/n}\, d\varphi \to \int_0^1 x(t,\varphi)^{(n+1)/n}\, d\varphi \quad \text{as} \;\; N \to \infty. \tag{5.26}$$

The convergence of $L_N(t)$ to $L(t)$ is an immediate consequence of (5.25) and (5.26).

Define $f_N = x_N(t,\varphi)^{(n-1)/n}, f = x(t,\varphi)^{(n-1)/n}$, and $p = n/(n-1)$. Equation (5.25) implies that $f_N^p \to f^p$ as $N \to \infty$. Thus Lemma 5.3 implies $f_N \to f$ in $L^p((0,1))$ and consequently $f_N \to f$ in $L^1((0,1))$. Hence

$$\int_0^1 x_N(t,\varphi)^{(n-1)/n}\, d\varphi \to \int_0^1 x(t,\varphi)^{(n-1)/n}\, d\varphi. \tag{5.27}$$

The convergence of $E_N(t)$ to $E(t)$ is an immediate consequence of (5.25) and (5.27). ∎

To enable us to take limit in the estimates (3.16) and (4.39) claimed in Theorems 3.3 and 4.3, we will prove the following boundedness lemma for $E_N(t)$ and $L_N(t)$ and then apply Lebesgue's dominated convergence theorem.

LEMMA 5.5. *Given $T > 0$, there exist positive constants M_1, m_2, and M_2, depending only on n and T, such that*

$$0 < E_N(t) \le M_1, \quad m_2 < L_N(t) < M_2$$

uniformly in N and $0 \le t \le T$, with M_1, m_2, and M_2 positive constants depending only on n and T.

Proof. By (5.25), there exist positive constants $\hat{m}_1$ and $\hat{M}_1$ such that for all N and all $t \ge 0$,

$$\hat{m}_1 \le \int_0^1 x_N(t,\varphi)\,d\varphi \le \hat{M}_1, \quad \hat{m}_1 \le \int_0^1 x(t,\varphi)\,d\varphi \le \hat{M}_1. \tag{5.28}$$

Then by Hölder's inequality,

$$\int_0^1 x_N(t,\varphi)^{(n-1)/n}\,d\varphi \le \left(\int_0^1 x_N(t,\varphi)\,d\varphi\right)^{(n-1)/n} \le \hat{M}_1^{(n-1)/n}. \tag{5.29}$$

Hence

$$E_N(t) = \int_0^1 x_N(t,\varphi)^{(n-1)/n}\,d\varphi \Big/ \int_0^1 x_N(t,\varphi)\,d\varphi \le \hat{M}_1^{(n-1)/n}/\hat{m}_1 =: M_1. \tag{5.30}$$

By Hölder's inequality,

$$\hat{m}_1 \le \int_0^1 x_N(t,\varphi)\,d\varphi \le \left\{\int_0^1 x_N(t,\varphi)^{(n+1)/n}\,d\varphi\right\}^{n/(n+1)}. \tag{5.31}$$

Thus

$$L_N(t) = \int_0^1 x_N(t,\varphi)^{(n+1)/n}\,d\varphi \Big/ \int_0^1 x_N(t,\varphi)\,d\varphi \ge \hat{m}_1^{(n+1)/n}/\hat{M}_1 =: m_2. \tag{5.32}$$

In the appendix, we will prove that there exists a positive increasing function $G(t)$ such that $\int_0^1 x_N(t,\varphi)^{(n+1)/n}\,d\varphi \le G(t) \le G(T)$. Thus, for all $0 \le t \le T$,

$$L_N(t) = \int_0^1 x_N(t,\varphi)^{(n+1)/n}\,d\varphi \Big/ \int_0^1 x_N(t,\varphi)\,d\varphi \le G(T)/\hat{m}_1 =: M_2. \tag{5.33}$$ ∎

The above boundedness results and Lebesgue's dominated convergence theorem guarantee that we can take limit as $N \to \infty$ in the estimates for coarsening rates for discrete systems (Theorems 3.3 and 4.3). This procedure gives us the estimates in Theorems 5.1 and 5.2, with E and L defined as in Lemma 5.4, for the coarsening rates for solutions of (5.12)+(5.13) with initial value $x_0 \in L_d^1 \cap L^{(n+1)/n}((0,1))$.

Our ultimate goal is to get estimates for coarsening rates for measure-valued solutions of the transport equations (1.2) and (1.8), respectively. To do this, we will establish the 1–1 correspondence between these measure-valued solutions, which are distributions of particle radii, and volume size-ranking solutions for (5.12)+(5.13). The estimates for coarsening rates for these measure-valued solutions are immediate consequences of this 1–1 correspondence and the estimates for these size-ranking solutions.

For any initial particle radius distribution $\mu(R) \in \mathcal{P}_{n+1}$, we define a particle volume distribution $\hat{\mu}(x) = (T\mu)(x)$ by requiring

$$\int_0^\infty f(x)\,d\hat{\mu}(x) = \int_0^\infty f(R^n)\,d\mu(R) \tag{5.34}$$

for all continuous functions f with compact support. Then $\hat{\mu} \in \mathcal{P}_{(n+1)/n}$.

The size-ranking function $x_0(\varphi) = \hat{x}(\hat{\mu})$ defined as in (5.9) belongs to $L^1_d \cap L^{(n+1)/n}((0,1))$. Hence the solution $x(t,\varphi)$ of problem (5.12)+(5.13) with $x(0,\cdot) = x_0(\cdot)$ belongs to L^1_d and we can get the estimates as in Theorems 5.1 and 5.2 by the procedure described above.

It is proved in [13] that the mapping (5.9) is invertible and that under the assumptions (H1)–(H5), the weak-star continuous mapping $\hat{\nu} : [0,\infty) \to \mathcal{P}_1$ related with $x(t,\varphi)$ through (5.9) is the unique measure-valued solution of the transport equation (5.8) in the sense of distributions with initial value $\hat{\mu}$. For any $t \in [0,\infty)$, we define a Borel measure ν_t by requiring

$$\int_0^\infty f(R)\, d\nu_t(R) = \int_0^\infty f(x^{1/n})\, d\hat{\nu}_t(x) \tag{5.35}$$

for all continuous functions f with compact support. Then $\nu : [0,\infty) \to \mathcal{P}_n$ is weak-star continuous and is a measure-valued solution of the transport equation (1.2) or (1.8), with initial value $\nu_0 = \mu$. Again, since we can approximate a power function $y \mapsto y^\alpha (\alpha > 0)$ by a monotonically increasing sequence of continuous functions with compact support, we get the following moment equivalence identity for ν and $\hat{\nu}$:

$$\int_0^\infty R^\alpha\, d\nu_t(R) = \int_0^\infty x^{\alpha/n}\, d\hat{\nu}_t(x) \quad \text{for any } \alpha > 0. \tag{5.36}$$

On the other hand, if we have a measure-valued solution $\nu : [0,\infty) \to \mathcal{P}_n$ for (1.2) or (1.8) with a given initial value μ, we can define $\hat{\nu} : [0,\infty) \to \mathcal{P}_1$ by (5.35) and $\hat{\nu}$ will be a measure-valued solution for (5.8) with initial value $\hat{\mu}$ defined by (5.34). The uniqueness of $\hat{\nu}$ then implies the uniqueness of ν.

The above analysis, together with the moment equivalence statements (5.11) and (5.36), gives us Theorems 5.1 and 5.2 on coarsening rates for mean-field models with general initial distributions of particle radii.

Appendix. In this appendix, we will establish a compactness result for solutions $x(t,\varphi)$ of problem (5.12)+(5.13) with $x(0,\cdot) = x_0(\cdot)$ for any $x_0 \in L^1_d \cap L^p((0,1))$ with $1 < p < \infty$ under the same assumptions (H1)–(H4) as in [13]. Note that the two models we considered fall into this category except for the case $n = 2, \beta = 0$ of the volume-diffusion–controlled growth model (and this is the reason why we do not include this case in our estimates for coarsening rates with general size distribution).

Proposition A.1. *Fix $T \in (0,\infty)$ and consider a sequence $\{x_k\}_{k=1}^\infty$ of solutions to* (5.12)+(5.13) *for $0 \le t \le T$ with initial values $x_k(0,\varphi) = x_{0k}(\varphi)$ $(\varphi \in (0,1))$. Assume that the sequence of initial data $\{x_{0k}\} \subset L^1_d \cap L^p((0,1))$ is compact in $L^p((0,1))$ for some $1 < p < \infty$ with $c_1 := \inf_k \int_0^1 x_{0k} > 0$. Then $\{x_k\}$ is compact in $C([0,T], L^p((0,1)))$ and any limit x is again a solution of* (5.12)+(5.13).

Proof. By Hölder's inequality, the assumption that $\{x_{0k}\} \subset L^1_d \cap L^p((0,1))$ is compact in $L^p((0,1))$ implies that $\{x_{0k}\}$ is compact in $L^1((0,1))$. Hence, by Proposition 6.1 in [13], $\{x_k\}$ is compact in $C([0,T], L^1_d)$ and any limit x is again a solution of (5.12)+(5.13). We will follow the strategy of the proof of Lemma 6.2 in [13] to prove that x_k is compact in $L^p((0,1))$.

It has been shown in [13] that $\theta(t)$ is uniformly bounded on [0,T]. The assumptions $(H1)$–$(H4)$ together with the boundedness of θ imply that there exists a positive constant C, depending only on T, such that

$$|a(x)\theta(t) - b(x)| \le C(1+x)$$

for all $x \geq 0$. By the generalized Arzelà–Ascoli theorem, to show that $\{x_k\}$ is compact in $C([0,T], L^p((0,1)))$, we need to prove the following three steps:

(1) uniform boundedness of $\{\int_0^1 x_k^p(t,\varphi)\, d\varphi\}$ for all $t \in [0,T]$ and all k,

(2) for fixed $t \in (0,T)$, $\{x_k(t,\cdot)\}$ is compact in $L^p((0,1))$,

(3) $\sup_k ||x_k(t_1,\cdot) - x_k(t_2,\cdot)||_{L^p((0,1))} \to 0$ as $|t_1 - t_2| \to 0$.

To show (1), define $F_\delta(t) = \int_\delta^1 x_k^p(t,\varphi)\, d\varphi$ for $\delta > 0$. Then $F_\delta < \infty$ since $x_k(t,\cdot)$ is decreasing and

$$\begin{aligned} F_\delta(t) &= \int_\delta^1 x_{0k}^p(\varphi)\, d\varphi + \int_\delta^1 \int_0^t p x_k^{p-1} \partial_s x_k(s,\varphi)\, ds\, d\varphi \\ &= \int_\delta^1 x_{0k}^p(\varphi)\, d\varphi + \int_\delta^1 \int_0^t p x_k^{p-1} (a(x_k)\theta - b(x_k))\, ds\, d\varphi \\ &\leq \int_\delta^1 x_{0k}^p(\varphi)\, d\varphi + Cp \int_\delta^1 \int_0^t x_k^{p-1}(1 + x_k)\, dt\, d\varphi. \end{aligned}$$

By Young's inequality,

$$x_k^{p-1} \leq \frac{p-1}{p} x_k^p + \frac{1}{p}. \tag{A.1}$$

Thus

$$\begin{aligned} F_\delta(t) &\leq \int_\delta^1 x_{0k}^p(\varphi)\, d\varphi + C \int_\delta^1 \int_0^t ((2p-1)x_k^p + 1)\, dt\, d\varphi \\ &\leq \int_\delta^1 x_{0k}^p(\varphi)\, d\varphi + CT + C(2p-1) \int_0^t \int_\delta^1 x_k^p\, d\varphi\, dt. \end{aligned}$$

By Gronwall's inequality,

$$F_\delta(t) \leq \int_\delta^1 x_{0k}^p(\varphi)\, d\varphi + CT + C(2p-1)e^{C(2p-1)t} \left(\int_\delta^1 x_{0k}^p(\varphi)\, d\varphi + CT \right). \tag{A.2}$$

The compactness of x_{0k} in $L^p((0,1))$ implies that there exists positive constant C_1 such that $\int_0^1 x_{0k}^p(\varphi)\, d\varphi \leq C_1$ for all k. Thus, by taking $\delta \to 0$ in (A.2) we get

$$\int_0^1 x_k^p(t,\varphi)\, d\varphi \leq C_1 + CT + C(2p-1)e^{C(2p-1)t}(C_1 + CT) =: G(t) \leq G(T). \tag{A.3}$$

Here G is an increasing function of t and does not depend on k. Hence (1) is proved.

It is shown in the proof of Lemma 6.2 in [13] that, for fixed t, there exists a pointwise convergent subsequence, still denoted as $\{x_k\}$ for simplicity. Therefore, to prove (2) we need only show that $\{x_k\}$ is equi-integrable. Since $x_k(t,\cdot)$ are decreasing, it is enough to show

$$\sup_k \int_0^\varepsilon x_k(t,\varphi)\, d\varphi \to 0 \quad \text{as } \varepsilon \to 0. \tag{A.4}$$

$$\begin{aligned} \int_0^\varepsilon x_k^p(t,\varphi)\, d\varphi &= \int_0^\varepsilon x_{0k}^p(\varphi)\, d\varphi + \int_0^\varepsilon \int_0^t p x_k^{p-1}(s,\varphi) \partial_s x_k(s,\varphi)\, ds\, d\varphi \\ &= \int_0^\varepsilon x_{0k}^p(\varphi)\, d\varphi + \int_0^\varepsilon \int_0^t p x_k^{p-1}(s,\varphi)(a(x_k)\theta - b(x_k))\, ds\, d\varphi \end{aligned} \tag{A.5}$$

$$\begin{aligned}
&\le \int_0^\varepsilon x_{0k}^p(\varphi)\,d\varphi + Cp\int_0^\varepsilon\int_0^t x_k^{p-1}(1+x_k)\,ds\,d\varphi \\
&\le \int_0^\varepsilon x_{0k}^p(\varphi)\,d\varphi + C\int_0^\varepsilon\int_0^t ((2p-1)x_k^p+1)\,ds\,d\varphi \quad \text{by (A.1)} \\
&\le \int_0^\varepsilon x_{0k}^p(\varphi)\,d\varphi + CT\varepsilon + C(2p-1)\int_0^t\int_0^\varepsilon x_k^p\,d\varphi\,ds.
\end{aligned}$$

By Gronwall's inequality,

$$\begin{aligned}
\text{(A.6)}\qquad \int_0^\varepsilon x_k^p(t,\varphi)\,d\varphi \le{}& \int_0^\varepsilon x_{0k}^p(\varphi)\,d\varphi + CT\varepsilon \\
&+(2p-1)Ce^{(2p-1)Ct}\left(\int_0^\varepsilon x_{0k}^p(\varphi)\,d\varphi + CT\varepsilon\right).
\end{aligned}$$

We can assume without loss of generality that $x_{0k}\to x_0$ in $L^p((0,1))$. Hence $\sup_k \int_0^\varepsilon x_{0k}^p(\varphi)\,d\varphi \to 0$ as $\varepsilon\to 0$. By (A.6), $\sup_k\int_0^\varepsilon x_k^p(t,\varphi)\,d\varphi\to 0$ as $\varepsilon\to 0$ and (2) is proved.

Now let us prove (3). Assume $t_1<t_2$.

$$\begin{aligned}
\text{(A.7)}\quad \int_0^1 |x_k^p(t_1,\varphi)-x_k^p(t_2,\varphi)|\,d\varphi &= p\int_0^1\left|\int_{t_1}^{t_2} x_k^{p-1}\partial_t x_k(t,\varphi)\,dt\right| d\varphi \\
&= p\int_0^1\left|\int_{t_1}^{t_2} x_k^{p-1}(a(x_k)\theta(t)-b(x_k))\,dt\right| d\varphi \\
&\le Cp\int_0^1\int_{t_1}^{t_2} x_k^{p-1}(1+x_k)\,dt\,d\varphi \\
&\le C\int_0^1\int_{t_1}^{t_2}((2p-1)x_k^p+1)\,dt\,d\varphi \quad \text{by (A.1)} \\
&\le C(2p-1)(G(T)+1)|t_2-t_1| \quad \text{by (A.3)}.
\end{aligned}$$

Thus (3) is true and the proposition is proved. ∎

REFERENCES

[1] N. C. Bartelt, W. Theis, and R. M. Tromp, *Ostwald ripening of two-dimensional islands on Si(001)*, Phys. Rev. B, 54 (1996), pp. 11741–11751.

[2] E. F. Beckenbach and R. Bellman, *Inequalities*, Springer-Verlag, Berlin, 1965.

[3] S. Dai and R. L. Pego, *An upper bound on the coarsening rate for mushy zones in a phase field model*, Interfaces Free Bound., to appear.

[4] L. C. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, American Mathematical Society, Providence, RI, 1990.

[5] H. Federer, *Geometric Measure Theory*, Grundlehren Math. Wissenschaften, Band 153, Springer-Verlag, Berlin, Heidelberg, 1969.

[6] R. V. Kohn and F. Otto, *Upper bounds on coarsening rates*, Comm. Math. Phys., 229 (2002), pp. 375–395.

[7] R. V. Kohn and X. Yan, *Coarsening rates for models of multicomponent phase separation*, Interfaces Free Bound., 6 (2004), pp. 135–149.

[8] R. V. Kohn and X. Yan, *Upper bounds on the coarsening rate for an epitaxial growth model*, Comm. Pure Appl. Math., 56 (2003), pp. 1549–1564.

[9] I. M. Lifshitz and V. V. Slyozov, *The kinetics of precipitation from supersaturated solid solutions*, J. Phys. Chem. Solids, 19 (1961), pp. 35–50.

[10] B. Niethammer, *Derivation of the LSW theory for Ostwald ripening by homogenization methods*, Arch. Ration. Mech. Anal., 147 (1999), pp. 119–178.

[11] B. Niethammer, *The LSW model for Ostwald ripening with kinetic undercooling*, Proc. Roy. Soc. Edinburgh Sect. A., 130 (2000), pp. 1337–1361.
[12] B. Niethammer and R. L. Pego, *Non-self-similar behavior in the LSW theory of Ostwald ripening*, J. Statist. Phys., 95 (1999), pp. 867–902
[13] B. Niethammer and R. L. Pego, *Well-posedness for measure transport in a family of nonlocal domain coarsening models*, Indiana Univ. Math. J., 54 (2005), pp. 499–530.
[14] J. Rubinstein and P. Sternberg, *Nonlocal reaction-diffusion equations and nucleation*, IMA J. Appl. Math., 48 (1992), pp. 249–264.
[15] V. V. Slezov and V. V. Sagalovich, *Diffusive decomposition of solid solutions*, Soviet Phys. Uspekhi, 30 (1987), pp. 23–45.
[16] C. Wagner, *Theorie der Alterung von Niederschlägen durch Umlösen*, Z. Elektrochem., 65 (1961), pp. 581–594.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 372–410

ON A HELE–SHAW-TYPE DOMAIN EVOLUTION WITH CONVECTED SURFACE ENERGY DENSITY*

MATTHIAS GÜNTHER† AND GEORG PROKERT‡

Abstract. Interest is directed to a moving boundary problem with a gradient flow structure which generalizes surface tension–driven Hele–Shaw flow to the case of nonconstant surface tension coefficient taken along with the liquid particles at the boundary. In the case with kinetic undercooling regularization, well-posedness of the resulting evolution problem in Sobolev scales is proved, including cases in which the surface tension coefficient degenerates. The problem is reformulated as a vector-valued, degenerate parabolic Cauchy problem. To solve this, we prove and apply an abstract result on Galerkin approximations with variable bilinear forms.

Key words. free boundary motion, degenerate nonlocal parabolic evolution

AMS subject classifications. 35R35, 76B07

DOI. 10.1137/S0036141004444846

1. Introduction. Various experimental studies investigate the influence of spatial variations of the surface energy density (corresponding to the surface tension coefficient γ) on surface tension–driven Hele–Shaw flows (cf., e.g., [15]). However, a mathematical model for such flows seems to be lacking. In this paper, a first step is attempted to close this gap. We derive and investigate a moving boundary problem which arises, at least from a mathematical point of view, as a natural generalization from the case where γ is a positive constant to the case of variable, nonnegative γ. Let us give an informal description of this generalization here; for details we refer to section 2.

The Hele–Shaw moving boundary problem with constant γ is well investigated. In particular, our starting point is the following observation [1, 8]: On the Fréchet manifold M of the surfaces Γ that bound a domain of fixed given volume, an evolution $t \mapsto \Gamma_t$ satisfying the moving boundary problem can be interpreted as a gradient flow with respect to (w.r.t.) the energy functional

$$\mathcal{E} = \mathcal{E}(\Gamma) := \gamma \operatorname{meas}(\Gamma) \tag{1.1}$$

and the Riemannian metric g given by (2.4).

In our generalization to nonconstant γ we use the energy functional

$$\mathcal{E} = \mathcal{E}(\Gamma, \gamma) := \int_\Gamma \gamma \, d\Gamma \tag{1.2}$$

and keep the demand that the evolution be given by a gradient flow w.r.t. the same Riemannian metric. (A parallel procedure applied to viscous free boundary flows leads to the usual description of the Marangoni effect.) This leads to two related difficulties: First, the functional $\mathcal{E}$ no longer depends on Γ only. This is resolved in

*Received by the editors June 25, 2004; accepted for publication (in revised form) December 30, 2004; published electronically October 14, 2005.
http://www.siam.org/journals/sima/37-2/44484.html

†Mathematisches Institut, Universität Leipzig, Augustusplatz 10/11, 04109 Leipzig, Germany (Matthias.Guenther@math.uni-leipzig.de).

‡Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (g.prokert@tue.nl).

the following way: Instead of the manifold M we consider the vector bundle F over M, having as fiber space at Γ the (smooth) functions on Γ. On this bundle, $\mathcal{E}$ is well defined. Second, one also has to prescribe an evolution law for γ as a function on the moving surface $t \mapsto \Gamma_t$. Again, we make a simple choice: We assume γ to be transported along with the velocity field at the boundary, and we allow the tangential transport to be diminished by a "slip factor" $\delta \in [0, 1]$. The case $\delta = 1$ describes a fixed coupling of the values of γ to the moving liquid particles. Physically, this would occur, e.g., if γ is temperature-dependent and heat conduction is negligible. On the other hand, the case $\delta = 0$ corresponds to transport in the normal direction only. In differential geometric terms, this transport law is realized by introducing a suitable connection D on F and demanding parallel transport of γ; see (2.7)–(2.10).

Let us note here that we do not claim that these assumptions are necessarily in accordance with the physics of an actual Hele–Shaw flow with nonuniform surface energy density, e.g., induced by the presence of a surfactant. It is well conceivable that the interface dynamics in such a situation might be dominated by more complex phenomena like the occurrence of boundary layers, thin surfactant films, or other effects. For instance, if a surfactant is present, one has to solve a transport equation for the surfactant concentration and determine γ from this. (See [17] for the case of Stokes flow; such a modification of our problem would not present new principal difficulties.)

To test our assumptions in a concrete situation, numerical work as well as comparison with experiments would be necessary. However, even our simple model is of mathematical interest in its own right and as a typical example for nonlocal, degenerate parabolic evolutions.

In section 2 of this paper we derive the moving boundary problem (2.12), (2.13) from the gradient flow formulation. In what follows, we prove our main result, namely, a local existence and uniqueness result for this problem in scales of Sobolev spaces. For the precise formulation and further results concerning continuous dependence on the initial data, see Theorems 3.1 and 3.2. If the surface Γ_t and the coefficient γ_t are known at some time t, then the velocity potential ϕ_t is completely determined by (2.12). If one parametrizes Γ_t over a fixed reference surface S, the moving boundary problem can be interpreted as an evolution equation with nonlinear, nonlocal pseudodifferential operators. The parametrization can be constructed in at least two different ways: On one hand, it is possible to parametrize the surfaces using one scalar function, e.g., the normal distance to the reference surface. Then γ_t has to satisfy a transport equation whose coefficients depend on the parametrization and on the velocity potential. On the other hand, the moving boundary can also be represented by mappings $u(\cdot, t) : S \to \mathbb{R}^m$ whose time derivatives are given by the velocity vector, i.e.,

$$\partial_t u = \mathcal{F}(u) := \big((\nabla_N + \delta\nabla_T)\phi_t\big) \circ u,$$

where ∇_N and ∇_T denote the normal and tangential component of the gradient, respectively.

This formulation, which we will use in what follows, is $\mathbb{R}^m$-valued. Therefore, the corresponding Cauchy problem will be necessarily degenerate, even if γ is strictly positive (or even constant). However, this is no crucial disadvantage, as our problem couples a transport equation with a parabolic evolution and we allow γ to degenerate as well. Our approach has two favorable properties: γ now appears only as a known, time-independent function on the reference domain, and the additional freedom in the choice of the diffeomorphisms can be used to derive generalized chain rules for our nonlocal operators which reduce the technical effort in the proofs of the necessary estimates.

As long as γ is nonnegative, the normal component of the linearization $\mathcal{F}'(u)v$ behaves as a degenerate elliptic second order operator on the normal component of v, so that we have, e.g., w.r.t. the L^2-inner product

$$\left\langle n\cdot v, n\cdot\mathcal{F}'(u)v\right\rangle_{L^2} \le C\|v\|_{L^2}^2.$$

An estimate like this does not hold for the complete linearization, including the tangential components. Due to the special structure of $\mathcal{F}$, however, it is possible to define inner products $\langle\ ,\ \rangle_u$ which define equivalent norms on L^2 and satisfy

$$\left\langle v, \mathcal{F}'(u)v\right\rangle_u \le C\|v\|_{L^2}^2.$$

Defining higher order inner products $\langle\ ,\ \rangle_{u,s}$ on the basis of $\langle\cdot,\cdot\rangle_u$, one finally can show an H^s-energy estimate

$$\left\langle \mathcal{F}(u), u\right\rangle_{u,s} \le C\|u\|_s^2 \tag{1.3}$$

for s sufficiently large, and, on the other hand, the dependence of these inner products on u can be controlled by a weaker Sobolev norm. In a suitable abstract functional analytic framework, these estimates can be used to obtain proofs for our main results. Moreover, these results imply the existence of a unique solution $w := (\mathrm{Id}-\lambda\mathcal{F})^{-1}$ of the equation $w-\lambda\mathcal{F}(w)=v$ provided $\lambda\ge 0$ sufficiently small. The solution of the Cauchy problem for the evolution equation is given by the exponential formula

$$u(t) = \lim_{n\to\infty}\Big(\mathrm{Id}-\frac{t}{n}\mathcal{F}\Big)^{-n}u(0)$$

with convergence in H^s provided $u(0)\in H^s$, s sufficiently large.

The structure of the paper after section 2 is as follows: In section 3, we introduce the necessary notation and announce our main results together with the abstract existence theorems which are used. Section 4 is devoted to the behavior of our (nonlocal) operators in scales of Sobolev spaces, and in section 5 the u-dependent inner products are introduced and the necessary estimates are shown. Finally, the main results (Theorems 3.1 and 3.2) are proved in section 6. The proof of a general abstract existence result (Theorem 3.4), which may be of independent interest, is given in the appendix.

2. The equations of motion. Here we characterize the moving boundary problem as abstract gradient flow on the manifold of the natural configuration space. We start by recalling the following general properties of incompressible, source-free Hele–Shaw flows. One looks for a family of domains $\Omega(t)\subset\mathbb{R}^m$ parametrized by time $t\ge 0$ and corresponding velocity fields $\mathbf{v}(\cdot,t)$ such that (according to Darcy's law)

$$\mathbf{v}(\cdot,t) = \nabla\varphi(\cdot,t)\ \text{ in }\ \Omega(t) \tag{2.1}$$

with a potential field $\varphi(\cdot,t)$ proportional to negative pressure. As we also demand that the boundary $\Gamma(t)$ of $\Omega(t)$ moves along with the velocity field, we find the kinematic boundary condition

$$V_n(t) = \partial_n\varphi(\cdot,t)\ \text{ on }\ \Gamma(t), \tag{2.2}$$

where $V_n(t)$ is the normal velocity of the moving boundary $\Gamma(t)$ and $\partial_n=\partial/\partial n$ is the derivative in direction of the unit outward normal $n(t)$ of $\Gamma(t)$. As $\mathbf{v}$ is divergence-free,

$$\Delta\varphi(\cdot,t) = 0\ \text{ in }\ \Omega(t). \tag{2.3}$$

Thus, in any Hele–Shaw flow, the complete velocity field is determined by the normal velocity at the boundary.

If the surface tension coefficient is a positive constant, the corresponding surface energy is proportional to the surface area and the Hele–Shaw flow driven by surface tension can be interpreted as abstract gradient flow of this functional w.r.t. an appropriately chosen inner product; cf. [1, 8]. As this formulation is a main ingredient in our derivation of the moving boundary problem below, we define this inner product more precisely. Consider for the time being a fixed smooth domain Ω with boundary Γ and define

$$V_\Gamma := \left\{ v \in C^\infty(\Gamma) \;\middle|\; \int_\Gamma v \, d\Gamma = 0 \right\}.$$

The space V_Γ can be interpreted as the space of all possible normal boundary velocities; the restriction expresses conservation of volume. We fix $\beta \geq 0$ and introduce on V_Γ the bilinear form g_Γ given by

$$g_\Gamma(v_1, v_2) := \int_\Omega \nabla\varphi_1 \nabla\varphi_2 \, dx + \beta \int_\Gamma v_1 v_2 \, d\Gamma, \tag{2.4}$$

where the φ_i, $i = 1, 2$, are (weak) solutions of the Neumann problems

$$\Delta\varphi_i = 0 \;\text{ in }\; \Omega, \quad \partial_n \varphi_i = v_i \;\text{ on }\; \Gamma.$$

To give a physical interpretation of the quadratic functional $v \mapsto g_\Gamma(v, v)$ we remark that the first term represents energy dissipation by the corresponding Hele–Shaw flow (cf. [8]), while for $\beta > 0$ the second term is a penalty for large normal boundary velocities. Note that, by Green's formula,

$$g_\Gamma(v_1, v_2) = \int_\Gamma \big(\varphi_1 + \beta \partial_n \varphi_1\big) v_2 \, d\Gamma. \tag{2.5}$$

In differential geometric terms, this inner product defines a Riemannian metric on the Fréchet manifold M of boundaries $\Gamma = \partial\Omega$ to smooth compact domains $\Omega \subset \mathbb{R}^m$ with given fixed volume. By interpreting a tangent vector at $\Gamma \in M$ as the normal velocity field of the boundary, there is a natural way of thinking of vector fields X on M as sections in the Fréchet vector bundle $E = \cup_{\Gamma \in M} V_\Gamma$ with base M and fiber V_Γ; i.e., there is a natural isomorphism $T_\Gamma M \simeq V_\Gamma$ and we have for any real functional $J \in C^\infty(M)$

$$(XJ)(\Gamma) = \partial_\varepsilon J(\Gamma_\varepsilon)|_{\varepsilon=0},$$

where $\varepsilon \mapsto \Gamma_\varepsilon \in M$ is a path of admissibles shapes with normal velocity v for $\varepsilon = 0$,

$$\Gamma_\varepsilon := \big\{ x_\varepsilon \;\big|\; x \in \Gamma \big\}, \qquad x_\varepsilon := x + \varepsilon\big(v(x) + O(\varepsilon)\big) n(x). \tag{2.6}$$

Thus, identifying a vector $X_\Gamma \in T_\Gamma M$ in this sense with its image $v \in V_\Gamma$ and considering smooth domain dependence of the solution to a Neumann problem, $\Gamma \mapsto g_\Gamma$ defines a Riemannian metric g on the manifold M. It is remarkable that in the case $\beta = 0$ a geodetic line w.r.t. this metric g represents the motion of an incompressible irrotational perfect fluid with a free boundary; for the corresponding Levi-Civita derivative, Riemannian curvature, and an analysis of the Jacobi equation from a differential geometric point of view, we refer to [3].

Now, considering first the surface energy (1.1) with constant γ, the normal velocity $V_n \in V_\Gamma$ of a surface tension–driven Hele–Shaw flow is determined by

$$g_\Gamma\big(V_n, v\big) = -\mathcal{E}'(\Gamma)\{v\} \;\text{ for all }\; v \in V_\Gamma,$$

where $\mathcal{E}'(\Gamma)\{v\} := (X\mathcal{E})(\Gamma)$ denotes the derivative of the energy in the direction of $X_\Gamma \simeq v$. As a consequence, at each instant of time t the flow reduces the surface energy as rapidly as possible among all normal velocities with prescribed norm corresponding to the inner product (2.4); in particular, the flow is volume preserving and surface area decreasing. By a well-known formula for the first variation of surface area, we find

$$\mathcal{E}'(\Gamma)\{v\} = -\int_\Gamma \kappa v \, d\Gamma,$$

where κ is the mean curvature of Γ with sign determined by the above variation formula (negative sign if Ω is convex); for notational convenience, throughout the paper the usual normalization of κ has been changed by a cofactor $m-1$. To model the influence of a variable surface tension coefficient which is coupled on a transport mechanism, it is now quite natural to consider a surface energy functional of the form (1.2) where $\gamma \geq 0$ denotes a surface energy density function along Γ, not necessarily constant. It should be noted that we don't assume a priori a one-to-one correspondence between the surface Γ and density γ, as is the case in simpler situations, e.g., where a known global function generates the density via restriction or where an anisotropic surface energy density is considered, i.e.,

$$\gamma = f|_\Gamma \quad \text{or} \quad \gamma = f \circ n \;\text{ on }\; \Gamma,$$

given $f \in C^\infty(\mathbb{R}^m)$ or $f \in C^\infty(\mathbb{S}^{m-1})$, respectively; the latter energy density is commonly used to model crystal growth problems. In fact, in our setting the functional $\mathcal{E}$ is uniquely defined on the vector bundle $F := \cup_{\Gamma \in M} C^\infty(\Gamma)$ with base M only. In such a case, computation of the derivative of the surface energy in the direction of a given vector field requires a law for the change of γ on the moving surface. Using differential geometric terms we make the following assumption: along a path c in M the energy density is transported by parallel displacement w.r.t. a given connection D_X which acts on sections γ in F, i.e.,

$$D_{\dot{c}}\gamma = 0 \;\text{ along }\; c. \tag{2.7}$$

In further considerations we restrict our attention to the connection D_X, defined as follows: Let X be any vector field, let $\Gamma \in M$, and let $v \in V_\Gamma$ with $X_\Gamma \simeq v$; then we set for any section γ in F

$$D_X\gamma|_\Gamma := \partial_\varepsilon \bar{\gamma}_\varepsilon\,|_{\varepsilon=0} + \delta \nabla_\Gamma \psi \nabla_\Gamma \gamma, \quad \delta \in [0,1], \tag{2.8}$$

where in terms of the notation (2.6)

$$\bar{\gamma}_\varepsilon(x) := \bar{\gamma}_{\Gamma_\varepsilon}(x_\varepsilon), \quad (\varepsilon, x) \in (-\varepsilon_0, \varepsilon_0) \times \Gamma, \tag{2.9}$$

and ψ is a solution of the Neumann problem

$$\Delta\psi = 0 \;\text{ in }\; \Omega, \quad \partial_n \psi = v \;\text{ on }\; \Gamma. \tag{2.10}$$

Interpretation of D_X and parallel transport w.r.t. D_X is quite obvious in terms of the underlying Hele–Shaw flow. In contrast to the case of constant γ, we also have to consider the influence of the tangential motion at the boundary which results from a normal variation of the boundary. As pointed out before, in a Hele–Shaw flow the velocity field corresponding to a normal boundary velocity $v \in V_\Gamma$ is $\nabla\psi$, where ψ solves (2.10). Hence, in the case $\delta = 1$, (2.7), (2.8) express that the surface energy density is transported along with the liquid particles, i.e., with the velocity field $\nabla\psi$ at the boundary. On the other hand, in the case of $\delta = 0$, transport in the normal direction without any tangential movement is expressed. The other cases are intermediate. On V_Γ we define the linear operator A_{ND} (Neumann-to-Dirichlet operator) by

$$A_{ND}v := \psi|_\Gamma,$$

where ψ satisfies (2.10) and $\int_\Gamma \psi \, d\Gamma = 0$. Hence, again in terms of the notation (2.6), the assumption $D_v\gamma|_\Gamma = 0$ implies

$$\partial_\varepsilon \bar{\gamma}_\varepsilon \,|_{\varepsilon=0} = -\delta \nabla_\Gamma \gamma \nabla_\Gamma A_{ND} v$$

and we obtain for

$$\mathcal{E}'(\gamma,\Gamma)\{v\} := \frac{d}{d\varepsilon}\mathcal{E}(\gamma_\varepsilon,\Gamma_\varepsilon)\Big|_{\varepsilon=0}$$

using again the formula for the first variation of area

$$\mathcal{E}'(\gamma,\Gamma)\{v\} = \int_\Gamma \big(\partial_\varepsilon \bar{\gamma}_\varepsilon \,|_{\varepsilon=0} - \kappa\gamma v\big)\, d\Gamma = -\int_\Gamma \big(\kappa\gamma v + \delta\nabla_\Gamma\gamma\nabla_\Gamma\psi\big)\, d\Gamma.$$

It easily follows from Green's formula that A_{ND} is symmetric w.r.t. the usual L^2-inner product on Γ, and thus

$$\mathcal{E}'(\gamma,\Gamma)\{v\} = -\int_\Gamma (\gamma\kappa v - \delta\Delta_\Gamma\gamma\, A_{ND}v)\, d\Gamma = -\int_\Gamma (\gamma\kappa - \delta A_{ND}\Delta_\Gamma\gamma) v \, d\Gamma. \tag{2.11}$$

We have to consider (2.7), (2.11) as a differential rule for the change of surface energy dependent upon surface and energy density. They allow the computation of the energy along any path in M starting from a known initial shape $\Gamma(0)$ with known energy density γ_0. But of course, in general, this computation is path-dependent; i.e., the resulting energy in the endpoint of the path will depend on the history along the whole path.

Now, as in the case of constant γ, we define the normal velocity $V_n \in V_\Gamma$ as a solution of the variational problem

$$g_\Gamma(V_n, v) = -\mathcal{E}'(\gamma,\Gamma)\{v\} \text{ for all } v \in V_\Gamma.$$

Together with (2.2), (2.5), and (2.11), this yields the dynamic boundary condition

$$\varphi + \beta\partial_n\varphi = \gamma\kappa - \delta A_{ND}\Delta_\Gamma\gamma.$$

Summarizing and using an auxiliary function ψ instead of the nonlocal operator A_{ND}, we have obtained the following moving boundary problem: For a given bounded domain $\Omega(0) \subset \mathbb{R}^m$ and a given nonnegative function γ_0 defined on $\partial\Omega(0)$ one looks for a family of C^2-domains $\Omega(t) \subseteq \mathbb{R}^m$, $t > 0$, and functions $\varphi(\cdot,t), \psi(\cdot,t) \in C^2\big(\overline{\Omega(t)}\big)$, $\gamma_t \in C^2\big(\Gamma(t)\big)$ such that

$$
\left\{
\begin{aligned}
\Delta\varphi(\cdot,t) &= 0 && \text{in } \Omega(t),\\
\Delta\psi(\cdot,t) &= 0 && \text{in } \Omega(t),\\
\partial_n\psi(\cdot,t) &= \Delta_{\Gamma(t)}\gamma_t && \text{on } \Gamma(t),\\
\varphi(\cdot,t) + \beta\partial_n\varphi(\cdot,t) &= \gamma_t\kappa(t) - \delta\psi(\cdot,t) && \text{on } \Gamma(t),\\
V_n(t) &= \partial_n\varphi(\cdot,t) && \text{on } \Gamma(t),
\end{aligned}
\right.
\tag{2.12}
$$

where $\kappa(t)$ is the curvature of $\Gamma(t)$. In the main part of this paper, we restrict our attention to the case $\delta = 1$. The generalization to $\delta \in [0,1)$ is sketched at the end of section 5. Additionally, we describe the transport of γ by (2.7), (2.8) with $\delta = 1$. Introducing Lagrangian coordinates $x = x(\xi,t)$, $\xi \in \Gamma(0)$ corresponding to the velocity field via

$$
\partial_t x(\xi,t) = \nabla\varphi\big(x(\xi,t),t\big) \quad \text{for} \quad t \geq 0, \qquad x(\xi,0) = \xi,
\tag{2.13}
$$

we obtain from (2.2) that $x = x(\cdot,t)$ is a diffeomorphism from $\Gamma(0)$ onto $\Gamma(t)$, and the transport law for γ_t takes the form

$$
\gamma_t\big(x(\xi,t)\big) = \gamma_0(\xi), \qquad \xi \in \Gamma(0),\ t \geq 0.
\tag{2.14}
$$

In (2.12), $\varphi(\cdot,t)$ and $\psi(\cdot,t)$ are determined up to a constant only, but this is without significance for the evolution of both $\Omega(t)$ and γ_t. Note that in the case $\beta = 0$, by setting $\Phi = \varphi + \psi$, (2.12) simplifies to

$$
\left\{
\begin{aligned}
\Delta\Phi(\cdot,t) &= 0 && \text{in } \Omega(t),\\
\Phi(\cdot,t) &= \gamma_t\kappa(t) && \text{on } \Gamma(t),\\
V_n &= \partial_n\Phi(\cdot,t) - \Delta_{\Gamma(t)}\gamma_t && \text{on } \Gamma(t).
\end{aligned}
\right.
\tag{2.15}
$$

In what follows, however, we will restrict our attention to the case $\beta > 0$. Without loss of generality, we can assume $\beta = 1$. In the case $\beta > 0$, $\delta = 1$, we can show well-posedness of our moving boundary problem even if γ is zero on parts of the boundary, provided its square root is smooth. This seems to be particular to this situation. We intend to discuss the case $\beta = 0$, which leads to a third order problem, in a forthcoming paper.

For $\gamma_t = \gamma = \text{const}$ and $\gamma > 0$, ψ is constant, and (2.12) is known as the so-called Hele–Shaw flow problem with kinetic undercooling and surface tension regularization. From a modeling point of view, this problem can be seen as the quasi-stationary version of the well-known Stefan problem. In this context, the boundary condition incorporates both the Gibbs–Thomson surface energy and a nonequilibrium effect of temperature decrease at the advancing phase boundary. A short-time existence proof for this problem and a proof that its solution is the limit for the solutions of the corresponding Stefan problems can be found in [20]. For existence results concerning a corresponding two-phase problem we refer to [5, 21]. Both effects are known to regularize the motion of the interface, and this is also true for Hele–Shaw flow problems [13, 18, 19]. In the case $\gamma \equiv 0$, with internal sources or sinks as driving forces, existence results are given in [11] for the two-dimensional case and analytic data and in [16] for arbitrary dimensions in the framework of Sobolev spaces.

If γ is a positive constant, the moving boundary has stable, attractive equilibria which are given by the spheres (see, e.g., [4, 6, 7] for the case $\beta = 0$). In general, however, after prescribing a nonconstant function γ on the reference domain and an initial diffeomorphism u, it is not a priori clear (even with γ near a constant and the moving domain near a ball) what the long-time evolution and the corresponding equilibrium will be. Instead, determining the equilibria belonging to a γ prescribed on the reference domain leads to a stationary free boundary problem in ψ whose solvability and stability (for Γ near a sphere and γ near a constant) we intend to discuss elsewhere.

3. Notation and main results. We list some notation. $C, C_1, \dots$, etc., denote generic constants; their dependences on other quantities is indicated only if not obvious from the context. Let $E \subseteq \mathbb{R}^m$, $m \geq 2$, be a bounded domain with smooth boundary $S := \partial E$ and ν the outer unit normal on S. For $M = S$ or $M = E$, we make constant use of the usual L^2-based Sobolev spaces $H^s(S)$, $H^s(S, \mathbb{R}^m)$ of order s with values in $\mathbb{R}$ and $\mathbb{R}^m$, respectively. If no confusion is likely, we just write H^s. The norms of these spaces will be denoted by $\|\cdot\|_s^M$; for $M = S$ the upper index M is dropped in most cases. When Fréchet derivatives of operator-valued mappings are considered, the additional arguments describing the variations are written in braces ($\{\,\}$).

3.1. Well-posedness for the moving boundary problem. Now, as already mentioned in the introduction, we reformulate the moving boundary problem (2.12)–(2.14) by describing $\Gamma(t)$ as an embedding $u(\cdot, t) : S \to \mathbb{R}^m$ such that the curves $t \mapsto u(y, t)$ for fixed $y \in S$ are trajectories belonging to the velocity field and γ_t is constant along these curves. This approach enables us to consider γ_t as a known function during the evolution at the cost of describing the moving boundary by m functions. To do so, let

$$U := \bigl\{u : S \to \mathbb{R}^m \bigm| u = w|_S \text{ with } w \in \mathrm{Diff}(\bar{E}, \Omega_u \cup \Gamma_u)\bigr\}, \tag{3.1}$$

where

$$\Omega_u := w(E) \quad \text{and} \quad \Gamma_u := \partial\Omega_u = u(S).$$

Throughout this paper, we use the abbreviation

$$U_s := U \cap H^s(S, \mathbb{R}^m).$$

Now, (2.12)–(2.14) is reduced to the following Cauchy problem, which will be investigated in what follows: Given $u_0 \in U_s$, s sufficiently large, we look for $T > 0$ and a mapping $[0, T] \ni t \mapsto u(t) \in U_s$, such that

$$u'(t) = \mathcal{F}\bigl(u(t)\bigr), \quad t \in [0, T], \tag{3.2}$$

$$u(0) = u_0. \tag{3.3}$$

Thereby, for $u \in U$, we have set

$$\mathcal{F}(u) := F(u)\bigl(\mathcal{G}(u)\bigr) \quad \text{with} \quad \mathcal{G}(u) := H(u) + G(u), \tag{3.4}$$

where, for any given function f on S,

$$F(u)f := \nabla\varphi(u, f) \circ u, \tag{3.5}$$

and $\varphi = \varphi(u, f)$ denotes the solution of the Robin boundary value problem

$$\Delta\varphi = 0 \text{ in } \Omega_u, \quad \partial_n\varphi + \varphi = f \circ u^{-1} \text{ on } \Gamma_u. \tag{3.6}$$

Further, $H(u)$, $G(u)$ are given by

$$H(u) := \gamma(\kappa_{\Gamma_u} \circ u), \quad G(u) := -A(u)\big(\Delta(u)\gamma\big). \tag{3.7}$$

Here $\gamma \in C^\infty(S)$ is a fixed and given nonnegative function, κ_{Γ_u} denotes the mean curvature of Γ_u with sign and scaling conventions as above,

$$\Delta(u)w := \Delta_{\Gamma_u}(w \circ u^{-1}) \circ u \tag{3.8}$$

is the pullback to S of the Laplace–Beltrami operator Δ_{Γ_u} on Γ_u, and

$$A(u)f := \varphi_N\big(u, f\big) \circ u \tag{3.9}$$

is the Neumann–Dirichlet operator, i.e., $\varphi_N = \varphi_N(u, f)$ solves the Neumann problem

$$\Delta\varphi_N = 0 \text{ in } \Omega_u, \quad \partial_n \varphi_N = c + f \circ u^{-1} \text{ on } \Gamma_u, \quad \textstyle\int_{\Gamma_u} \varphi_N \, dx = 0. \tag{3.10}$$

The constant $c = c(u, f) \in \mathbb{R}$ in (3.10) is determined by the solvability condition

$$\int_{\Gamma_u} (f \circ u + c)\, d\Gamma_u = 0; \tag{3.11}$$

clearly $c(u, f) = 0$ for $f = \Delta(u)\gamma$. For fixed smooth γ on S, the mappings $u \mapsto H(u)$ and $u \mapsto \Delta(u)\gamma$ constitute quasi-linear second order differential operators on S. Moreover, the solutions of the boundary value problems (3.6), (3.10) depend smoothly on the domain Ω_u, i.e., on $u \in H^s$, $s > (m+1)/2$, and $f \mapsto F(u)f$, $f \mapsto A(u)f$ represent pseudodifferential operators of order zero and minus one, respectively. In particular, G is a pseudodifferential operator of lower order than H and may be considered as a correction term to ensure the gradient flow structure of the evolution problem. For precise formulations of the mapping properties of F and A and detailed proofs, see section 4. As a consequence, this leads to

$$[u \mapsto \mathcal{F}(u)] \in C^\infty\big(U_s, H^{s-2}(S, \mathbb{R}^m)\big) \tag{3.12}$$

for $s > (m+3)/2$. Now we are in position to formulate our main results.

THEOREM 3.1 (short-time existence and uniqueness). *Fix an integer $s_0 > (m+5)/2$ and assume $\gamma = \rho^2$ with $\rho \in C^\infty(S)$. Let $s \geq s_0$ be an integer and let $u_0 \in U_s$. Then there exists $T > 0$ and a unique solution*

$$u \in C\big([0, T], U_s\big) \cap C^1\big([0, T], H^{s-2}(S, \mathbb{R}^m)\big) \tag{3.13}$$

of the initial value problem (3.2), (3.3). *Additionally, any given $\bar{u}_0 \in U_{s_0}$ has a suitable H^{s_0}-neighborhood K, such that for initial values u_0 varying in $K \cap H^s$, there are $T > 0$ and C independent of u_0 such that*

$$\|u(t)\|_s \leq C(1 + \|u(0)\|_s) \text{ for all } t \in [0, T]. \tag{3.14}$$

THEOREM 3.2 (regularity and continuous dependence on initial values). *Under the assumptions of Theorem* 3.1 *let u be any solution to* (3.2) *in the class* (3.13) *with some $T > 0$. Then the following holds:*

(i) $u(0) \in H^{s+1}(S, \mathbb{R}^m)$ *implies*

$$u \in C\big([0, T], U_{s+1}\big) \cap C^1\big([0, T], H^{s-1}(S, \mathbb{R}^m)\big).$$

(ii) *Assume $u_0^n \to u_0$ in $H^s(S, \mathbb{R}^m)$ for $n \to \infty$. Then, for n sufficiently large, there exist solutions u_n of* (3.2) *in the class* (3.13) *with initial values $u_n(0) = u_0^n$, and there holds $u_n \to u$ in $C\big([0, T], H^s(S, \mathbb{R}^m)\big)$.*

The proofs of both theorems are given in section 6.

3.2. An existence result for abstract evolution equations. Here we consider (3.2), (3.3) as an abstract nonlinear Cauchy problem for an unknown function $u = u(t)$ with values in a Banach space and prove existence of a solution if the nonlinearity $\mathcal{F}$ satisfies a certain condition of semiboundedness w.r.t. a family of bilinear forms. As a general framework we adopt the following assumptions:

$$\text{(H)}\quad \begin{cases} \text{Let } X \subseteq Y \subseteq Z \text{ be real, separable Banach spaces with dense and} \\ \text{continuous embeddings and } \mathcal{U} \subseteq Y \text{ open. For every } u \in \mathcal{U} \text{ let } \langle \cdot,\cdot \rangle_u : \\ X \times Z \to \mathbb{R} \text{ be a continuous and nondegenerate bilinear form, such} \\ \text{that with fixed constants } C \geq 1,\ M \geq 0, \\ \text{(H1)} \quad \langle v, w\rangle_u = \langle w, v\rangle_u \text{ for all } v, w \in X; \\ \text{(H2)} \quad C^{-1}\|v\|_Y^2 \leq \langle v,v\rangle_u \leq C\|v\|_Y^2 \text{ for all } v \in X,\ u \in \mathcal{U}; \\ \text{(H3)} \quad \langle v,v\rangle_u \leq \langle v,v\rangle_w \big(1 + M\|u-w\|_Z\big) \text{ for all } v \in X,\ u,w \in \mathcal{U}; \\ \text{(H4)} \quad \text{weak convergences } u_n \rightharpoonup u \text{ in } Y,\ u_n, u \in \mathcal{U}, \text{ and } w_n \rightharpoonup w \text{ in} \\ \qquad\quad Z \text{ imply } \langle v, w_n\rangle_{u_n} \to \langle v, w\rangle_u \text{ for all } v \in X. \end{cases}$$

Assuming (H) holds, by the dense embedding $X \subseteq Y$ and

$$\big|\langle v, w\rangle_u\big|^2 \leq \langle v,v\rangle_u \langle w,w\rangle_u \leq C^2\|v\|_Y^2\|w\|_Y^2 \ \text{ for } \ v, w \in X$$

there exists for each $u \in \mathcal{U}$ an inner product $(\cdot,\cdot)_u$ on Y, which is compatible with $\langle\cdot,\cdot\rangle_u$; i.e., we have

$$(v,w)_u = \langle v,w\rangle_u \ \text{ for } \ v \in X, w \in Y.$$

Moreover, for $u_n, u \in \mathcal{U}$, $u_n \rightharpoonup u$, $w_n \rightharpoonup w$ in Y implies

$$(v, w_n)_{u_n} \to (v,w)_u \ \text{ for all } \ v \in X.$$

In further considerations, for the sake of brevity we put

$$\|v\|_u = (v,v)_u^{1/2}, \quad |||u||| = (u,u)_u^{1/2}\ .$$

Assumption (H2) implies that $\|\cdot\|_Y$ and $\|\cdot\|_u$ are equivalent, and hence Y has all topological properties of a Hilbert space—in particular, Y is reflexive. From $u_n, u \in \mathcal{U}$, $u_n \rightharpoonup u$ in Y it follows that

$$|||u||| \leq \underline{\lim}_{\,n\to\infty} |||u_n|||\ ;$$

if $|||u||| = \lim_{n\to\infty} |||u_n|||$, one concludes hereby that $u_n \to u$ in Y.

THEOREM 3.3. *Let* (H) *be valid and let* $\mathcal{F} : \mathcal{U} \to Z$ *be a weakly sequentially continuous mapping such that for every* $u_0 \in \mathcal{U}$ *there exists a neighborhood* $B(u_0) \subset \mathcal{U}$ *of* u_0 *in* Y *with*

$$\sup\big\{\langle u, \mathcal{F}(u)\rangle_u \ \big|\ u \in B(u_0) \cap X\big\} < +\infty. \tag{3.15}$$

Then for any $u_0 \in \mathcal{U}$, *there exist* $T > 0$ *and a solution* u *of* (3.2), (3.3) *in the class*

$$C_w\big([0,T],\mathcal{U}\big) \cap C_w^1\big([0,T], Z\big). \tag{3.16}$$

Additionally, this solution satisfies $u(t) \to u_0$ *in* Y *for* $t \to +0$. *Moreover,* $T > 0$ *can be chosen uniformly for initial values taken from a suitable neighborhood of* u_0 *in* Y.

In (3.16), we denote by $C_w\big([0,T],\mathcal{U}\big)$ the space of functions from $[0,T]$ to $\mathcal{U}$ which are continuous w.r.t. weak convergence in Y. Similarly, $C_w^1\big([0,T],Z\big)$ denotes the set of weakly differentiable functions from $[0,T]$ to Z with the derivative in $C_w\big([0,T],Z\big)$. It should be noted that in general there is no uniqueness and no continuous dependence on initial data in any sense in Theorem 3.3. This theorem can be easily derived from a more elaborate, quantitative formulation given in the next theorem. Note that, for the limit case $R=+\infty$ and bilinear forms independent of u, this theorem coincides with Theorem A in [14], but, as already mentioned in the introduction, our application requires only the generalization to such variable bilinear forms.

Theorem 3.4. *Assume* (H) *is satisfied with some ball*

$$\mathcal{U}=B:=\big\{u\in Y\;\big|\;\|u\|_Y<R\big\},\quad R>0,$$

and $\mathcal{G}:B\to Z$ is a weakly sequentially continuous mapping such that

$$2\langle u,\mathcal{G}(u)\rangle_u+M\,\|\mathcal{G}(u)\,\|_Z|||u|||\le\beta\big(|||u|||^2\big)\ \text{ for all }\ u\in X\cap B \tag{3.17}$$

with a C^1-function $\beta:\mathbb{R}_+\to\mathbb{R}_+=[0,\infty)$. Let $u_0\in B$,

$$|||u_0|||<r:=R/(2C^3)^{1/2},$$

and $T>0$ such that the solution ρ of the scalar Cauchy problem

$$d\rho/dt=\beta\big(\rho(t)\big),\quad \rho(0)=|||u_0|||^2 \tag{3.18}$$

exists on $[0,T]$ and satisfies $\rho(t)<r^2$ there. Then the Cauchy problem

$$u'(t)=\mathcal{G}(u(t)),\qquad u(0)=u_0 \tag{3.19}$$

possesses a solution u in the class (3.16) *for which additionally*

$$\begin{gathered}|||u(t)|||^2\le\rho(t)\ \text{ for all }\ t\in[0,T],\\ u(t)\to u_0\ \text{ in }\ Y\ \text{ for }\ t\to+0.\end{gathered}$$

The proof of this theorem will be given in the appendix.

Proof of Theorem 3.3. Let $u_0\in\mathcal{U}$ and $B(u_0)$ as in Theorem 3.3 be given. We set

$$\begin{gathered}\mathcal{G}(v)=\mathcal{F}(v+w_0)\ \text{ for }\ v\in B:=\big\{v\in Y\;\big|\;\|v\|_Y<R\big\},\\ \langle\cdot,\cdot\rangle_{v,1}:=\langle\cdot,\cdot\rangle_{v+w_0},\quad |||v|||_1:=(v,v)_{v+w_0}^{1/2},\end{gathered}$$

whereby the density of X in Y enables us to choose $w_0\in X$ and $R>0$ such that

$$\|w_0-u_0\|_Y<R/(32C^5)^{1/2},\quad \big\{w_0+v\;\big|\;v\in B\big\}\subseteq B(u_0).$$

Clearly, the bilinear form $\langle\cdot,\cdot\rangle_{v,1}$, $v\in B$, satisfies the assumptions (H) again (with the same constants as $\langle\cdot,\cdot\rangle_u$, $u\in\mathcal{U}$). Further, by (3.15), there exists $L>0$ such that

$$\langle v+w_0,\mathcal{G}(v)\rangle_{v,1}\le L\ \text{ for all }\ v\in B\cap X$$

and, by the weak sequential continuity of $\mathcal{F}$, the reflexivity of Y and (H4),

$$\big|\langle w_0,\mathcal{G}(v)\rangle_{v,1}\big|,\|\mathcal{G}(v)\|_Z\le L\ \text{ for all }\ v\in B.$$

Thus

$$2\langle v, \mathcal{G}(v)\rangle_{v,1} + M\|\mathcal{G}(v)\|_Z |||v|||_1 \le K \text{ for all } v \in B \cap X$$

with $K := L(4 + MCR)$. Now, for any given $w \in Y$ with

$$\|w - u_0\|_Y \le R/(32C^5)^{1/2},$$

we apply Theorem 3.4 to solve the initial value problem

$$dv/dt = \mathcal{G}(v), \quad v(0) = w - w_0,$$

which corresponds to (3.2) with initial value $u(0) = w$. As

$$|||w - w_0|||_1 \le C\big(\|w - u_0\|_Y + \|u_0 - w_0\|_Y\big) < r/2, \quad r := R/(2C^3)^{1/2},$$

Theorem 3.4 ensures the existence of a solution in the class (3.16) with $T = 3r^2/(4K)$ and $v(t) \to w - w_0$ in Y for $t \to +0$. ∎

4. Smooth domain dependence of the nonlocal operators. In this section we study the domain dependence, i.e., dependence on $u \in U_s$, of the Robin problem (3.6) by transforming it into a boundary value problem on the fixed reference domain E. In particular, we derive multilinear estimates for the Fréchet derivatives w.r.t. u. The crucial tools here will be estimates for (multiple) pointwise products in our scale of Sobolev spaces and a differentiation rule based on invariance properties of the nonlocal operators.

To begin with, we recall some well-known basic properties and estimates concerning Sobolev spaces. For $s > d/2$, where M is E or S and $d = \dim M/2$, the spaces $H^s(M)$ turn into Banach algebras and the pointwise product of functions $w_1, \dots, w_n \in H^s(M)$ allows the estimate

$$\|w_1 w_2 \cdots w_n\|_s^M \le C \sum_{i=1}^{n} \Big(\|w_i\|_s^M \prod_{j\ne i} \|w_j\|_{s_0}^M\Big) \tag{4.1}$$

if $s \ge s_0 > d/2$. Moreover, for such values of s the composition of C^∞-functions with H^s-functions leads to H^s-functions again: e.g., $\Psi \in C^\infty(\bar{M} \times \mathbb{R})$ and $w \in H^s(M)$ imply $\Psi(\cdot, w(\cdot)) \in H^s(M)$ (note the continuity of w by Sobolev's embedding),

$$\big[w \mapsto \Psi(\cdot, w(\cdot))\big] \in C^\infty\big(H^s(M), H^s(M)\big), \tag{4.2}$$

and it holds that

$$\|\Psi(\cdot, w(\cdot))\|_s^M \le C\big(1 + \|w\|_s^M\big) \tag{4.3}$$

for all w from bounded subsets of $H^{s_0}(M)$, where the constant depends, in addition to s_0, s, and M, on Ψ and on upper bounds of $\|w\|_{s_0}$. In particular,

$$\|1/w\|_s^M \le C\big(\alpha, \|w\|_{s_0}^M\big)\|w\|_s^M \tag{4.4}$$

for all $w \in H^s(M)$ with $w \ge \alpha > 0$ on M. On the other hand, we have the following counterpart of (4.1) for the product of functions $w_1 \in H^{s_1}(M), \dots, w_n \in H^{s_n}(M)$:

$$\|w_1 w_2 \cdots w_n\|_s^M \le c\, \|w_1\|_{s_1}^M \|w_2\|_{s_2}^M \cdots \|w_n\|_{s_n}^M \tag{4.5}$$

if $0 \le s \le s_1, \dots, s_n \le s_0$ with $s_1 + \dots + s_n \ge s + (n-1)s_0$ and $s_0 > d/2$.

In the following, for functions w defined on S let $\mathcal{E}w$ be an extension into $\bar{E}$, i.e., $\mathcal{E}w|_S = w$, whereby the trace mapping theorem permits us to choose

$$\mathcal{E} \in \mathcal{L}\big(H^s(S), H^{s+1/2}(E)\big) \text{ for all } s > 0. \tag{4.6}$$

For $\mathbb{R}^m$-valued functions we apply $\mathcal{E}$ componentwise.

Our first technical concern is the extension of the mapping $u \in U_s$ to a suitable diffeomorphism $\widetilde{u}$ from $\bar{E}$ to $\Omega_u \cup \Gamma_u$. For fixed, smooth $u_0 \in \operatorname{Diff}(\bar{E}, \Omega_u \cup \Gamma_u)$, (4.7) clearly defines a possible extension for all u in U_s such that $\|u_0|_S - u\|_s < \varepsilon$ for sufficiently small $\varepsilon > 0$. However, ε depends on u_0 in an uncontrolled way. Eventually, this would restrict our existence results to evolutions in an open and dense subset of U_s containing $U_s \cap C^\infty(S, \mathbb{R}^m)$ but being uncharacterized otherwise. The following lemma provides a way to avoid this unnecessary restriction.

Lemma 4.1. *Let $v \in U_s$, $s > (m+1)/2$. Then there exist an H^s-neighborhood $V_s \subseteq U_s$ of v and a map $u_0 \in C^\infty(\bar{E}, \mathbb{R}^m)$ such that for every $u \in V_s$ the mapping*

$$\widetilde{u} := u_0 + \mathcal{E}(u - u_0) \tag{4.7}$$

defines a diffeomorphism of $\bar{E}$ onto $\bar{\Omega}_u$.

Proof. By the definition (3.1) of U, every $v \in U$ has an extension $v_1 \in \operatorname{Diff}(\bar{E}, \bar{\Omega}_v)$ and there exists an $\varepsilon > 0$ such that $w \in \operatorname{Diff}(\bar{E}, \bar{\Omega}_{w|_S})$ for all $w \in C^1(\bar{E}, \mathbb{R}^n)$ with $\|w - v_1\|_{C^1}^{\bar{E}} \le \varepsilon$. Thus it suffices to find u_0 and V_s with

$$\|\widetilde{u} - v_1\|_{C^1}^{\bar{E}} \le \varepsilon \text{ for all } u \in V_s, \tag{4.8}$$

where $\widetilde{u}$ is given by (4.7). Let

$$\mathcal{E}_1 \in \mathcal{L}\big(C^1(S, \mathbb{R}^m), C^1(\bar{E}, \mathbb{R}^m)\big) \tag{4.9}$$

be an extension operator which maps $C^\infty(S, \mathbb{R}^m)$ into $C^\infty(\bar{E}, \mathbb{R}^m)$. Setting

$$u_0 = w_1 + \mathcal{E}_1 w_2 \quad \text{with} \quad w_1 \in C^\infty(\bar{E}, \mathbb{R}^m),\ w_2 \in C^\infty(S, \mathbb{R}^m)$$

to be chosen later, we get by Sobolev embedding $H^{s+1/2}(E) \hookrightarrow C^1(\bar{E})$ and (4.6), (4.9)

$$\begin{aligned}
\|\widetilde{u} - v_1\|_{C^1}^{\bar{E}} &\le C\|u - u_0\|_{H^s}^S + \|u_0 - v_1\|_{C^1}^{\bar{E}} \\
&\le C\left(\|w_2\|_{C^1}^S + \|u - u_0\|_{H^s}^S\right) + \|w_1 - v_1\|_{C^1}^{\bar{E}} \\
&\le C\left(\|w_1 - v_1\|_{C^1}^{\bar{E}} + \|w_2 + w_1 - v\|_{H^s}^S + \|v - u_0\|_{H^s}^S + \|u - v\|_{H^s}^S\right).
\end{aligned}$$

Hence, letting $\delta = \varepsilon/(4C)$ and choosing first w_1 with $\|w_1 - v_1\|_{C^1}^{\bar{E}} \le \delta$ and, afterwards, w_2 with $\|w_2 + w_1 - v\|_{H^s}^S \le \delta$, then (4.8) is valid with $V_s = \big\{u \mid \|u - v\|_{H^s}^S < \delta\big\}$. □

Fix $s > (m+1)/2$, $v \in U_s$, and V_s according to Lemma 4.1. Maintaining notation and construction of this lemma, let

$$\bar{E} \ni x \to y = \widetilde{u}(x) = \big(\widetilde{u}_1(x), \dots, \widetilde{u}_m(x)\big) \in \bar{\Omega}_u, \quad u \in V_s, \tag{4.10}$$

be the corresponding diffeomorphism (4.7), $J = (\partial_i \widetilde{u}_j)$ its Jacobian, and $(g_{ij}) = J^\top J$ the Euclidean metric tensor relative to the above coordinates. Furthermore let (g^{ij}) be the inverse of (g_{ij}) and $g = \det(g_{ij})$. Then we have $(g^{ij}) = g^{-1}(\operatorname{Cof} J)^\top(\operatorname{Cof} J)$, where

$\operatorname{Cof} J = (a_{ij})$ and a_{ij} is the algebraic complement of $\partial_i \widetilde{u}_j$ in J. Note that, uniformly in $u \in V_s$, the function g is strictly positive in E. Moreover, for the transformation $\omega = d\Gamma_u/dS$ of surface area elements via (4.10) and the outer normals n of Ω_u and ν of S, the following holds:

$$\omega = |(\operatorname{Cof} J)\nu|, \quad n \circ \widetilde{u} = (\operatorname{Cof} J)\nu / |(\operatorname{Cof} J)\nu|.$$

By definition, all of the quantities g, g_{ij}, a_{ij}, and g^{ij} are polynomials in the first derivatives of $\widetilde{u}$ and, in the case of g^{ij}, additionally in $1/g$. Consequently, remembering (4.1)–(4.5) and the construction (4.7) of $\widetilde{u}$, we obtain smooth dependence of these quantities on u. More precisely, we have

$$[u \mapsto q] \in C^\infty\big(V_s, H^{s-1/2}(E)\big), \quad q = g, g_{ij}, a_{ij}, \text{ or } g^{ij}, \tag{4.11}$$

and (4.5) implies an estimate of the kth Fréchet derivative:

$$\|q^{(k)}(u)\{u_1, \dots, u_k\}\|^E_{t-1/2} \le C\|u_1\|_{s_1} \cdots \|u_k\|_{s_k} \tag{4.12}$$

if $1/2 \le t \le s_1, \dots, s_k \le s$ and $s_1 + \cdots + s_k \ge t + (k-1)s$. The constant is independent of $u \in V_s$ and of $u_1 \in H^{s_1}(S), \dots, u_k \in H^{s_k}(S)$. Similar arguments lead to

$$[u \mapsto p] \in C^\infty\big(V_s, H^{s-1}(S)\big), \quad p = \omega \text{ or } n \circ \widetilde{u}, \tag{4.13}$$

with an estimate of the derivatives corresponding to (4.12).

Now, introducing the transformed velocity potential $\psi = \psi(u)f = \varphi(u, f) \circ \widetilde{u}$ and the transformed Laplace and boundary operator according to

$$L(u)\psi = \partial_i\big(\sqrt{g} g^{ij} \partial_j \psi\big), \quad B(u)\psi = \omega\psi + \nu_i \sqrt{g} g^{ij} \partial_j \psi,$$

the Robin problem (3.6) may be written as

$$L(u)\psi = 0 \ \text{ in } \ E, \quad B(u)\psi = \omega f \ \text{ on } \ S. \tag{4.14}$$

Note that the values of $\psi(u)f$ in E depend not only on u and f, but also on the diffeomorphism $\tilde{u}$, i.e., on the chosen V_s. On the other hand, $\psi(u)f|_S$ is completely determined by u and f. The symmetry of the operator $h \mapsto \varphi(u)h$ w.r.t. the L^2-inner product on Γ_u implies

$$\int_S w\, \psi(u)f\, dS = \int_S \omega f\, \psi(u)(w\omega^{-1})\, dS \tag{4.15}$$

(recall that $\omega = d\Gamma_u/dS$) and the operator F from (3.5) gets the form

$$F(u)f = (F_1(u)f, \dots, F_m(u)f), \quad F_i(u)f = a_{ij}\partial_j \psi(u)f/\sqrt{g}. \tag{4.16}$$

We start the investigation of F by discussing a generalized version of (4.14). Note that we have to deal with two technical difficulties here concerning nonsmooth coefficients and uniformity of the estimates for “large” subsets of U_s. Therefore, we need the following preparation.

Lemma 4.2. *Let $\Omega \subset \mathbb{R}^m$ be a bounded smooth domain, let $x_0 \in \Omega$, $\chi \in C_0^\infty(\mathbb{R}^m)$, and let*

$$\chi_\varepsilon(x) := \chi((x - x_0)/\varepsilon), \qquad x \in \mathbb{R}^m.$$

Moreover, let $\mu \in H^s(\Omega)$ with $s > m/2$ and $\mu(x_0) = 0$. Then there is an $s_1 \in (m/2, s)$ such that

$$\lim_{\varepsilon\downarrow 0} \|\chi_\varepsilon \mu\|^\Omega_{s_1} = 0.$$

Proof. Note at first that

$$\|\chi_\varepsilon\|^{\mathbb{R}^m}_s \le C\varepsilon^{m/2-s}.$$

This is immediately clear for integer s; the general case follows by interpolation. By Sobolev's embedding, we have $\mu \in C^\alpha(\overline{\Omega})$ for some $\alpha > 0$, and consequently, due to $\mu(x_0) = 0$,

$$|\mu(x)| \le C\varepsilon^\alpha, \qquad x \in \mathrm{supp}\chi_\varepsilon \cap \Omega.$$

Thus,

$$\|\chi_\varepsilon\mu\|^\Omega_0 \le C\varepsilon^\alpha \|\chi_\varepsilon\|^\Omega_0 \le C\varepsilon^{\alpha+m/2}$$

and

$$\|\chi_\varepsilon\mu\|^\Omega_s \le C\|\chi_\varepsilon\|^\Omega_s \|\mu\|^\Omega_s \le C\varepsilon^{m/2-s}.$$

The assertion follows now from interpolation. ◻

LEMMA 4.3. *Let $s > (m+1)/2$, $s_0 \in ((m+1)/2, s)$ be given.*

For any $v \in U_s$ there is an H^{s_0}-neighborhood V_{s_0} such that the boundary value problem

$$L(u)w = \partial_i h_i \text{ in } E, \quad B(u)w = \omega e + \nu_i h_i \text{ on } S$$

is uniquely solvable for $u \in V_{s_0} \cap H^s(S, \mathbb{R}^m)$, $e \in H^{s-1}(S)$, $h \in H^{s-1/2}(E, \mathbb{R}^m)$. Moreover, we have

$$\|w\|^E_t \le C\big(\|h\|^E_{t-1} + \|e\|^S_{t-3/2}\big) \tag{4.17}$$

for $t \in [1, s+1/2]$ with C independent of h, e, and u varying in H^s-bounded subsets of $V_{s_0} \cap H^s(S, \mathbb{R}^m)$.

Proof. 1. Fix $v \in U_s$ and choose V_{s_0} according to Lemma 4.1. Fix $u \in V_{s_0} \cap H^s(S, \mathbb{R}^m)$. For $t = 1$, the assertions are easily seen from the variational formulation. For $t = s + 1/2$, the assertions follow from the H^s-regularity theory of elliptic boundary value problems (with operators in divergence form). Our coefficients $\sqrt{g}g^{ij}$, however, are only in $H^{s-1/2}(E)$, which is slightly nonstandard. To prove the necessary regularity result in this case, we can proceed as in the proof of Theorem A.14 in [12], replacing the Hölder norms there by Sobolev norms. To control the error terms occurring from the freezing of coefficients, we use the estimate

$$\|\mu_{ij}\partial_j w\|^E_{s-1/2} \le C\big(\|\mu_{ij}\|^E_{s_1}\|w\|^E_{s+1/2} + \|\mu_{ij}\|^E_{s-1/2}\|w\|^E_{s_1+1/2}\big)$$

(and a corresponding one for the boundary term) with s_1 from Lemma 4.2. Recalling that μ_{ij} has a form to which that lemma applies, (4.17) can be established for $t = s + 1/2$ by a usual perturbation argument, with a constant $C = C(u)$. The general case follows by interpolation.

2. To show uniformity w.r.t. $u \in V_{s_0} \cap H^s(S, \mathbb{R}^m)$, we proceed in a similar way: For $t = s + 1/2$, pick $u_1, u_2 \in V_{s_0} \cap H^s(S, \mathbb{R}^m)$, denote the corresponding coefficients by $\sqrt{g_k} g_k^{ij}$, $k = 1, 2$, and estimate

$$\begin{aligned} &\big\|\big(\sqrt{g_1} g_1^{ij} - \sqrt{g_2} g_2^{ij}\big)\partial_j w\big\|_{s-1/2}^E \\ &\qquad \le C\big(\|\widetilde{u}_1 - \widetilde{u}_2\|_{s_0+1/2}^E \|w\|_{s+1/2}^E + \|\widetilde{u}_1 - \widetilde{u}_2\|_{s+1/2}^E \|w\|_{s_0+1/2}^E\big) \\ &\qquad \le C\big(\|u_1 - u_2\|_{s_0}^S \|w\|_s^E + \|w\|_1^E\big), \end{aligned}$$

where an interpolation inequality has been used. A similar estimate can be given for the boundary term. After shrinking V_{s_0} if necessary, one can show the uniformity by another perturbation argument. ∎

Under the assumptions of Lemma 4.3, as a first trivial consequence we obtain the estimate

$$\|\psi(u) f\|_t^E, \|\psi(u) f\|_{t-1/2}^S \le C\|f\|_{t-3/2} \tag{4.18}$$

for $t \in [1, s + 1/2]$. Note for later reference that these estimates continue to hold for $t \in [0, s + 1/2]$, provided $s > \max\{m + 1, 5\}/2$. To see this, it is sufficient to show (4.18) for $t = 0$; the general case follows by interpolation again. Fix u, pick $\phi \in L^2(S)$ arbitrary and define $w \in H^{3/2}(E)$ by

$$L(u)w = 0 \text{ in } E, \quad B(u)w = \phi \text{ on } S.$$

Then, by Green's formula rewritten in the form

$$\int_E \big(\phi_1 L(u)\phi_2 \, dx - \phi_2 L(u)\phi_1\big)\, dx = \int_S \big(\phi_1 B(u)\phi_2 - \phi_2 B(u)\phi_1\big)\, dS$$

and (4.17) with $t = 2$,

$$\begin{aligned} \int_S \phi\psi(u) f \, dS = \int_S B(u) w \psi(u) f \, dS &= \int_S w\omega f \, dS \\ &\le C\|w\|_{3/2}\|f\|_{-3/2} \le C\|\phi\|_{1/2}\|f\|_{-3/2}. \end{aligned}$$

This proves the second estimate in (4.18). Analogously, picking $\zeta \in L^2(E)$ and defining $v \in H^2(E)$ by

$$L(u)v = \zeta \text{ in } E, \quad B(u)w = 0 \text{ on } S,$$

we get

$$\int_e \zeta\psi(u) f \, dx = \int_E L(u) w \psi(u) f \, dx = -\int_S v\omega_f \, dS \le C\|\zeta\|_0 \|f\|_{-3/2}.$$

This proves the first estimate in (4.18).

Furthermore, concerning the smooth dependence of $\psi(u)f$ on u, Lemma 4.3 together with (4.11), (4.13) implies via a perturbation argument

$$[u \mapsto \psi(u)] \in C^\infty\big(V_s, \mathcal{L}\big(H^{t-3/2}(S), H^t(E)\big)\big). \tag{4.19}$$

Replacing t by $t - 3/2$ and considering (4.16), leads to the following corollary.

Corollary 4.4. *Let $s > (m+1)/2$ and $-1/2 \le t \le s-1$. Then*

$$[u \mapsto F(u)] \in C^\infty\big(U_s, \mathcal{L}\big(H^t(S), H^t(S,\mathbb{R}^m)\big)\big).$$

Recall that our ultimate goal is to prove the energy estimate (1.3). Since we will translate spatial derivatives into Fréchet derivatives later, higher order Fréchet derivatives will have to be estimated. The operator $(u,f) \mapsto F(u)f$ is of order one w.r.t. u and of order zero w.r.t. f; note that the estimates (4.20) are standard for *local* operators of this type, e.g.,

$$(u,f) \mapsto [x \mapsto \Psi(\nabla u(x))f(x)].$$

The nonstandard aspect here is that F involves the solution of an elliptic boundary value problem, and therefore the same is true for its Fréchet derivatives.

Lemma 4.5. *Let $s > (m+1)/2$, $u \in U_s$, and $t \in [1,s]$ be given. Then for any choice of $s_1, \dots s_{k+1} \in [t,s]$ with $s_1 + \dots + s_{k+1} \ge t + ks$ there exists a constant $C > 0$ such that for all $f \in H^{s-1}(S)$ and all $u_1, \dots, u_k \in H^s(S,\mathbb{R}^m)$ the following holds:*

$$\big\|F^{(k)}(u)\{u_1,\dots,u_k\}f\big\|_{t-1} \le C\|u_1\|_{s_1}\cdots\|u_k\|_{s_k}\|f\|_{s_{k+1}-1}. \tag{4.20}$$

The constant may be chosen independently of u as long as u varies in H^s-bounded subsets of U_s which are H^{s_0}-closed for some $s_0 \in ((m+1)/2, s)$.

Proof. 1. Fix $v \in U_s$ and a neighborhood V_{s_0} according to Lemma 4.3. We show (4.20) with C independent of $u \in V_{s_0} \cap H^s(S,\mathbb{R}^m)$. To begin with, recall the estimate (4.18) in the form

$$\|\nabla\psi(u)f\|^S_{t-1},\ \|\psi(u)f\|^E_{t+1/2} \le C\|f\|_{t-1} \tag{4.21}$$

for $t \in [1,s]$; concerning the estimate of $\nabla\psi$ along S in the limit case $t=1$, note that the boundary condition allows a representation of $\nabla\psi$ as a suitable linear combination of ψ, f, and tangential derivatives of ψ. In view of (4.16) this implies the asserted estimate (4.20) for the simplest case $k=0$. To obtain similar estimates for the Fréchet derivatives $\psi^{(k)} = \psi^{(k)}(u)\{u_1,\dots,u_k\}f$, $k = 1,2,\dots$, we have to examine the corresponding derivatives of the coefficients in the transformed Laplacian and the boundary terms. For $\psi^{(k)}$ we get

$$\begin{aligned} L(u)\psi^{(k)} &= -\sum L^{(l)}(u)\{u_{i_1},\dots,u_{i_j}\}\psi^{(k-l)}(u)\{u_{i_{j+1}},\dots,u_{i_k}\}f \ \text{ in } E,\\ B(u)\psi^{(k)} &= -\sum B^{(l)}(u)\{u_{i_1},\dots,u_{i_j}\}\psi^{(k-l)}(u)\{u_{i_{j+1}},\dots,u_{i_k}\}f\\ &\qquad + \omega^{(k)}\{u_1,\dots,u_k\}f \ \text{ on } S, \end{aligned} \tag{4.22}$$

where

$$\begin{aligned} L^{(l)}(u)\{u_{i_1},\dots,u_{i_j}\}\varphi &= \partial_i\big((\sqrt{g}g^{ij})^{(k)}\{u_{i_1},\dots,u_{i_j}\}\partial_j\varphi\big),\\ B^{(l)}(u)\{u_{i_1},\dots,u_{i_j}\}\varphi &= \nu_i(\sqrt{g}g^{ij})^{(k)}\{u_{i_1},\dots,u_{i_j}\}\partial_j\varphi + \omega^{(l)}\{u_{i_1},\dots,u_{i_j}\}\varphi, \end{aligned}$$

and the sums are extended over $1 \le l \le k$ and all decompositions $i_1 < \dots < i_l$ and $i_{j+1} < \dots < i_k$ of the indices $1,2,\dots,n$. In particular, if $k=1$, we obtain for $\psi' = \psi'(u)\{u_1\}f$ the boundary value problem

$$\begin{aligned} L(u)\psi' &= -\partial_i\big((\sqrt{g}g^{ij})'\{u_1\}\partial_j\psi\big) \ \text{ in } E,\\ B(u)\psi' &= -\nu_i\big(\sqrt{g}g^{ij}\big)'\{u_1\}\partial_j\psi + \omega'\{u_1\}(f-\psi) \ \text{ on } S. \end{aligned} \tag{4.23}$$

Thus, for any $t \in [1, s]$, Lemma 4.3 implies

$$\|\psi'(u)\{u_1\}f\|^E_{t+1/2} \le C\big(\|(\sqrt{g}g^{ij})'\{u_1\}\partial_j\psi\|^E_{t-1/2} + \|\omega'\{u_1\}(f-\psi)\|^S_{t-1}\big).$$

To estimate the terms on the right-hand side, by (4.5) we obtain

$$\|(\sqrt{g}g^{ij})'\{u_1\}\partial_j\psi\|^E_{t-1/2} \le C\|(\sqrt{g}g^{ij})'\{u_1\}\|^E_{s_1-1/2}\|\partial_j\psi\|^E_{s_2-1/2},$$

and accordingly

$$\|\omega'\{u_1\}(f-\psi)\|^S_{t-1} \le C\|\omega'\{u_1\}\|^S_{s_1-1}\big(\|f\|_{s_2-1} + \|\psi\|^S_{s_2-1}\big)$$

for any choice of $s_1, s_2 \in [t, s]$ with $s_1 + s_2 \ge t + s$. As

$$\|\psi\|^S_{s_2-1}, \|\partial_j\psi\|^E_{s_2-1/2} \le C\|\psi\|^E_{s_2+1/2} \le C'\|f\|_{s_2-1},$$

by (4.21), using (4.12), (4.13), we find

$$\big\|(\sqrt{g}g^{ij})'\{u_1\}\partial_j\psi\big\|^E_{t-1/2}, \big\|\omega'\{u_1\}(f-\psi)\big\|^S_{t-1} \le C'\|u_1\|_{s_1}\|f\|_{s_2-1},$$

and hence

$$\|\nabla\psi'(u)\{u_1\}f\|^S_{t-1},\ \|\psi'(u)\{u_1\}f\|^E_{t+1/2} \le C\|u_1\|_{s_1}\|f\|_{s_2-1},$$

where the same remark applies to the estimate of $\nabla\psi'$ along S as to (4.21). Using (4.22), these estimates are extended inductively to $\psi^{(k)}$:

$$\|\nabla\psi^{(k)}\|^S_{t-1},\ \|\psi^{(k)}\|^E_{t+1/2} \le C\|u_1\|_{s_1}\cdots\|u_k\|_{s_k}\|f\|_{s_{k+1}-1}, \tag{4.24}$$

provided $t \in [1, s]$, $s_1, \dots, s_{k+1} \in [t, s]$ with $s_1 + \cdots + s_{k+1} \ge t + ks$. In view of (4.16), these estimates together with (4.11) and (4.5) imply the asserted estimate (4.20).

2. Let $K \subset U_s$ be H^{s_0}-closed and bounded in $H^s(S, \mathbb{R}^m)$. As shown in part 1 of this proof, K can be covered by H^{s_0}-open sets $V_{s_0,v}$, $v \in K$, such that (4.20) holds uniformly for $u \in V_{s_0,v} \cap K$. Now the assertion follows from the compactness of K in $H^{s_0}(S, \mathbb{R}^m)$. □

Now we use invariance properties w.r.t. diffeomorphisms (cf., e.g., [10]). Let $\tau \in \mathrm{Diff}(S)$. Then by definition

$$\varphi(u, f) = \varphi(u \circ \tau, f \circ \tau) \text{ in } \Omega_u. \tag{4.25}$$

Recalling the definition of F, we have

$$\begin{aligned}(F(u)f) \circ \tau &= \big(\nabla\varphi(u, f)\big) \circ (u \circ \tau),\\ F(u \circ \tau)(f \circ \tau) &= \big(\nabla\varphi(u \circ \tau, f \circ \tau)\big) \circ (u \circ \tau)p;\end{aligned}$$

consequently (4.25) implies

$$(F(u)f) \circ \tau = F(u \circ \tau)(f \circ \tau) \text{ on } S. \tag{4.26}$$

Any smooth vector field D on S, identified with a first order differential operator, generates a one-parameter group of smooth diffeomorphisms $t \mapsto \tau_t$ with $\tau_t = id$ for $t = 0$. Setting $\tau = \tau_t$ in (4.26) and differentiating w.r.t. t at $t = 0$ gives

$$DF(u)f = F'(u)\{Du\}f + F(u)Df \tag{4.27}$$

for $u \in U_s$ and $f \in H^{s-1}(S)$, $s > (m+1)/2$. Furthermore, by differentiation w.r.t. u,

$$\begin{aligned} DF^{(k)}(u)\{\dots\}f &= F^{(k+1)}(u)\{Du,\dots\}f + F^{(k)}(u)\{\dots\}Df \\ &\quad + \sum_{j=1}^{k} F^{(k)}(u)\{u_1,\dots,u_{j-1},Du_j,u_{j+1},\dots,u_k\}f, \end{aligned} \tag{4.28}$$

where the dots indicate the arguments $u_1,\dots,u_k \in H^s(S,\mathbb{R}^m)$. We choose m smooth vector fields $D_1,\dots,D_m$ on S such that

$$\operatorname{span}\{D_1,\dots,D_m\} = T_x \text{ for all } x \in S$$

and use the multi-index notation $D^\alpha = D_1^{\alpha_1}\dots D_m^{\alpha_m}$, $\alpha = (\alpha_1,\dots,\alpha_m)$, for higher order derivatives. Note that, for $s \ge 0$ integer, we can use

$$(u,v)_s = \sum_{|\alpha|\le s} (D^\alpha u, D^\alpha v)_{L^2(S)}$$

as the inner product generating the norm in $H^s(S)$. As a consequence of (4.27), (4.28), by induction we obtain a differentiation rule which resembles Leibniz's rule at an abstract level: For any multi-index α and $u \in U_s$, $f \in H^{s-1}(S)$, $s > |\alpha|+(m+1)/2$, it holds that

$$D^\alpha F(u)f = \sum c_{\beta_1,\dots,\beta_{k+1}} F^{(k)}(u)\{D^{\beta_1}u,\dots,D^{\beta_k}u\}D^{\beta_{k+1}}f, \tag{4.29}$$

where the sum has to be extended over all integers k and systems of nonnegative multi-indices $\beta_1,\dots,\beta_{k+1}$ with

$$0 \le k \le |\alpha|, \quad 1 \le |\beta_1|,\dots,|\beta_k|, \quad \beta_1+\dots+\beta_{k+1} = \alpha. \tag{4.30}$$

The coefficients are nonnegative integers, in particular, $c_\alpha = c_{\alpha,0} = 1$.

The differentiation rule (4.29) and Lemma 4.5 enable us to prove estimates involving spatial derivatives of F. Concerning the second part of the following proposition, note that (4.31) provides a splitting of $D^\alpha F(u)f$ according to (4.29), with $R_\alpha(u)f$ containing the lower order terms, i.e., the terms involving spatial derivatives up to order $|\alpha|-1$ only. Again, the results (as well as the techniques used in the proof) are standard for local operators of corresponding types.

PROPOSITION 4.6.

(i) *Let $s \ge s_0 > (m+1)/2$ integer, $u \in U_{s+1}$. Then*

$$\|F(u)f\|_s \le C\big(\|u\|_{s+1}\|f\|_{s_0} + \|f\|_s\big).$$

(ii) *Assume additionally $s \ge s_0+1$ and let α be any multi-index with $|\alpha| = s$. Then we have*

$$D^\alpha F(u)f = F(u)D^\alpha f + F'(u)\{D^\alpha u\}f + R_\alpha(u)f, \tag{4.31}$$

where the remainder term allows the estimate

$$\|R_\alpha(u)f\|_0 \le C\big(\|u\|_s\|f\|_{s_0} + \|f\|_{s-1}\big).$$

The constants in both estimates can be chosen uniformly as u varies in H^{s_0}-closed, H^{s_0+1}-bounded subsets of U_{s+1}.

Proof. We consider the more complicated situation (ii) only. According to (4.29), the remainder term has a representation as a sum of terms

$$I_\beta = F^{(k)}(u)\{D^{\beta_1}u, \dots, D^{\beta_k}u\}D^{\beta_{k+1}}f,$$

where the multi-indices satisfy (4.30) and additionally $|\beta_i| < s = |\alpha|$. Thus $k \geq 1$ and $|\beta_i| \geq 1$ for at least two indices, say $i = i_1, i_2$. We estimate I_β using (4.20) with $s_i = 1 + (1-\theta_i)(s_0-1)$ and

$$\theta_i = (|\beta_i| - 1)/(s-2), \quad i = i_1, i_2, \qquad \theta_i = |\beta_i|/(s-2), \quad i \notin \{i_1, i_2\}.$$

Then $s_i \in [1, s_0]$ and $s_1 + \cdots + s_{k+1} = 1 + ks_0$; hence applying (4.20) (with $t = 1$, $s = s_0$) yields

$$\|I_\beta\|_0 \leq C\|u\|_{|\beta_1|+s_1} \dots \|u\|_{|\beta_1|+s_k}\|f\|_{|\beta_{k+1}|+s_{k+1}-1}.$$

Note that $\theta_1 + \cdots + \theta_{k+1} = 1$ and set $\lambda := \theta_1 + \cdots + \theta_k$. From

$$|\beta_i| + s_i \leq (1-\theta_i)(s_0+1) + \theta_i s$$

we get by norm convexity and Young's inequality

$$\begin{aligned}\|I_\beta\|_0 &\leq C\|u\|_{s_0+1}^{k-1}\big(\|u\|_{s_0+1}\|f\|_{s-1}\big)^{1-\lambda}\big(\|u\|_s\|f\|_{s_0}\big)^{\lambda} \\ &\leq C\|u\|_{s_0+1}^{k-1}\big(\|u\|_{s_0+1}\|f\|_{s-1} + \|u\|_s\|f\|_{s_0}\big).\end{aligned}$$

This proves the assertion. ∎

We conclude this section with remarks concerning the Neumann–Dirichlet operator A defined by (3.9), (3.10). It is obvious that the regularity properties of $u \mapsto A(u)$ are the same as for $u \mapsto \psi(u)|_S$; hence (4.19) reappears as

$$[u \mapsto A(u)] \in C^\infty\big(U_s, \mathcal{L}\big(H^t(S), H^{t+1}(S)\big)\big) \tag{4.32}$$

for $s > (m+1)/2$ and $-1/2 \leq t \leq s-1$. Moreover, the differentiation rule (4.28) also holds for A, and $\psi_N(u)f := \varphi_N(u)f \circ \tilde{u}$ satisfies estimates parallel to (4.24). Hence we get

$$\|A^{(k)}(u)\{u_1, \dots, u_k\}f\|_t \leq C\|u_1\|_{s_1} \cdots \|u_k\|_{s_k}\|f\|_{s_{k+1}-1}, \tag{4.33}$$

provided $s_1, \dots, s_{k+1} \in [t, s]$ with $s_1 + \cdots + s_{k+1} \geq t + ks$. Thus we have the following analogue to Proposition 4.6.

Proposition 4.7.

(i) *Let $s \geq s_0 > (m+1)/2$ integer, $u \in U_s$, and $f \in H^{s-1}(S)$. Then*

$$\|A(u)f\|_s \leq C\big(\|u\|_s\|f\|_{s_0-1} + \|f\|_{s-1}\big)$$

with a uniform constant as long as u varies in H^{s_0}-closed, H^{s_0}-bounded subsets of U_s.

(ii) *Assume additionally $s \geq s_0+1$, and let α be any multi-index with $|\alpha| = s-1$. Then we have*

$$\|D^\alpha A(u)f - A(u)D^\alpha f\|_1 \leq C\big(\|u\|_s\|f\|_{s_0-1} + \|f\|_{s-2}\big),$$

where now the constant can be chosen uniformly as u varies in H^{s_0}-closed, H^{s_0+1}-bounded subsets of U_{s+1}.

Finally, for later reference we point out the simple commutator estimate

$$\|A(u)(hf) - hA(u)f\|_1 \leq C\|h\|_s\|f\|_{-1} \tag{4.34}$$

for $u \in U_s,\ f, h \in H^s(S),\ s > (m+1)/2$. Fixing any neighborhood V_s according to Lemma 4.1, fixing $u \in V_s$ with corresponding diffeomorphism (4.10), and considering

$$A(u)f = \psi(u)\big(f - \omega A(u)f\big)|_S$$

reduces (4.34) to

$$\|h\psi(u)f - \psi(u)(hf)\|_1^S \leq C\|h\|_s\|f\|_{-1}. \tag{4.35}$$

Let $\tilde h$ be the extension of h into E determined by solving the Dirichlet problem

$$L(u)\tilde h = 0 \ \text{ in } \ E, \quad \tilde h = h \ \text{ on } \ S.$$

Clearly $\|\tilde h\|_{s+1/2}^E \leq C\|h\|_s$ and $\tilde\psi := h\psi(u)f - \psi(u)(hf)$ solves the boundary value problem

$$\begin{aligned} L(u)\tilde\psi &= 2\partial_i\big(\sqrt{g}g^{ij}\partial_i\tilde h\,\psi(u)f\big) \ \text{ in } \ E,\\ B(u)\tilde\psi &= -\omega\nu_i g^{ij}\partial_i\tilde h\,\psi(u)f \ \text{ on } \ S. \end{aligned}$$

Hence by Lemma 4.3

$$\|\tilde\psi\|_{3/2}^E \leq C\|\sqrt{g}g^{ij}\partial_i\tilde h\|_{s-1/2}\|\psi(u)f\|_{1/2}^E \leq C\|h\|_s\|\psi(u)f\|_{1/2}^E.$$

Together with (4.18) this implies (4.35).

5. The main estimate. In this section we prove H^s a priori estimates for the nonlinear operator $\mathcal{F}$ w.r.t. variable bilinear forms, which we define in what follows. As already mentioned in the introduction, these estimates are the main ingredient in the existence proof.

To begin with, for $u \in U_s,\ s > (m+1)/2$, we define

$$P(u)v := v \cdot \big(n(u) \circ u\big), \quad N(u)w := w\,\big(n(u) \circ u\big), \tag{5.1}$$

$$\Lambda(u)w := \nabla_{\Gamma_u}(w \circ u^{-1}) \circ u \tag{5.2}$$

as the Euclidean inner product and multiplication with outer normal $n(u)$ of Γ_u and pullback of tangential gradient ∇_{Γ_u} along Γ_u, respectively. If $P(u)$, $N(u)$, and $\Lambda(u)$ are considered as operators in v and w, their coefficients are smooth functions of u and its first derivatives. Thus, using (4.1)–(4.5),

$$P(u) \in \mathcal{L}\big(H^t(S,\mathbb{R}^m), H^t(S)\big), \quad N(u) \in \mathcal{L}\big(H^t(S), H^t(S,\mathbb{R}^m)\big), \tag{5.3}$$

$$\Lambda(u) \in \mathcal{L}\big(H^t(S), H^{t-1}(S,\mathbb{R}^m)\big) \tag{5.4}$$

depend smoothly on $u \in U_s$ for $0 \leq t \leq s-1$ and $1 \leq t \leq s$, respectively. Clearly, the operators P, N, Λ satisfy invariance properties as stated for F in (4.26). As a consequence, the differentiation rule (4.27) and its corollary (4.28) are also true for P, N, Λ; we make use of that without explicit mention. Further, recall that the pullback $\Delta(u)w$ of the Laplace–Beltrami operator Δ_{Γ_u} on Γ_u according to (3.8) and the operator $H(u)$ according to (3.7) may be expressed as

$$\Delta(u)w = \Lambda_i(u)\big(\Lambda_i(u)w\big), \quad H(u) = -\gamma\Lambda_i(u)\big(n_i(u) \circ u\big),$$

respectively (see, e.g., [9, sect. 15.1]). Thus, recalling (4.32), the operator $\mathcal{G}$ defined by (3.4), (3.7) satisfies

$$[u \mapsto \mathcal{G}(u)] \in C^\infty\big(U_s, H^{s-2}(S)\big),$$

provided $s > (m+3)/2$. Together with Corollary 4.4 this implies the smoothness of $\mathcal{F}$ as stated in (3.12).

In further considerations of this section we fix the integer $s_0 := [(m+5)/2] + 1$ and set

$$\widetilde{U}_s := U_s \cap K \ \text{ for all } \ s \geq s_0$$

with an H^{s_0}-bounded and L^2-closed subset $K \subseteq U_{s_0}$. Note that

$$1 \leq C\|u\|_{s_0} \leq C'\|u\|_s, \quad \|u\|_{C^3(S)} \leq C$$

for all $u \in \widetilde{U}_s$, $s \geq s_0$. By transforming the well-known integration by parts formula for the differential operator ∇_{Γ_u} onto S, we get the form

$$\int_S \omega \Lambda_i(u) f \, dS = -\int_S \omega (\kappa_{\Gamma_u} \circ u)(n_i(u) \circ u) f \, dS.$$

Consequently, for $u \in \widetilde{U}_s$, $s \geq s_0$, and any $f \in C^1(S)$, we have

$$\left| \int_S \Lambda_i(u) f \, dS \right| \leq C \int_S |f| \, dS. \tag{5.5}$$

Furthermore, note the estimates

$$\|\mathcal{G}(u)\|_{s-2}, \|\mathcal{F}(u)\|_{s-2} \leq C\|u\|_s \ \text{ for all } \ u \in \widetilde{U}_s, s \geq s_0.$$

The following lemma is crucial, as it identifies the leading (first) order term in the linearization of $u \mapsto F(u)$ in an explicit way.

Lemma 5.1. *Let $s \geq s_0$. Then for $u \in U_s$, $v \in H^s(S, \mathbb{R}^m)$, and $f \in H^{s-1}(S)$ it holds that*

$$F'(u)\{v\}f = F(u)\big(\Lambda(u)\big(P(u)v\big) \cdot F(u)f\big) + R(u)\{v\}f,$$

where R allows the estimate

$$\|R(u)\{v\}f\|_0 \leq C\|f\|_{s-1}\|v\|_0.$$

The constant is independent of u as long as u varies in $\widetilde{U}_s$.

Proof. As in the proof of Lemma 4.5 we can assume $u \in V_s$. We have

$$F_i'(u)\{v\}f = \partial_i \varphi' \circ u + v_j \partial_i \partial_j \varphi \circ u,$$

where $\varphi' = \varphi'(u,f)\{v\}$ denotes the derivative w.r.t. u of the velocity potential $\varphi = \varphi(u,f)$ in Ω_u. As

$$\|\varphi(u,f)\|_{C^2(\bar{\Omega}_u)} \leq C_1\|\psi(u)f\|_{C^2(\bar{E})} \leq C_2\|\psi(u)f\|_{H^{s+1/2}(E)} \leq C_3\|f\|_{s-1}$$

by Sobolev's embedding and (4.21), we obtain

$$\|v_j \partial_i \partial_j \varphi \circ u\|_0^S \leq C\|f\|_{s-1}\|v\|_0.$$

Furthermore, φ' satisfies $\Delta\varphi' = 0$ in Ω_u and the boundary condition

$$\varphi' + n \cdot \nabla\varphi' + n' \cdot \nabla\varphi + (\partial_i\varphi + n_j\partial_i\partial_j\varphi)v_i \circ u^{-1} = 0 \text{ on } \Gamma_u,$$

where we have used the abbreviation

$$n' = \partial_\varepsilon\big(n(u+\varepsilon v) \circ (id + \varepsilon v \circ u^{-1})\big)|_{\varepsilon=0}$$

for the variation of the outer normal on Γ_u. A simple calculation (cf. Lemma 1.1 in [3]) shows

$$n' = -\nabla_{\Gamma_u}\big(n \cdot v \circ u^{-1}\big) + v_i \circ u^{-1}\nabla_{\Gamma_u} n_i. \tag{5.6}$$

By retransformation onto the reference domain E, for $\tilde{\psi}' = \varphi' \circ u$ to satisfy the boundary value problem, this implies

$$L(u)\tilde{\psi}' = 0 \text{ in } E, \quad B(u)\tilde{\psi}' = \Lambda(u)\big(P(u)v\big) \cdot F(u)f + R_1(u)\{v\}f \text{ on } S.$$

The operator R_1 acts by pointwise multiplications w.r.t. the components of v, and hence by the same reasoning as above we get the estimate

$$\|R_1(u)\{v\}f\|_0^S \le C\|f\|_{s-1}\|v\|_0.$$

Thus the result follows. $\square$

For $u \in U_s$ let $M(u) \in \mathcal{L}\big(L^2(S,\mathbb{R}^m)\big)$ be the operator defined by

$$M(u)v := v - \Lambda(u)\big(\psi(u)P(u)v\big). \tag{5.7}$$

By (4.19) and (5.4),

$$M(u) \in \mathcal{L}\big(H^t(S,\mathbb{R}^m), H^t(S,\mathbb{R}^m)\big), \quad 0 \le t \le s-1,$$

depends smoothly on $u \in U_s$, $s > (m+1)/2$; for later reference we state explicitly

$$\|M^{(k)}(u)\{u_1,\dots,u_k\}v\|_t \le C\|u_1\|_s \cdots \|u_k\|_s\|v\|_t. \tag{5.8}$$

Because of $P(u)\Lambda(u) = 0$ the operator $M(u)$ constitutes an isomorphism in $L^2(S,\mathbb{R}^m)$ with inverse

$$M(u)^{-1}v = v + \Lambda(u)\big(\psi(u)P(u)v\big), \tag{5.9}$$

and we have

$$C^{-1}\|v\|_0 \le \|M(u)v\|_0 \le C\|v\|_0 \tag{5.10}$$

for all $v \in L^2(S,\mathbb{R}^m)$ with a positive constant C independent of $u \in \widetilde{U}_s$.

The operator M will be used for the definition of our variable inner products; see (5.12) below. The technique used here is comparable to the symmetrization of hyperbolic systems. The following lemma exhibits the crucial property on which the choice of M is based. Note that it relates, up to lower order terms, an inner product for vector-valued functions to an inner product for scalar-valued ones.

Lemma 5.2. *Let $s > (m+3)/2$. There exists a positive constant C such that for all $u \in \widetilde{U}_s$ and $f \in L^2(S)$, $w \in L^2(S,\mathbb{R}^m)$*

$$\big|\big(M(u)F(u)f, M(u)w\big)_0 - \big(f, P(u)w\big)_0\big| \le C\|f\|_{-1}\|w\|_0. \tag{5.11}$$

Proof. Reformulating the boundary condition satisfied by $\psi(u)f$, we have

$$P(u)\big(F(u)f\big) = f - \psi(u)f, \quad F(u)f - \Lambda(u)\big(\psi(u)f\big) = N(u)\big(f - \psi(u)f\big),$$

and consequently

$$\begin{aligned} M(u)F(u)f &= F(u)f - \Lambda(u)\psi(u)\big(f - \psi(u)f\big) \\ &= N(u)\big(f - \psi(u)f\big) + \Lambda(u)\psi(u)^2 f. \end{aligned}$$

Further, recalling $P(u)\Lambda(u) = 0$,

$$\big(N(u)f, M(u)w\big)_0 = \big(f, P(u)M(u)w\big)_0 = \big(f, P(u)w\big)_0,$$

and we obtain

$$\big(M(u)F(u)f, M(u)w\big)_0 = \big(f - \psi(u)f, P(u)w\big)_0 + \big(\Lambda(u)\psi(u)^2 f, M(u)w\big)_0.$$

Together with

$$\|\Lambda(u)\psi(u)^2 f\|_0^S, \|\psi(u)f\|_0^S \le C\|f\|_{-1}$$

from (4.18), this immediately implies (5.11). ∎

In view of (5.10), for every fixed $u \in U_s$, $s \ge s_0$,

$$(v, w)_{s,u} := \big(M(u)v, M(v)w\big)_0 + \sum_{|\alpha|=s} \big(M(u)D^\alpha v, M(u)D^\alpha w\big)_0 \tag{5.12}$$

defines a inner product on $H^s(S, \mathbb{R}^m)$, which is equivalent to the usual one. This inner product (and corresponding bilinear forms) will be used when we apply the abstract existence theorem, Theorem 3.4, to our evolution problem. The next two lemmas provide the properties necessary for this.

LEMMA 5.3. *Let $s \ge s_0$ and $u \in \widetilde{U}_s$.*

(i) *There exists a $C > 0$ independent of u such that for all $v \in H^{s+2}(S, \mathbb{R}^m)$ and $w \in H^s(S, \mathbb{R}^m)$*

$$(v, w)_{s,u} \le C\|v\|_{s+2}\|w\|_{s-2}. \tag{5.13}$$

(ii) *There exist $\lambda_0, c_0 > 0$ independent of u such that for all $v \in H^{s+4}(S, \mathbb{R}^m)$ and $\lambda \ge \lambda_0$*

$$\big(v, (\Delta_0^2 + \lambda)v\big)_{s,u} \ge c_0\|v\|_{s+2}^2 \tag{5.14}$$

with the elliptic operator $\Delta_0 := D_i D_i$ on S.

Proof. (i) We consider a typical term of (5.12) and show

$$I_\alpha(v, w) := \big(M(u)D^\alpha v, M(u)D^\alpha w\big)_0 \le C\|v\|_{s+2}\|w\|_{s-2} \tag{5.15}$$

for smooth v, w and multi-indices α with $|\alpha| = s$. Recalling (5.8), we have

$$\|M^{(k)}(u)\{D^{\alpha_1}u, \dots, D^{\alpha_k}u\}w\|_0 \le C\|w\|_0$$

if $|\alpha_1|, \dots, |\alpha_k| \le 2$. Thus, writing $D^\alpha w = D^\beta D^\delta w$ with $|\beta| = 2$ and $|\delta| = s - 2$, multiple application of the differentiation rule gives a representation

$$M(u)D^\alpha w = \sum (-1)^{|\alpha_0|} D^{\alpha_0} M^{(k)}(u)\{D^{\alpha_1}u, \dots, D^{\alpha_k}u\} D^\delta w, \tag{5.16}$$

where $|\alpha_i| \le 2$ (in fact $\alpha_0 + \cdots + \alpha_k = \beta$); hence

$$\|M^{(k)}(u)\{D^{\alpha_1}u, \dots, D^{\alpha_k}u\}D^\delta w\|_0 \le C\|w\|_{s-2}.$$

Furthermore, using the differentiation rule again, we have

$$\|M(u)D^\alpha v\|_2 \le C\|v\|_{s+2},$$

and consequently

$$\big(M(u)D^\alpha w, D^{\alpha_0}M^{(k)}(u)\{D^{\alpha_1}u, \dots, D^{\alpha_k}u\}D^\delta v\big)_0 \le C\|v\|_{s+2}\|w\|_{s-2}.$$

This implies (5.15).

(ii) Using the same type of argument as in the proof of part (i), we obtain

$$\big(v, \Delta_0^2 v\big)_{s,u} \ge \big(D_iD_jv, D_iD_jv\big)_{s,u} - C\|v\|_{s+1}\|v\|_{s+2},$$

and consequently

$$\big(v, (\Delta_0^2 + \lambda)v\big)_{s,u} \ge c_0\big(\|v\|_{s+2}^2 + \lambda\|v\|_s^2\big) - C\|v\|_{s+1}^2$$

with a positive constant c_0. Hence applying

$$\|v\|_{s+1}^2 \le \sigma\|v\|_{s+2}^2 + C(\sigma)\|v\|_s^2$$

with $\sigma = c_0/2$ and choosing λ sufficiently large, we get the claim. ∎

An immediate consequence of Lemma 5.3(i) is the existence of a continuous bilinear form $\langle\cdot,\cdot\rangle_{s,u}$ on $H^{s+2}(S,\mathbb{R}^m) \times H^{s-2}(S,\mathbb{R}^m)$ compatible with $(\cdot,\cdot)_{s,u}$; i.e., it holds that $\langle v,w\rangle_{s,u} = (v,w)_{s,u}$ for all $v, w \in H^{s+2}(S,\mathbb{R}^m)$. Further, we put for $\varepsilon \in (0,1]$

$$\langle v,w\rangle_{s,u}^\varepsilon := \langle v,w\rangle_{s_0,u} + \varepsilon^2\langle v,w\rangle_{s,u}. \tag{5.17}$$

LEMMA 5.4. *We assume as above that $s \ge s_0$, $\varepsilon \in (0,1]$.*

(i) *For fixed $u \in U_s$, the mapping $\langle\cdot,\cdot\rangle_{s,u}^\varepsilon : H^{s+2}(S,\mathbb{R}^m) \times H^{s-2}(S,\mathbb{R}^m) \to \mathbb{R}$ constitutes a continuous, nondegenerate bilinear form, symmetric on $H^{s+2}(S,\mathbb{R}^m) \times H^{s+2}(S,\mathbb{R}^m)$.*

(ii) *With constants $C > 0$ independent of ε, u, v, w, one has for $u, w \in \tilde{U}_s$ and $v \in H^{s+2}(S,\mathbb{R}^m)$*

$$C\big(\|v\|_{s_0}^2 + \varepsilon^2\|v\|_s^2\big) \le \langle v,v\rangle_{s,u}^\varepsilon \le C^{-1}\big(\|v\|_{s_0}^2 + \varepsilon^2\|v\|_s^2\big), \tag{5.18}$$

$$\langle v,v\rangle_{s,u}^\varepsilon \le \langle v,v\rangle_{s,w}^\varepsilon\big(1 + C\|u-w\|_{s_0-2}\big). \tag{5.19}$$

(iii) *The weak convergences $u_n \rightharpoonup u \in H^s$, $w_n \rightharpoonup w \in H^{s-2}$ imply for each $v \in H^{s+2}$*

$$\langle v,w_n\rangle_{s,u_n}^\varepsilon \to \langle v,w\rangle_{s,u}^\varepsilon.$$

Proof. (i) It remains only to show nondegeneracy. First note that Lemma 5.3(ii) implies for every $v \in H^{s+2}$ and $\lambda \ge \lambda_0$

$$\big\langle v, \Delta_0^2 v + \lambda v\big\rangle_{s,u}^\varepsilon \ge c_0\|v\|_0^2.$$

Let ε, u, w be fixed such that $\langle v,w\rangle_{s,u}^\varepsilon = 0$ for every $v \in H^{s+2}$. Let λ be sufficiently large and let $v \in H^{s+2}$ be the unique solution of the fourth order elliptic equation

$$\Delta_0^2 v + \lambda v = w \ \text{ on } \ S.$$

Thus we have

$$0 = \big\langle v, w\big\rangle_{s,u}^{\varepsilon} = \big\langle v, \Delta_0^2 v + \lambda v\big\rangle_{s,u}^{\varepsilon} \geq c_0 \|v\|_0^2$$

for our special v; consequently it follows that $v = 0$ and then $w = 0$.

(ii) The estimates (5.18) are immediate consequences of (5.10). Concerning (5.19) we note only that by (5.8)

$$\|M(u)f - M(w)f\|_0 \leq C\|u - w\|_{s_0-2}\|f\|_0,$$

from which the assertion can easily be derived.

(iii) Fix $v \in H^{s+2}$, $u \in U_s$, and, for the time being, $w \in H^s$. Using the representation (5.14), we get for $|\alpha| = s$

$$\begin{aligned}&\big(M(u)D^\alpha v, M(u)D^\alpha w\big)_0 \\ &\qquad = \sum (-1)^{|\alpha_0|}\big(M(u)D^\alpha v, D^{\alpha_0}M^{(k)}(u)\{D^{\alpha_1}u, \dots, D^{\alpha_k}u\}D^\delta w\big)_0\end{aligned}$$

with $|\alpha_i| \leq 2$, $|\delta| = s - 2$. Now assume $u_n \rightharpoonup u$ in U_s; thus $u_n \to u$ in $H^{s'}$ with $s \in [0, s)$, and $w_n \rightharpoonup w$ in H^{s-2}. According to the above remark, $\big\langle v, w_n\big\rangle_{s,u_n}$ can essentially be represented as a sum of terms

$$\sum (-1)^{|\alpha_0|}\big\langle M(u_n)D^\alpha v, D^{\alpha_0}M^{(k)}(u_n)\{D^{\alpha_1}u_n, \dots, D^{\alpha_k}u_n\}D^\delta w_n\big\rangle_{H^2\times H^{-2}}$$

with α, α_i, and δ as above, where $\langle\cdot,\cdot\rangle_{H^2\times H^{-2}}$ denotes the L^2-duality map on $H^2(S, \mathbb{R}^m) \times H^{-2}(S, \mathbb{R}^m)$. From the smoothness properties of M we conclude that

$$M(u_n)D^\alpha v \to M(u)D^\alpha v \ \text{ in } \ H^2(S, \mathbb{R}^m). \tag{5.20}$$

Similarly, uniform boundedness of $\|w_n\|_{s-2}$ and convergence $u_n \to u$ in H^{s_0-2} imply via (5.8)

$$M^{(k)}(u_n)\{D^{\alpha_1}u_n, \dots, D^{\alpha_k}u_n\}D^\delta w_n - M^{(k)}(u)\{D^{\alpha_1}u, \dots, D^{\alpha_k}u\}D^\delta w_n \to 0$$

in L^2. Since strong continuity of linear operators implies weak continuity, we have

$$M^{(k)}(u)\{D^{\alpha_1}u, \dots, D^{\alpha_k}u\}D^\delta w_n \rightharpoonup M^{(k)}(u)\{D^{\alpha_1}u, \dots, D^{\alpha_k}u\}D^\delta w$$

weakly in L^2, and consequently

$$D^{\alpha_0}M^{(k)}(u_n)\{D^{\alpha_1}u_n, \dots, D^{\alpha_k}u_n\}D^\delta w_n \rightharpoonup D^{\alpha_0}M^{(k)}(u)\{D^{\alpha_1}u, \dots, D^{\alpha_k}u\}D^\delta w$$

weakly in H^{-2}. Together with (5.20), this completes the proof. ∎

The following estimates, given in Lemma 5.5 and Proposition 5.6, form the core of the existence proof. The techniques are quite standard (cf. (4.1), (4.5)), but it is crucial to use the structure of $\mathcal{G}$ given by (5.25), which provides "coercivity w.r.t. the normal component"; cf. (5.24). Finally, in the proof of Proposition 5.6 we will couple the results concerning the operators F and $\mathcal{G}$ and use an integration by parts to deal with the possible degeneration of γ.

Lemma 5.5. *Let $s \geq s_0$ be an integer, let $u \in \widetilde{U}_s$, and assume $\gamma = \rho^2$ with $\rho \in C^\infty(S)$.*

(i) *There exist positive constants* c, C*, independent of* u*, such that*

$$\|\mathcal{G}'(u)v\|_0 \le C\big(\|\rho P(u)v\|_2 + \|v\|_1\big), \tag{5.21}$$

$$\big(DP(u)v, D\mathcal{G}'(u)v\big)_0 \le -c\|\rho P(u)Dv\|_1^2 + C\|v\|_1^2 \tag{5.22}$$

for all $v \in H^2(S, \mathbb{R}^m)$ *and any derivative* $D = D^\alpha$, $|\alpha| \le 1$.

(ii) *Moreover, for* $|\alpha| = s$ *we have*

$$\|D^\alpha \mathcal{G}(u)\|_{-1} \le C\big(\|\rho P(u) D^\alpha u\|_1 + \|u\|_s\big), \tag{5.23}$$

$$\big(P(u)D^\alpha u, D^\alpha \mathcal{G}(u)\big)_0 \le -c\|\rho P(u)D^\alpha u\|_1^2 + C\|u\|_s^2. \tag{5.24}$$

Proof. (i) To show (5.21), (5.22) it suffices to construct a representation of the form

$$\mathcal{G}'(u)v = \gamma\Delta(u)(P(u)v) + \rho R_1(u)v + R_2(u)v \tag{5.25}$$

with operators $R_1(u)$, $R_2(u)$ such that

$$\|R_1(u)v\|_0, \|R_2(u)v\|_1 \le C\|v\|_1. \tag{5.26}$$

For the part $H'(u)v$ of $\mathcal{G}'(u)v$, which is a second order differential operator in v, this is quite clear using the well-known fact that the linearization of the mean curvature has $\Delta(u)(P(u)v)$ as its main part. Concerning $G'(u)v$ we note

$$\begin{aligned} -G'(u)v &= A(u)\Delta'(u)\{v\}\gamma + A'(u)\{v\}\Delta(u)\gamma, \\ \Delta'(u)\{v\}\gamma &= 2\rho\Delta'(u)\{v\}\rho + 4\Lambda_i'(u)\{v\}\rho\Lambda_i(u)\rho; \end{aligned}$$

hence we have the representation $-G'(u)v = \rho R_1(u) + R_2(u)$ with

$$\begin{aligned} R_1(u) &:= 2A(u)\Delta'(u)\{v\}\rho, \\ R_2(u) &:= 2\big(A(u)(\rho\Delta'(u)\{v\}\rho) - \rho A(u)\Delta'(u)\{v\}\rho\big) \\ &\quad + 4A(u)\Lambda_i'(u)\{v\}\rho\Lambda_i(u)\rho + A'(u)\{v\}\Delta(u)\gamma. \end{aligned}$$

Due to

$$\|\Delta(u)\rho\|_{s_0-2} \le C, \quad \|\Delta'(u)\{v\}\rho\|_{-1}, \|\Lambda_i'(u)\{v\}\rho\|_0 \le C\|v\|_1,$$

the estimate (5.26) for R_1 is now a consequence of

$$\|A(u)f\|_0 \le C\|f\|_{-1}, \quad \|\Delta'(u)\{v\}\rho\|_{-1} \le C\|v\|_1,$$

whereas the estimate for R_2 follows from the commutator estimate (4.34) together with

$$\|A(u)f\|_1 \le C\|f\|_0, \quad \|A'(u)\{v\}f\|_1 \le C\|v\|_1\|f\|_{s_0-2}.$$

(ii) Similar to part (i), it suffices to show the existence of a decomposition

$$D^\alpha\mathcal{G}(u) = \gamma\Delta(u)(P(u)D^\alpha u) + \rho R_1(u) + R_2(u),$$

with operators R_1, R_2 allowing the estimates

$$\|R_1(u)\|_{-1}, \|R_2(u)\|_0 \le C\|u\|_s.$$

Again, for the part $D^\alpha H$ of $D^\alpha \mathcal{G}$ this is quite clear, where R_1, R_2 are now local differential operators w.r.t. u of order $s+1$ and s, respectively. Concerning $D^\alpha G(u)$ we write $\alpha = \beta + \delta$ with $|\beta| = 1$, $|\delta| = s-1$ and calculate

$$-D^\alpha G(u) = 2\rho D^\beta A(u)\Delta'(u)\{D^\delta u\}\rho + Q_1 + \cdots + Q_5$$

with

$$\begin{aligned}
Q_1 &:= D^\beta\big(D^\delta A(u)\Delta(u)\gamma - A(u)D^\delta\Delta(u)\gamma\big),\\
Q_2 &:= D^\beta A(u)\big(D^\delta\Delta(u)\gamma - \Delta'(u)\{D^\delta u\}\gamma\big),\\
Q_3 &:= 4D^\beta A(u)\Lambda_i'(u)\{D^\delta u\}\rho\Lambda_i(u)\rho,\\
Q_4 &:= 2D^\beta\big(A(u)(\rho\Delta'(u)\{D^\delta u\}\rho) - \rho A(u)\Delta'(u)\{D^\delta u\}\rho\big),\\
Q_5 &:= 2(D^\beta\rho)A(u)\Delta'(u)\{D^\delta u\}\rho.
\end{aligned}$$

Now we set $R_1 := 2D^\beta A(u)\Delta'(u)\{D^\delta u\}\rho$ and $R_2 := Q_1 + \cdots + Q_5$. The necessary estimates follow from the properties of A—in particular, Proposition 4.7 (ii) and (4.34)—and from the additional commutator estimate

$$\|D^\delta\Delta(u)\gamma - \Delta'(u)\{D^\delta u\}\gamma\|_0 \le C\|u\|_s. \qquad \square$$

Now we are prepared to formulate and prove the following a priori estimates for $\mathcal{F}$ w.r.t. the bilinear forms $\langle\cdot,\cdot\rangle_{s,u}$.

PROPOSITION 5.6. *Let $s \ge s_0$ be an integer. Then*

$$\big\langle v, \mathcal{F}'(u)v\big\rangle_{1,u} \le C\|v\|_1^2, \tag{5.27}$$

$$\big\langle u, \mathcal{F}(u)\big\rangle_{s,u} \le C\|u\|_s^2 \tag{5.28}$$

for all $u \in \widetilde{U}_s$ and $v \in H^2(S,\mathbb{R}^m)$ with constants independent of u and v.

Proof. We start with the proof of (5.27). Due to (5.21), for any derivative D

$$\big\|D\mathcal{F}'(u)v - F'(u)\{Dv\}\mathcal{G}(u) - F(u)\big(D\mathcal{G}'(u)v\big)\big\|_0 \le C\big(\|v\|_1 + \|\rho P(u)v\|_2\big),$$

and consequently by Lemma 5.1

$$D\mathcal{F}'(u)v = F(u)\big(\mathcal{F}(u)\cdot\Lambda(u)(P(u)Dv) + D\mathcal{G}'(u)v\big) + R(u)v,$$

where the remainder term satisfies

$$\|R(u)v\|_0 \le C\big(\|v\|_1 + \|\rho P(u)v\|_2\big).$$

Further, by (5.21) we have

$$\|D\mathcal{G}'(u)v\|_{-1} \le C\big(\|v\|_1 + \|\rho P(u)v\|_2\big),$$

and moreover

$$\|\mathcal{F}(u)\cdot\Lambda(u)(P(u)Dv)\|_{-1} \le C\|v\|_1.$$

Hence by Lemma 5.2 it follows that

$$\big\langle Dv, D\mathcal{F}'(u)v\big\rangle_{0,u} \le \big(P(u)Dv, \mathcal{F}(u)\cdot\Lambda(u)(P(u)Dv) + D\mathcal{G}'(u)v\big)_0 + I(u)v^2,$$

where now

$$I(u)v^2 \le C\big(\|v\|_1 + \|\rho P(u)v\|_2\big)\|v\|_1.$$

Writing

$$\begin{aligned}\big(P(u)Dv, \mathcal{F}(u)\cdot\Lambda(u)(P(u)Dv)\big)_0 \\ = \frac{1}{2}\int_S \Big(\Lambda_i(u)\big(\mathcal{F}_i(u)(P(u)Dv)^2\big) - (P(u)Dv)^2\Lambda_i(u)\mathcal{F}_i(u)\Big)\, dS,\end{aligned}$$

an integration by parts on S using (5.5) yields

$$\big|\big(P(u)Dv, \mathcal{F}(u)\cdot\Lambda(u)(P(u)Dv)\big)_0\big| \le C\|v\|_1^2;$$

hence together with Lemma 5.5 and (5.22) we obtain the estimate (5.27).

Further, to prove (5.28) we use the abbreviation

$$\|u\|'_{s+1} := \left(\|u\|_s + \sum_{|\alpha|=s}\|\rho P(u)D^\alpha u\|_1\right).$$

Using Proposition 4.6(ii) we write

$$D^\alpha\mathcal{F}(u) = F(u)D^\alpha\mathcal{G}(u) + F'(u)\{D^\alpha u\}\mathcal{G}(u) + R_1(u),$$

where R_1 allows the estimate

$$\|R_1(u)\|_0 \le C\big(\|u\|_s\|\mathcal{G}(u)\|_{s_0-2} + \|\mathcal{G}(u)\|_{s-1}\big) \le C\|u\|'_{s+1}$$

because of

$$\|\mathcal{G}(u)\|_{s-1} \le C\left(\sum_{|\alpha|=s}\|D^\alpha\mathcal{G}(u)\|_{-1} + \|\mathcal{G}(u)\|_0\right)$$

and (5.23). Further, using Lemma 5.1 we have

$$D^\alpha\mathcal{F}(u) = F(u)\big(D^\alpha\mathcal{G}(u) + \mathcal{F}(u)\cdot\Lambda(u)(P(u)D^\alpha u)\big) + R_1(u) + R_2(u),$$

where again

$$\|R_2(u)\|_0 \le C\|D^\alpha u\|_0 \le C\|u\|_s,$$

and consequently

$$\big\langle R_1(u) + R_2(u), D^\alpha u\big\rangle_{0,u} \le C\|u\|'_{s+1}\|u\|_s.$$

By (5.11) we obtain

$$\big\langle D^\alpha u, D^\alpha\mathcal{F}(u)\big\rangle_{0,u} = \big(P(u)D^\alpha u, \mathcal{F}(u)\cdot\Lambda(u)(P(u)D^\alpha u) + D^\alpha\mathcal{G}(u)\big)_0 + I(u),$$

where $I(u)$ allows the estimate

$$\begin{aligned}I(u) &\le C\|D^\alpha u\|_0\big(\|D^\alpha\mathcal{G}(u)\|_{-1} + \|\mathcal{F}(u)\cdot\Lambda(u)(P(u)D^\alpha u)\|_{-1}\big) \\ &\le C\big(1 + \|u\|'_{s+1}\big)\|u\|_s.\end{aligned}$$

Finally, by an integration by parts as above we get

$$\big(D^\alpha u, \mathcal{F}(u)\cdot\Lambda(u)(P(u)D^\alpha u\big)_0 \le C\|u\|_s^2.$$

Together with (5.24), this completes the proof. ∎

The structure of $F'(u)$ as stated in Lemma 5.1 and the integration by parts argument used in the above proof are necessary to cover the case of a γ which can degenerate. If γ is strictly positive, the argumentation can be simplified by using Lemma 5.1 to obtain the estimate

$$\|F'(u)\{v\}f\|_0 \le C\big(\|P(u)v\|_1 + \|v\|_0\big)\|f\|_{s_0-1}. \tag{5.29}$$

To conclude this section we add some remarks about the case of a slip factor δ (introduced in (2.8)) different from one. The nonlinear operator of the evolution equation is now

$$\mathcal{F}_1(u) := F_1(u)\big(\mathcal{G}_1(u)\big)$$

with

$$F_1(u)f := \big(\delta\, id + (1-\delta)N(u)P(u)\big)F(u)f, \quad \mathcal{G}_1(u) = H(u) + \delta G(u).$$

Clearly, Lemma 4.5 and Proposition 4.6 continue to hold also for F_1. To see that $F_1'(u)$ satisfies an estimate parallel to (5.29) as well, note that due to (5.6) we have

$$\|P'(u)\{v\}w\|_0 \le C\big(\|P(u)v\|_1 + \|v\|_0\big)\|w\|_{s_0-2},$$
$$\|N'(u)\{v\}z\|_0 \le C\big(\|P(u)v\|_1 + \|v\|_0\big)\|z\|_{s_0-2}.$$

This implies such an estimate for F_1. Hence, by changing the definition (5.7) of M into

$$M(u)v := v - \delta\Lambda(u)\big(\psi(u)P(u)v\big)$$

and $\langle\cdot,\cdot\rangle_{s,u}$ accordingly, we obtain the crucial estimates (5.27), (5.28) of Proposition 5.6 also for $\mathcal{F}_1$, at least in the case of strictly positive γ. Note that for $\delta = 0$ the bilinear forms $\langle\cdot,\cdot\rangle_{s,u}$ are in fact independent of u.

6. Proofs of Theorems 3.1 and 3.2. As pointed out earlier, the abstract existence results, Theorems 3.3 and 3.4, provide neither uniqueness of the solution nor strong continuity. The corresponding statements of Theorem 3.1 have to be proved separately. We start with a result on (Lipschitz) continuous dependence on the initial data in a weak norm which immediately implies uniqueness but will also be used in the proof of strong continuity. The techniques (Gronwall's lemma and Taylor expansion, together with the use of estimates obtained earlier) are quite standard.

LEMMA 6.1. *Fix $\widetilde{U}_{s_0} \subset U_{s_0}$. Let $u, v \in C_w\big([0,T], H^{s_0}\big) \cap C_w^1\big([0,T], H^{s_0-2}\big)$ be two solutions of* (3.2) *with*

$$u(t), v(t) \in \widetilde{U}_{s_0} \ \text{ for } \ t \in [0,T].$$

There exists a real number C depending only on T and $\widetilde{U}_{s_0}$ such that

$$\|u(t) - v(t)\|_1 \le C\|u(0) - v(0)\|_1 \ \text{ for all } \ t \in [0,T]. \tag{6.1}$$

Proof. We put $w(t) := v(t) - u(t)$ and remark

$$u, v \in C\big([0,T], H^s\big) \cap C^1\big([0,T], H^{s-2}\big) \ \text{ for } \ 2 \le s < s_0;$$

in particular, the mapping $[0,T] \ni t \mapsto \big\langle w(t), w(t)\big\rangle_{0,u(t)}$ is differentiable and we will show

$$\frac{d}{dt}\big\langle w(t), w(t)\big\rangle_{1,u(t)} \le C\big\langle w(t), w(t)\big\rangle_{1,u(t)}, \tag{6.2}$$

which implies (6.1) via Gronwall's lemma. Recalling that H is a quasi-linear second order differential operator, we have

$$\|H'(z)w\|_1 \le C\|w\|_3, \quad \|H''(z)\{w,w\}\|_1 \le C\|w\|_3\|w\|_{s_0-2}$$

and, accordingly by (4.33),

$$\|G'(z)w\|_1 \le C\|v\|_3, \quad \|G''(z)\{w,w\}\|_1 \le C\|w\|_3\|w\|_{s_0-2}.$$

Consequently, together with Lemma 4.5 and (4.20), we obtain

$$\|\mathcal{F}''(z)\{w,w\}\|_1 \le C\|w\|_3\|w\|_{s_0-2}. \tag{6.3}$$

Using Taylor's theorem we have

$$w'(t) := \frac{d}{dt}w(t) = \mathcal{F}\big(v(t)\big) - \mathcal{F}\big(u(t)\big) = \mathcal{F}'\big(u(t)\big)w(t) + R\big(u(t), v(t)\big);$$

the remainder term therein can be estimated by (6.3) and norm convexity

$$\big\|R\big(u(t), v(t)\big)\big\|_1 \le C_1\|w(t)\|_{s_0-2}\|w(t)\|_3 \le C_2\|w(t)\|_{s_0}\|w(t)\|_1 \le C_3\|w(t)\|_1.$$

From this and (5.27), we obtain

$$\big\langle w(t), w'(t)\big\rangle_{1,u(t)} = \big\langle w(t), \mathcal{F}'\big(u(t)\big)w(t) + R\big(u(t), v(t)\big)\big\rangle_{1,u(t)} \le C\|w(t)\|_1^2. \tag{6.4}$$

Furthermore, recalling (5.8), we have

$$\big\|M'\big(u(t)\big)\{u'(t)\}w(t)\big\|_1 \le C_2\|u'(t)\|_{s_0-2}\|w(t)\|_1.$$

Hence

$$\|u'(t)\|_{s_0-2} = \|\mathcal{F}(u(t))\|_{s_0-2} \le C$$

gives

$$\big\|M'\big(u(t)\big)\{u'(t)\}w(t)\big\|_1 \le C\|w(t)\|_1. \tag{6.5}$$

Consequently, considering

$$\begin{aligned}\frac{1}{2}\tfrac{d}{dt}\big\langle w(t), w(t)\big\rangle_{1,u(t)} \\ &= \big\langle w(t), w'(t)\big\rangle_{1,u(t)} + \Big(M\big(u(t)\big)w(t), M'\big(u(t)\big)\{u'(t)\}w(t)\Big)_1,\end{aligned}$$

we obtain the desired estimate (6.1) from (6.4), (6.5). $\square$

We note a result on nonlinear interpolation, whose proof can be found in [2, Prop. A.1 and Rem. A.2]. It will be crucial in the proof of strong continuity of the solution in time.

LEMMA 6.2. *Let $\mathcal{U} \subseteq H^s(S,\mathbb{R}^m)$, $s \geq 1$, be an open set. Let $T_\alpha : \mathcal{U} \to H^1(S,\mathbb{R}^m)$ be mappings with $T_\alpha(\mathcal{U}\cap H^{s+1}) \subseteq H^{s+1}$; α runs through a certain index set I. Further, assume Lipschitz continuity of T_α in H^1 and boundedness of T_α in H^{s+1}:*

$$\|T_\alpha(u) - T_\alpha(v)\|_1 \leq C\|u - v\|_1 \ \textit{ for all } \ u, v \in \mathcal{U},$$
$$\|T_\alpha(u)\|_{s+1} \leq C\big(1 + \|u\|_{s+1}\big) \ \textit{ for all } \ u \in \mathcal{U} \cap H^{s+1}$$

with a constant C independent of u, v and $\alpha \in I$. Then $T_\alpha(\mathcal{U} \cap H^s) \subseteq H^s$ and the mappings $T_\alpha : \mathcal{U} \subseteq H^s \to H^s$ are continuous, uniformly w.r.t. $\alpha \in I$.

Now we are prepared for the proof of our theorems. In many respects, it is parallel to the proof of the main results in [10].

Proof of Theorem 3.1.

Step 1. We show that for any given $\bar{u}_0 \in U_{s_0}$ and any integer $s \geq s_0$ there exist $T = T(\bar{u}_0, s) > 0$ and $\delta = \delta(\bar{u}_0, s) > 0$ such that the Cauchy problem (3.2) has a unique solution in the class

$$u \in C_w\big([0,T], U_s\big) \cap C_w^1\big([0,T], H^{s-2}\big)$$

for all initial values $u_0 \in H^s$ with $\|u_0 - \bar{u}_0\|_{s_0} \leq \delta$. The uniqueness of the solution follows immediately from Lemma 6.1. In order to prove the existence we use Theorem 3.4. With a fixed $s \geq s_0$ and an $\varepsilon \in (0,1]$ which will be fixed below, we put

$$X = H^{s+2}(S,\mathbb{R}^m), \quad \|\cdot\|_X = \|\cdot\|_{s_0+2} + \varepsilon\|\cdot\|_{s+2};$$
$$Y = H^{s}(S,\mathbb{R}^m), \quad \|\cdot\|_Y = \|\cdot\|_{s_0} + \varepsilon\|\cdot\|_{s};$$
$$Z = H^{s-2}(S,\mathbb{R}^m), \quad \|\cdot\|_Z = \|\cdot\|_{s_0-2} + \varepsilon\|\cdot\|_{s-2}.$$

Further, let $\widetilde{U}_s$ be as in section 5 and assume that the given $\bar{u}_0$ is an interior point. Then, according to the results of section 5, for $u \in \widetilde{U}_s$ the bilinear forms $\langle v, w\rangle_{s,u}^\varepsilon$: $X \times Z \to \mathbb{R}$ satisfy the requirements (H) of section 3; note that the constants C, M in (H) can be chosen independently of ε. As in the proof of Theorem 3.3 we choose $w_0 \in C^\infty(S,\mathbb{R}^m)$ and $R > 0$ (both independent of ε) such that

$$\|w_0 - u_0\|_{s_0} \leq R/(32C^5)^{1/2}, \quad \{w_0 + v \mid v \in B\} \subseteq \widetilde{U}_{s_0}$$

with the ball $B := \big\{v \in Y \mid \|v\|_Y < R\big\}$. We set

$$\langle v, w\rangle_u := \langle v, w\rangle_{s,w_0+u}^\varepsilon, \quad |||v||| = \langle v, v\rangle_{w_0+v}$$

and define a map $\mathcal{H} : B \subseteq Y \to Z$ by

$$\mathcal{H}(v) := \mathcal{F}(v + w_0), \quad u \in B.$$

Further, the mapping $\mathcal{H} : B \subseteq Y \to Z$ is weakly sequentially continuous and

$$\langle w_0 + v, \mathcal{H}(v)\rangle \leq C_1\|w_0 + v\|_Y^2 \leq C_2$$

by Proposition 5.6. Moreover we have

$$\|\mathcal{H}(v)\|_Z \leq C_3\|v + w_0\|_Y \leq C_4$$

and

$$|\langle w_0, \mathcal{H}(v)\rangle_v| \le C_5 \|w_0\|_X \|\mathcal{H}(v)\|_Z \le C_6.$$

These estimates hold for all $v \in B$ with constants $C_1, \ldots,$ which may depend on C, M, R, s and $\bar{u}_0, w_0$, but not on v. Gathering them, we obtain the inequality

$$2\langle v, \mathcal{H}(v)\rangle_v + M\|\mathcal{H}(v)\|_Z |\!|\!| v |\!|\!| \le C_7 \ \text{ for all } \ v \in B \cap X.$$

Now, let $u_0 \in H^s(S, \mathbb{R}^m)$ be given such that

$$\|u_0 - \bar{u}_0\|_{s_0} \le R/(32C^5)^{1/2}. \tag{6.6}$$

Hence, with $r := R/(2C^3)^{1/2}$ we find

$$|\!|\!| u_0 - w_0 |\!|\!| \le C\big(\|u_0 - \bar{u}_0\|_{s_0} + \|\bar{u}_0 - w_0\|_{s_0} + \varepsilon(\|u_0\|_s + \|w_0\|_s)\big) \le r$$

if ε is chosen according to

$$\varepsilon := \min\big\{1, r/4C(\|u_0\|_s + \|w_0\|_s)\big\}. \tag{6.7}$$

By Theorem 3.4, applied to $\mathcal{H}$, there exist $T > 0$, independent of u_0 with (6.6), and a solution

$$v \in C_w\big([0,T], B \cap H^s\big) \cap C^1_w\big([0,T], H^{s-2}\big)$$

of

$$dv(t)/dt = \mathcal{G}(v(t)) \ \text{ for } \ t \in [0,T], \quad v(0) = u_0 - w_0.$$

Then $u := v + w_0$ is a solution of (3.2) with initial value $u(0) = u_0$, and we have

$$\|u(t)\|_s \le \|w_0\|_s + \|v(t)\|_s \le \|w_0\|_s + \varepsilon^{-1}\|v(t)\|_Y,$$

which in view of (6.7) implies

$$\|u(t)\|_s \le C(1 + \|u(0)\|_s). \tag{6.8}$$

Step 2. Let $u, \tilde{u}$ be two solutions of (3.2) in $[0,T]$ according to Step 1 with initial values

$$u(0), \tilde{u}(0) \in \mathcal{U}, \quad \mathcal{U} := \big\{v \in H^s \mid \|v - \bar{u}_0\|_{s_0} \le \delta\big\},$$

$\delta > 0$ sufficiently small. Lemma 6.1 gives

$$\|u(t) - \tilde{u}(t)\|_1 \le C\|u(0) - \tilde{u}(0)\|_1. \tag{6.9}$$

For fixed $t \in [0,T]$ we consider the evolution operator

$$\mathcal{U} \ni u_0 \mapsto T_t(u_0) := u(t) \in H^s,$$

assigning to any initial value u_0 the value of the corresponding solution of (3.2) at time t. By Step 1 with s replaced by $s+1$ we obtain $T_t(\mathcal{U} \cap H^{s+1}) \subseteq H^{s+1}$ and the estimate

$$\|T_t(u_0)\|_{s+1} \le C\big(1 + \|u_0\|_{s+1}\big). \tag{6.10}$$

Equations (6.9) and (6.10) together with the interpolation result from Lemma 6.2 show the continuity of the mapping

$$\mathcal{U} \cap H^s \ni u_0 \mapsto u(t) \in H^s \ \text{ for } \ s \geq s_0,$$

uniformly w.r.t. $t \in [0, T]$.

Step 3. To complete the proof of Theorem 3.1 it remains to show that the solutions according to Step 1 actually belong to

$$u \in C\big([0,T], H^s\big) \cap C^1\big([0,T], H^{s-2}\big). \tag{6.11}$$

To do this, we approximate the initial value $u_0 = u(0)$ by a sequence $u_0^n \in H^{s+1}$ such that $u_0^n \to u_0$ in H^s. Then by Step 1, for n sufficiently large, there exist solutions u_n of (3.2) with $u_n(0) = u_0^n$ in the class

$$u_n \in C_w\big([0,T], H^{s+1}\big) \cap C_w^1\big([0,T], H^{s-1}\big),$$

which in particular implies

$$u_n \in C\big([0,T], H^s\big) \cap C^1\big([0,T], H^{s-2}\big).$$

On the other hand, by Step 2, we have $u_n(t) \to u(t)$ in H^s uniformly w.r.t. $t \in [0,T]$. Since the uniform limit of continuous functions is continuous again, this implies (6.11). ∎

Proof of Theorem 3.2. Let a solution $u \in C([0,T], U_s) \cap C^1([0,T], H^{s-2})$ be given. The set $\{u(t) \,|\, t \in [0,T]\}$ is compact in H^s and can be covered by the open sets $\{v \in H^s \,|\, \|v - u(t)\|_s < \delta(u(t), s+1)\}$, $t \in [0,T]$, where $\delta(u(t), s+1)$ are the same as in the proof of Theorem 3.1. Choosing a finite subcover, we find from this theorem and the autonomous character of (3.2) that there is a $T_0 > 0$ such that for any $t \in [0,T]$ with $u(t) \in H^{s+1}$, we have

$$u|_{[t,T_1]} \in C([t,T_1], U_{s+1}) \cap C^1([t,T_1], H^{s-1}), \quad T_1 := \min\{t + T_0, T\}.$$

Proceeding stepwise, we obtain (i).

A similar compactness argument together with Theorem 3.1 and its proof ensures the existence of $T_2 > 0$ such that the following is true for all $t \in [0,T]$: Problem (3.2) is solvable on the time interval $[0, T_2]$ (in the class (3.13)) for all initial values z sufficiently near $u(t)$, and the mapping which assigns to z its corresponding solution $V(\cdot, z)$ is continuous with values in $C([0,T_2], H^s)$. We choose $t_i \in [0,T]$ such that $0 = t_0 < \cdots < t_n = T$, $t_i - t_{i-1} < T_2$, and open H^s-neighborhoods K_i of $u(t_i)$ small enough to ensure that V is defined on K_i and $V(t_i - t_{i-1}, K_{i-1}) \subset K_i$, $i = n-1, \dots, 1$. Now (ii) follows from the continuity of the composition of continuous maps. ∎

Appendix. Proof of Theorem 3.4. We will construct a solution of (3.19) by implicit time discretization, solving the nonlinear problems in each timestep by Galerkin approximations. For this purpose, we need the following lemma.

LEMMA A.1. *For any $K \in (0, r^2)$ there is an $\varepsilon_0 > 0$ such that for any $\varepsilon \in (0, \varepsilon_0]$ and any $v \in Y$ satisfying $|||v|||^2 \leq K$ there is a $u^* \in B$ satisfying*

$$u^* = v + \varepsilon \mathcal{G}(u^*) \tag{A.1}$$

and the estimate

$$|||u^*|||^2 \leq |||v|||^2 + \varepsilon \beta\big(|||u^*|||^2\big) \leq 2K. \tag{A.2}$$

Proof. For arbitrary $v \in Y$, $u \in X \cap B$ we have

$$\begin{aligned}\langle u, u - \varepsilon\mathcal{G}(u) - v\rangle_u &= |||u|||^2 - \varepsilon\langle u, \mathcal{G}(u)\rangle_u - (u,v)_u \\ &\geq |||u|||^2 - \frac{\varepsilon}{2}\beta\big(|||u|||^2\big) + \frac{\varepsilon M}{2}\|\mathcal{G}(u)\|_Z|||u|||^2 - |||u|||\|v\|_u \\ &\geq \frac{1}{2}\Big(|||u|||^2 - \varepsilon\beta\big(|||u|||^2\big) - \|v\|_u^2 + \varepsilon M\|\mathcal{G}(u)\|_Z|||u|||^2\Big). \end{aligned} \tag{A.3}$$

Choose $\varepsilon_0 > 0$ such that for all $\varepsilon \in (0, \varepsilon_0]$ and for all $s \in [0, 2C^4K]$

$$K - \varepsilon\beta(s) \geq 0, \tag{A.4}$$

$$1 - \varepsilon\beta'(s) \geq 0. \tag{A.5}$$

Assume now that $v \in B$, $|||v|||^2 \leq K$. Let

$$\mathcal{B} := \{u \in Y \mid \|u\|_Y^2 \leq 2KC^3\}$$

and note that $\mathcal{B}$ is a closed convex subset of B. Assume $\|u\|_Y^2 = 2KC^3$. Then

$$\begin{gathered} 2C^2K = C^{-1}\|u\|_Y^2 \leq |||u|||^2 \leq C\|u\|_Y^2 = 2C^4K, \\ \|v\|_u^2 \leq C\|v\|_Y^2 \leq C^2|||v|||^2 \leq C^2K. \end{gathered}$$

Therefore, for $\varepsilon \in (0, \varepsilon_0]$,

$$\langle u, u - \varepsilon\mathcal{G}(u) - v\rangle_u \geq \frac{1}{2}(C^2K - \varepsilon\beta(|||u|||^2)) \geq 0. \tag{A.6}$$

Let $\{M_n\}$ be an increasing sequence of finite-dimensional subspaces of X whose union is dense in X. We fix n, choose a basis $\{e_1, \ldots, e_n\}$ of M_n, and show that the variational equality

$$\langle w, u_n - \varepsilon\mathcal{G}(u_n) - v\rangle_u = 0 \;\text{ for all }\; w \in M_n \tag{A.7}$$

has a solution $u_n \in M_n \cap \mathcal{B}$. Note that (A.7) is equivalent to $g(u_n) = 0$, where $g : M_n \cap \mathcal{B} \to M_n$ is defined by

$$g(u) := P_u(u - \varepsilon\mathcal{G}(u) - v) \quad \text{with} \quad P_u(z) := \sum_{i=1}^{n}\langle e_i, z\rangle_u e_i.$$

Due to (H4), g is continuous. Assume now that $g(u) \neq 0$ for all $u \in M_n \cap \mathcal{B}$. Then we define the continuous operator $f : M_n \cap \mathcal{B} \to M_n$ by

$$f(u) := -\sqrt{2KC^3}g(u)/\|g(u)\|_Y.$$

As $\|f(u)\|_Y^2 = 2KC^3$, f maps the closed convex set $M_n \cap \mathcal{B}$ into itself. Therefore, by Brouwer's fixed point theorem, there is a $\overline{u} \in M_n \cap \mathcal{B}$ such that $\overline{u} = f(\overline{u})$. Consequently, $\|\overline{u}\|_Y^2 = 2KC^3$, and from (A.6) we obtain the contradictory inequality

$$\begin{aligned} 0 < |||\overline{u}|||^2 = \big\langle \overline{u}, f(\overline{u})\big\rangle_{\overline{u}} &= -\frac{\sqrt{2KC^3}}{\|g(u)\|_Y}\big\langle \overline{u}, g(\overline{u})\big\rangle_{\overline{u}} \\ = -\frac{\sqrt{2KC^3}}{\|g(u)\|_Y}\big\langle \overline{u}, \overline{u} - \varepsilon\mathcal{G}(\overline{u}) - v\big\rangle_{\overline{u}} &\leq 0. \end{aligned}$$

Therefore, (A.7) is solvable for every n, and as $\{u_n\}$ is bounded in Y, we can assume without loss of generality that $u_n \rightharpoonup u^*$ in Y for some $u^* \in \mathcal{B}$. Passage to the limit in (A.7) yields by (H4)

$$\langle w, u^* - \varepsilon\mathcal{G}(u^*) - v\rangle_{u^*} = 0 \text{ for all } w \in M_n, n = 1, 2, \dots,$$

and consequently by the density assumption

$$\langle w, u^* - \varepsilon\mathcal{G}(u^*) - v\rangle_{u^*} = 0 \text{ for all } w \in X.$$

The nondegeneracy of $\langle\cdot,\cdot\rangle_{u^*}$ yields (A.1). To show the estimate (A.2), note first that

$$|||u^*|||^2 \le \underline{\lim}_{\,n\to\infty} |||u_n|||^2 \le 2C^4K.$$

Thus, the second inequality in (A.2) follows from (A.4). To show the first inequality we assume without loss of generality that $|||v||| \le |||u^*|||$ and use (A.5), (A.3), and (H4) to obtain

$$\begin{aligned}
|||u^*|||^2 - \varepsilon\beta(|||u^*|||^2) &\le \underline{\lim}_{\,n\to\infty}\big(|||u_n|||^2 - \varepsilon\beta\big(|||u_n|||^2\big)\big)\\
&\le \underline{\lim}_{\,n\to\infty}\big(\|v\|_{u_n}^2 - M\varepsilon\|\mathcal{G}(u_n)\|_Z|||u_n|||^2\big) \le \|v\|_{u^*}^2 - M\varepsilon\|\mathcal{G}(u^*)\|_Z|||u^*|||^2\\
&\le |||v|||^2 + M\|u^* - v\|_Z|||v|||^2 - M\varepsilon\|\mathcal{G}(u^*)\|_Z|||u^*|||^2\\
&= |||v|||^2 + M\varepsilon\|\mathcal{G}(u^*)\|_Z|||v|||^2 - M\varepsilon\|\mathcal{G}(u^*)\|_Z|||u^*|||^2 \le |||v|||^2. \qquad \square
\end{aligned}$$

As further preparation for the proof of Theorem 3.4 we need the following simple result on approximate solutions of the ordinary differential equation (3.18).

LEMMA A.2. *Assume $u_0 \in B$ and let $\rho \in C^1[0,T]$ be the solution of (3.18). There is an $n_0 \in \mathbb{N}$ such that for $n \ge n_0$ and $k = 1, \dots, n$ there are $\rho_n^k, r_n \in \mathbb{R}$ such that*

$$\rho_n^0 = |||u_0|||^2, \quad \rho_n^k + \delta_n\beta(\rho_n^{k+1}) \le \rho_n^{k+1} \le \rho\big((k+1)\delta_n\big) + r_n, \quad r_n \to 0,$$

where $\delta_n := T/n$.

Proof. If n_0 is sufficiently large, $n \ge n_0$, there exist solutions $\rho_n \in C^1[0,T]$ to the initial value problems

$$\rho_n'(t) = \beta\big(\rho_n(t)\big) + 1/\sqrt{n}, \quad \rho_n(0) = |||u_0|||^2.$$

We set

$$\rho_n^k := \rho_n(k\delta_n), \qquad k = 0, \dots, n.$$

Then

$$\rho_n^{k+1} - \rho_n^k = \delta_n\rho_n'(\xi) = \delta_n\beta(\rho_n(\xi)) + \delta_n n^{-1/2}$$

for some $\xi \in \big(k\delta_n, (k+1)\delta_n\big)$. Moreover,

$$\big|\beta\big(\rho_n(\xi)\big) - \beta\big(\rho_n^{k+1}\big)\big| \le S\big|\rho_n(\xi) - \rho_n^{k+1}\big| \le S'n^{-1}$$

with constants S, S' independent of n. Thus

$$\rho_n^{k+1} - \rho_n^k \ge \delta_n\beta(\rho_n^{k+1}) + \delta_n n^{-1/2} - S'\delta_n n^{-1} \ge \delta_n\beta(\rho_n^{k+1})$$

for $n \geq n_0$, n_0 sufficiently large. Moreover, well-known results on the dependence of the solution of ordinary differential equation's on their right-hand sides ensure

$$r_n := \max_{t\in[0,T]} |\rho_n(t) - \rho(t)| \to 0, \qquad n \to \infty,$$

and hence

$$\rho_n(t) \leq \rho(t) + r_n, \qquad t \in [0,T].$$

This proves the lemma. □

Proof of Theorem 3.4. In a first step, we construct approximations u_n^k for the solution at time kT/n. Choose $K \in \big(\max_{t\in[0,T]} \rho(t), r^2\big)$ and choose $\varepsilon_0 > 0$ such that the assertions of Lemma A.1 and (A.5) hold. Let $n_0 \in \mathbb{N}$ be at least as large as in Lemma A.2 and assume additionally that $n_0 \geq T/\varepsilon_0$ and

$$\rho(t) + r_n \leq K \quad \text{for } n \geq n_0 \text{ and } t \in [0,T].$$

Now we fix $n \geq n_0$ and show the existence of $u_n^k \in B$, $k = 0, \dots, n$, such that

$$\begin{aligned} u_n^{k+1} &= u_n^k + \delta_n \mathcal{G}(u_n^{k+1}), \quad k = 0, \dots, n-1, \\ u_n^0 &= u_0, \\ |||u_n^k|||^2 &\leq \rho_n^k, \end{aligned}$$

where the ρ_n^k are given by Lemma A.2. For $k = 0$, existence and the estimate are clear. Assume now that $u_n^0, \dots, u_n^k$ are constructed according to these conditions for $0 \leq k \leq n-1$. Our assumptions imply $\delta_n \leq \varepsilon_0$ and $|||u_n^k|||^2 \leq K$; hence the existence of u_n^{k+1} follows from Lemma A.1. Moreover, by (A.2), $|||u_n^{k+1}|||^2 \leq 2K$ and

$$|||u_n^{k+1}|||^2 \leq |||u_n^k|||^2 + \delta_n \beta\big(|||u_n^{k+1}|||^2\big) \leq \rho_n^k + \delta_n \beta\big(|||u_n^{k+1}|||^2\big);$$

hence

$$|||u_n^{k+1}|||^2 - \delta_n \beta\big(|||u_n^{k+1}|||^2\big) \leq \rho_n^{k+1} - \delta_n \beta\big(\rho_n^{k+1}\big).$$

Note that (A.5) implies that the mapping $s \mapsto s - \delta_n \beta(s)$ is monotone increasing on $[0, 2K]$, and hence $|||u_n^{k+1}|||^2 \leq \rho_n^{k+1}$.

In a second step, we approximate u on $[0,T]$ by piecewise linear functions u_n and piecewise constant functions $\overline{u}_n$, $n \geq n_0$, given by

$$\begin{aligned} u_n(t) &:= u_n^k + \delta_n^{-1}(t - k\delta_n)(u_n^{k+1} - u_n^k) \ \text{ for } \ k\delta_n \leq t \leq (k+1)\delta_n, \\ &\qquad k = 0, \dots, n-1, \\ \overline{u}_n(t) &:= u_n^{k+1} \text{ for } k\delta_n < t \leq (k+1)\delta_n, \ k = 0, \dots, n-1, \ \overline{u}_n(0) = u_n^0. \end{aligned}$$

Then

$$u_n(t) = u_0 + \int_0^t \mathcal{G}\big(\overline{u}_n(\tau)\big)\, d\tau, \quad t \in [0,T],$$

and with a suitable constant S independent of $t \in [0,T]$ and $n \geq n_0$,

$$\|u_n(t)\|_Y, \|\overline{u}_n(t)\|_Y \leq S.$$

Consequently, $\|\mathcal{G}\big(\overline{u}_n(t)\big)\|_Z$ is bounded independently of n and thus

$$\|u_n(t) - u_n(t')\|_Z \le L|t - t'|$$

with L independent of n. Hence, the sequence $\{u_n\}$ is bounded and equicontinuous with values in Z, and hence by Ascoli's theorem, we can assume without loss of generality that

$$u_n \to u \ \text{ in } \ C\big([0,T], Z\big).$$

Moreover,

$$u_n(t) \rightharpoonup u(t) \ \text{ in } \ Y, \quad t \in [0,T]. \tag{A.8}$$

To show this, fix $t \in [0,T]$ and choose an arbitrary subsequence $\{u_{n'}(t)\}$. As it is bounded in Y, it has a weakly convergent subsequence $\{u_{n''}(t)\}$ for which $u_{n''}(t) \rightharpoonup u^\star$ in Y, and hence also in Z, and thus $u^\star = u(t)$. Now (A.8) follows from a standard argument. An analogous argument shows

$$u \in C_w\big([0,T], Y\big).$$

Furthermore, for $t \in \big(k\delta_n, (k+1)\delta_n\big]$ we have

$$\|\overline{u}_n(t) - u_n(t)\|_Z = \|u_n\left((k+1)\delta_n\right) - u_n(t)\|_Z \le L\delta_n,$$

hence also

$$\overline{u}_n \to u \ \text{ in } \ C\big([0,T], Z\big),$$

and, by the same arguments as for u_n above,

$$\overline{u}_n(t) \rightharpoonup u(t) \ \text{ in } \ Y, \quad t \in [0,T].$$

As $\mathcal{G}$ is weakly sequentially continuous,

$$\mathcal{G}\big(\overline{u}_n(t)\big) \rightharpoonup \mathcal{G}\big(u(t)\big) \ \text{ in } \ Z, \quad t \in [0,T],$$

and $\mathcal{G} \circ u \in C_w\big([0,T], Z\big)$. If f is any bounded linear functional on Z, it follows that

$$f\left(\int_0^t \mathcal{G}(\overline{u}_n(\tau))\, d\tau\right) = \int_0^t f\big(\mathcal{G}(\overline{u}_n(\tau))\big)\, d\tau \to \int_0^t f\big(\mathcal{G}(u(\tau))\big)\, d\tau, \qquad n \to \infty,$$

and hence

$$f\big(u(t)\big) = f(u_0) + \int_0^t f\big(\mathcal{G}(u(\tau))\big)\, d\tau, \quad t \in [0,T].$$

Consequently,

$$f\left(\frac{u(t+h) - u(t)}{h}\right) \to f\big(\mathcal{G}(u(t))\big), \quad h \to 0,$$

i.e.,

$$\frac{u(t+h) - u(t)}{h} \rightharpoonup \mathcal{G}\big(u(t)\big) \ \text{ in } \ Z, \quad h \to 0.$$

Therefore $u \in C^1_w([0,T],Z)$ and u satisfies (3.19). Finally, for $t \in (k\delta_n,(k+1)\delta_n]$ we get

$$|||\overline{u}_n(t)|||^2 = |||u_n^{k+1}|||^2 \le \rho\big((k+1)\delta_n\big) + r_n,$$

and hence

$$|||\overline{u}_n(t)|||^2 \le \rho(t+\delta_n) + r_n \quad \text{for } \; 0 \le t \le T-\delta_n.$$

Thus

$$|||u(t)|||^2 \le \underline{\lim}_{\,n\to\infty} |||\overline{u}_n(t)|||^2 \le \rho(t), \quad t \in [0,T].$$

For $t \to 0$ this implies, in particular,

$$\overline{\lim}_{\,t\to 0} |||u(t)|||^2 \le \lim_{t\to 0} \rho(t) = |||u(0)|||^2 \le \underline{\lim}_{\,t\to 0} |||u(t)|||^2;$$

hence $|||u(t)||| \to |||u(0)|||$ and consequently $u(t) \to u(0)$ in Y as $t \to 0$. $\square$

REFERENCES

[1] R. Almgren, *Singularity formation in Hele-Shaw bubbles*, Phys. Fluids, 8 (1996), pp. 344–352.

[2] K. Beyer and M. Günther, *On the Cauchy problem for a capillary drop. Part* I: *Irrotational motion*, Math. Methods Appl. Sci., 21 (1998), pp. 1149–1183.

[3] K. Beyer and M. Günther, *The Jacobi equation for irrotational free boundary flows*, Analysis, 20 (2000), pp. 237–254.

[4] P. Constantin and M. Pugh, *Global solution for small data to the Hele-Shaw problem*, Nonlinearity, 6 (1993), pp. 393–415.

[5] K. Deckelnick and C. Elliott, *Local and global existence results for anisotropic Hele-Shaw flows*, Proc. Roy. Soc. Edinburgh Sect. A, 129 (1999), pp. 265–294.

[6] J. Escher and G. Simonett, *A center manifold analysis for the Mullins-Sekerka model*, J. Differential Equations, 143 (1998), pp. 267–292.

[7] A. Friedman and F. Reitich, *Nonlinear stability of a quasi-static Stefan problem with surface tension: A continuation approach*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 30 (2001), pp. 341–403.

[8] L. Giacomelli and F. Otto, *Variational formulation for the lubrication approximation of the Hele-Shaw flow*, Calc. Var. Partial Differential Equations, 3 (2001), pp. 377–403.

[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, New York, 1977.

[10] M. Günther and G. Prokert, *On Stokes flow with variable and degenerate surface tension coefficient*, NoDEA Nonlinear Differential Equations Appl., 12 (2005), pp. 21–60.

[11] Yu. E. Hohlov and M. Reissig, *On classical solvability for the Hele-Shaw moving boundary problems with kinetic undercooling regularization*, European J. Appl. Math., 6 (1995), pp. 421–439.

[12] L. Hörmander, *The boundary problems of physical geodesy*, Arch. Ration. Mech. Anal., 62 (1976), pp. 1–52.

[13] S. D. Howison, *Complex variable methods in Hele-Shaw moving boundary problems*, European J. Applied Math., 3 (1992), pp. 209–224.

[14] T. Kato and C. Y. Lai, *Nonlinear evolution equations and the Euler flow*, J. Funct. Anal., 56 (1984), pp. 15–28.

[15] S. R. K. Maruvada, C. W. Park, and D. Y. Yoon, *The influence of surfactant on the bubble motion in Hele-Shaw cells*, Phys. Fluids, 6 (1994), pp. 3267–3275.

[16] G. Prokert, *Hyperbolic evolution equations for moving boundary problems*, European J. Applied Math., 10 (1999), pp. 607–622.

[17] G. Prokert, *On Stokes flow driven by surface tension in the presence of a surfactant*, European J. Applied Math., 15 (2004), pp. 791–819.

[18] L. A. Romero, *The Fingering Problem in a Hele-Shaw Cell*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1981.

[19] P. G. Saffman, *Viscous fingering in Hele-Shaw cells*, J. Fluid Mech., 173 (1986), pp. 73–94.

[20] F. Yi, *Asymptotic behaviour of the solutions of the supercooled Stefan problem*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 181–190.

[21] W. Yu, *A quasisteady Stefan problem with curvature correction and kinetic undercooling*, J. Partial Differential Equations, 9 (1996), pp. 55–70.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 411–437

LIMITS OF SOLUTIONS OF -LAPLACE EQUATIONS AS GOES TO INFINITY AND RELATED VARIATIONAL PROBLEMS*

HITOSHI ISHII† AND PAOLA LORETI‡

Abstract. We show that the convergence, as $p \to \infty$, of the solution u_p of the Dirichlet problem for $-\Delta_p u(x) = f(x)$ in a bounded domain $\Omega \subset \mathbf{R}^n$ with zero-Dirichlet boundary condition and with continuous f in the following cases: (i) one-dimensional case, radial cases; (ii) the case of no balanced family; and (iii) two cases with vanishing integral. We also give some properties of the maximizers for the functional $\int_\Omega f(x)v(x)\,\mathrm{d}x$ in the space of functions $v \in C(\overline{\Omega}) \cap W^{1,\infty}(\Omega)$ satisfying $v|_{\partial\Omega} = 0$ and $\|Dv\|_{L^\infty(\Omega)} \le 1$.

Key words. p-Laplace equation, asymptotic behavior, variational problem, L^∞ variational problem, eikonal equation, ∞-Laplace equation

AMS subject classifications. 35B40, 35J60, 35J20, 35F30

DOI. 10.1137/S0036141004432827

1. Introduction. We study the asymptotic behavior, as $p \to \infty$, of the solution u_p of the Dirichlet problem

$$\begin{cases} -\Delta_p u(x) = f(x) & \text{in } \Omega, \\ u(x) = 0 & \text{for } x \in \partial\Omega. \end{cases} \tag{1.1}$$

Here and henceforth Δ_p denotes the p-Laplacian, i.e.,

$$\Delta_p u(x) := \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(|Du|^{p-2}\frac{\partial u}{\partial x_i}\right),$$

$\Omega \subset \mathbf{R}^n$ is a bounded open set, the exponent p satisfies $p > 1$, and $f \in C(\overline{\Omega})$.

The PDE in (1.1) is the Euler–Lagrange equation of the maximization problem for the functional

$$I_p(u) := \int_\Omega \left(f(x)u(x) - \frac{1}{p}|Du(x)|^p\right)\mathrm{d}x \quad \text{over } W_0^{1,p}(\Omega). \tag{1.2}$$

As is well known, the two problems (1.1) and (1.2) are equivalent. The problem (1.1) has a unique solution $u \in W_0^{1,p}(\Omega)$, and so does (1.2). For the existence and uniqueness of a solution of (1.1), we refer to [L]. According to the regularity results for (1.1), the solution u_p has Hölder continuous derivatives in Ω. That is, $u_p \in C^{1,\gamma}(\Omega)$ for some constant $\gamma \in (0,1)$ which depends on p. Moreover, if the boundary $\partial\Omega$ is smooth, then $u_p \in C^{1,\gamma}(\overline{\Omega})$. See [U, D, Lb, T] for these regularity properties.

The asymptotic problem for (1.1) as $p \to \infty$ appears in modeling of a torsional creep phenomenon for a prismatic elastoplastic rod. This corresponds to the case

*Received by the editors March 16, 2004; accepted for publication (in revised form) January 12, 2005; published electronically October 14, 2005.
http://www.siam.org/journals/sima/37-2/43282.html

†Department of Mathematics, School of Education, Waseda University, Nishi-Waseda 1-6-1, Shinjuku, Tokyo 169-8050, Japan (ishii@edu.waseda.ac.jp). This author was supported in part by Grant-in-Aid for Scientific Research grants 15340051 and 14654032 and by Waseda University Grant for Special Research Projects, grant 2001A-838.

‡Dipartimento di Metodi e Modelli Matematici per le Scienze Applicate, Università degli Studi di Roma "La Sapienza," Via Scarpa n. 16, 00161 Roma, Italy (loreti@dmmm.uniroma1.it).

where $n = 2$ and f is a positive constant (see, for instance, [BDM, K, PP]). In fact, if $f > 0$, then the limit of u_p in $C(\overline{\Omega})$ exists and is the distance function from the boundary $\partial\Omega$, i.e., the function $d(x) := \operatorname{dist}(x, \partial\Omega)$. See [IK, IL1, IL2, FIN, BK, JLM] for some related topics.

This convergence result is then generalized to the case of general nonnegative functions f by using the ∞-Laplace equation in the region ω where f vanishes, i.e, solving the problem

$$\begin{cases} -\Delta_\infty w(x) = 0 & \text{in } \omega, \\ \qquad w(x) = d(x) & \text{on } \partial\omega, \end{cases} \tag{1.3}$$

where

$$\Delta_\infty w(x) := \sum_{i,j=1}^{n} \frac{\partial w(x)}{\partial x_i} \frac{\partial w(x)}{\partial x_j} \frac{\partial^2 w(x)}{\partial x_i \partial x_j} \quad \text{and} \quad \omega := \operatorname{int}\{x \in \Omega \mid f(x) = 0\}.$$

Due to [J] (see also [BB]), the problem (1.3) has a unique viscosity solution $w \in C(\overline{\omega})$ which is Lipschitz continuous in ω and satisfies $\|Dw\|_{L^\infty(\omega)} \leq 1$. If we assume that $f \geq 0$ in Ω and define $U \in C(\overline{\Omega})$ by

$$U(x) = \begin{cases} d(x) & \text{for } x \in \overline{\Omega} \setminus \omega, \\ w(x) & \text{for } x \in \omega, \end{cases}$$

where w is the unique viscosity solution of (1.3), then U gives the limit of u_p in $C(\overline{\Omega})$ as $p \to \infty$. See Remark 5.2 in [BDM], where the above idea of finding the limit function appears. See also [CIL] for an introduction to viscosity solutions.

In 1967 G. Aronsson initiated the study of the ∞-Laplace equation in his study of absolutely minimal Lipschitz extensions (AMLE), also called canonical Lipschitz extensions, to a domain ω of a function given on $\partial\omega$. The AMLE and ∞-Laplace equation are subjects which have seen intensive research activities recently. For these developments, we refer to [A1, A2, CEG, ACJ].

As we will see in section 5, the family $\{u_p\}_{p>q}$, with $q > n$, is precompact in $C(\overline{\Omega})$. Therefore, $\{u_p\}_{p>1}$ has a sequence $\{u_{p_j}\}_{j\in\mathbf{N}}$ convergent in $C(\overline{\Omega})$, where $p_j \to \infty$ as $j \to \infty$. However, it is not clear whether the whole family $\{u_p\}_{p>1}$ is convergent in $C(\overline{\Omega})$ or not, except in the case where $f \geq 0$.

In this paper we address ourselves to the question of whether the whole family $\{u_p\}_{p>1}$ is convergent in $C(\overline{\Omega})$ as $p \to \infty$. We present only partial positive answers to this question in this paper.

In the cases where $n = 1$ or when Ω is an open ball and f is a radial function, we show the convergence of u_p in $C(\overline{\Omega})$ and identify the limit function. In these cases, our proof relies heavily on an explicit formula for u_p.

In the general situation we do not know any convenient formula for u_p and in our approach we make a careful study (especially the structure of its maximizers) of the variational problem for the functional

$$I_\infty(u) := \int_\Omega f(x)u(x)\mathrm{d}x \tag{1.4}$$

over the set $X := \{v \in C(\overline{\Omega}) \cap W^{1,\infty}(\Omega) \mid v|_{\partial\Omega} = 0,\ \|Dv\|_{L^\infty(\Omega)} \leq 1\}$. This variational problem appears as the limit problem of (1.2). (See Proposition 5.3 below.) This problem may be conceived of an L^∞ variational problem because of the L^∞ bound on the gradient Du and because it appears as the limit problem for the variational problem (1.2) as $p \to \infty$.

As a generalization of the case where $f \geq 0$, we show the convergence of u_p in $C(\overline{\Omega})$ under the condition of *no balanced family*, i.e., under the assumption that for any nonempty family $\mathcal{C}$ of Lipschitz-connected components of $\{x \in \Omega \mid f(x) \neq 0\}$ that stay away from $\partial\Omega$ and $\omega := \bigcup\{U \in \mathcal{C}\}$ (the union of the sets U, where U ranges over all $U \in \mathcal{C}$),

$$\int_\omega f(x)\,\mathrm{d}x \neq 0.$$

Here the standard definition of connected components is not appropriate and we have used the notion of Lipschitz-connected (L-connected, for short) components. See section 2 for the precise assumption, (2.4), and for the definition of L-connected components.

We also consider the case when

$$\int_\Omega f(x)\,\mathrm{d}x = 0 \quad \text{and} \quad f \neq 0. \tag{1.5}$$

This is the case when the above assumption (the assumption of no balanced family) is not satisfied. Also, this is the case related to the Monge–Kantorovich mass transfer problem. The Monge–Kantorovich mass transfer problem has received much attention in the last decade. We refer to [EG, BBD, ACBBV] for the recent developments of the Monge–Kantorovich mass transfer problem and the role of the asymptotic problem for (1.1) as $p \to \infty$ in the mass transfer problem.

In the Monge–Kantorovich case, i.e., the case where (1.5) holds, we have only two special results besides those in the cases when $n = 1$ or when Ω is an open ball and f is radial. One of them says that if Ω is symmetric with respect to the origin and f is an odd function, u_p converges in $C(\overline{\Omega})$, and the other roughly says that if the diameter of Ω is relatively small compared with support of f, $\operatorname{spt} f$, then the convergence of u_p in $C(\overline{\Omega})$ is valid.

The main results of this paper, concerned with convergence of u_p, are precisely stated in section 2. The proof of convergence in the one-dimensional case and the radial case are presented in sections 3 and 4, respectively. Section 5 is devoted to general properties of $\{u_p\}$, the set $\mathcal{M}$ of maximizers of the variational problem (1.4), the set $\mathcal{A}$ of the limits of u_p, i.e.,

$$\mathcal{A} = \{U \in C(\overline{\Omega}) \mid \text{there is a sequence } p_j \to \infty \text{ such that } u_{p_j} \to U \text{ in } C(\overline{\Omega})\}. \tag{1.6}$$

Section 6 is devoted to further properties of the set $\mathcal{M}$ which are useful in our study of convergence of u_p. These observations on $\mathcal{M}$ comprise the main results of this paper together with our results on the convergence of u_p.

We prove our convergence results in the case of no balanced family and in the vanishing integral case (the case of (1.5)), respectively, in sections 7 and 8.

We explain the notation in this paper. For $a, b \in \mathbf{R}$ we write $a \vee b = \max\{a,\, b\}$, $a \wedge b = \min\{a,\, b\}$, $a^+ = a \vee 0$, and $a^- = a \wedge 0$. We use the same notation for functions. We denote by $\mu(A)$ the Lebesgue measure of the measurable set $A \subset \mathbf{R}^n$. If needed, we denote by $\mu_n(A)$ in order to specify the dimension of the space where A lives. We denote by $B(x, a)$ the closed ball of radius a with x as its center.

Finally, we remark that most of the results in this paper have already been announced in [IL4].

2. Main results on the convergence. In this section we state our results concerning the limit, as $p \to \infty$, of the solution u_p of the Dirichlet problem

$$\begin{cases} -\Delta_p u(x) = f(x) & \text{in } \Omega, \\ u(x) = 0 & \text{for } x \in \partial\Omega. \end{cases} \tag{2.1}$$

Here, as before, $\Omega \subset \mathbf{R}^n$ is a bounded, open subset of $\mathbf{R}^n$ and $f \in C(\overline{\Omega})$.

To begin with, let us recall that the problem (2.1) has a unique solution $u_p \in W_0^{1,p}(\Omega)$ for any $p \in (1, \infty)$. See, e.g., [L].

Let $X = \{v \in C(\overline{\Omega}) \cap W^{1,\infty}(\Omega) \mid v|_{\partial\Omega} = 0,\ \|Dv\|_{L^\infty(\Omega)} \le 1\}$. We will recall that the family $\{u_p\}_{p>r}$ is bounded in $W^{1,q}(\Omega)$ for any $q > 1$ and $r > 1$, which guarantees that for any sequence $p_j \to \infty$, there is a subsequence $\{p_{j_k}\}_{k\in\mathbf{N}}$ such that $u_{p_{j_k}}$ converges to a function $U \in X$ uniformly in $\overline{\Omega}$ as $k \to \infty$.

We are interested in whether the following claim (C) is true or not:

(C) *The solution u_p converges uniformly to a function $U \in X$ as $p \to \infty$.*

We are not yet able to determine if the claim (C) is always true or not, and in what follows we present a couple of sufficient conditions for (C) to hold as our main results in this paper.

First we treat the case when $n = 1$. In this case we can not only show that (C) holds but also identify the limit, as the next theorem states.

Let $n = 1$ and $\Omega = (0, a)$, where $a > 0$ is a constant. We define the function $F \in C^1([0,\ a])$ by

$$F(x) = \int_0^x f(t)\, \mathrm{d}t.$$

We define

$$h(r) = \mu(\{x \in \Omega \mid F(x) < r\}), \quad \beta^* = \sup\left\{r \in \mathbf{R} \mid h(r) \le \frac{a}{2}\right\},$$
$$O_- = \{x \in \Omega \mid F(x) < \beta^*\},\ O_+ = \{x \in \Omega \mid F(x) > \beta^*\},\ O_0 = \{x \in \Omega \mid F(x) = \beta^*\},$$
$$k = \begin{cases} 0 & \text{if } \mu(O_0) = 0, \\ \frac{\mu(O_+) - \mu(O_-)}{\mu(O_0)} & \text{if } \mu(O_0) > 0. \end{cases}$$

Then we introduce the function $U \in C([0,\ a])$ by

$$U(x) = \int_0^x (\mathbf{1}_{O_-}(t) - \mathbf{1}_{O_+}(t) + k\mathbf{1}_{O_0}(t))\, \mathrm{d}t. \tag{2.2}$$

Here and henceforth $\mathbf{1}_A$ denotes the characteristic function of the set A. We will see in the next section (Lemma 3.5) that $|k| \le 1$, which assures that $U \in X$.

THEOREM 2.1. *If $n = 1$ and $\Omega = (0,\ a)$, then* (C) *holds and, moreover, the limit function U is given by* (2.2).

As above in the radial case we can show that (C) is valid and give an explicit formula for the limit function.

Let $a > 0$ be a constant and assume that $\Omega = \operatorname{int} B(0, a)$ and $f(x) = g(|x|)$ for some $g \in C([0, a])$.

We define $O_\pm \subset \mathbf{R}^n$ by

$$O_+ = \left\{t \in (0, a) \middle| \int_{B(0,t)} f(x)\, \mathrm{d}x > 0\right\}, \quad O_- = \left\{t \in (0, a) \middle| \int_{B(0,t)} f(x)\, \mathrm{d}x < 0\right\}$$

and $U \in X$ by

$$U(x) = \int_{|x|}^{a} (\mathbf{1}_{O_+}(t) - \mathbf{1}_{O_-}(t))\,\mathrm{d}t. \tag{2.3}$$

THEOREM 2.2. *If $\Omega = \operatorname{int} B(0,a)$ and $f(x) = g(|x|)$ is a radial function, then* (C) *holds and the limit function U is given by* (2.3).

The next condition under which (C) holds is a generalization of the well-known observation due to [BDM] and [J] (see Remark 5.2 of [BDM] and the uniqueness result of [J]) that if $f \geq 0$ in Ω, then (C) holds.

In order to make a precise statement, we need to introduce some notation.

We write

$$\Omega_+ = \{x \in \Omega \mid f(x) > 0\}, \quad \Omega_- = \{x \in \Omega \mid f(x) < 0\}, \quad \text{and} \quad \Omega_* = \Omega_+ \cup \Omega_-.$$

Note that $\operatorname{spt} f = \overline{\Omega}_*$. Let $\mathcal{O}_*$ denote the sets of all connected components of Ω_*.

We modify the notion of "connectedness" for a better formulation, as follows. Let $A, B \subset \mathbf{R}^n$. Define $\rho(A,B) \in [0,\infty]$ by setting

$$\rho(A,B) = \inf\{d(A,U_1) + d(U_1,U_2) + \cdots + d(U_m,B) \mid U_1,\ldots,U_m \in \mathcal{O}_*\},$$

where $d(U,V) = \inf\{|x-y| \mid x \in U, y \in V\}$. Notice that $\rho(A,B) = \infty$ if and only if either $A = \emptyset$ or $B = \emptyset$. Since, as is easily checked,

$$\rho(A,B) = \rho(B,A) \geq 0, \qquad \rho(A,B) \leq \rho(A,C) + \rho(C,B)$$

for any $A, B, C \subset \mathbf{R}^n$, if we write $A \sim B$ for $A, B \subset \mathbf{R}^N$ when $\rho(A,B) = 0$, then this relation $\sim$ defines an equivalence relation in $\mathcal{O}_*$.

Using the above equivalence relation, we classify $\mathcal{O}_*$ as

$$\mathcal{O}_* = \bigcup \{\mathcal{O}_\lambda \mid \lambda \in \Lambda\},$$

where

(i) for each $\lambda \in \Lambda$, $\mathcal{O}_\lambda \neq \emptyset$;
(ii) for each $\lambda \in \Lambda$, if $U \in \mathcal{O}_\lambda$, then $\mathcal{O}_\lambda = \{V \in \mathcal{O}_* \mid V \sim U\}$;
(iii) if $\lambda_1, \lambda_2 \in \Lambda$ and $\lambda_1 \neq \lambda_2$, then $\mathcal{O}_{\lambda_1} \cap \mathcal{O}_{\lambda_2} = \emptyset$.

We set

$$G_\lambda = \bigcup \{U \mid U \in \mathcal{O}_\lambda\} \quad \text{for } \lambda \in \Lambda$$

and define

$$\Lambda_0 = \{\lambda \in \Lambda \mid \rho(G_\lambda, \partial\Omega) = 0\}.$$

We note that $\{G_\lambda \mid \lambda \in \Lambda\}$ classifies the set Ω_*. Each G_λ, with $\lambda \in \Lambda$, is called an L-connected component of Ω_*.

As the proof of Lemma 7.1 below shows, if w is a Lipschitz continuous function on Ω and $Dw(x) = 0$ for almost every (a.e.) $x \in \Omega_*$, then w is constant on each G_λ, with $\lambda \in \Lambda$. Conversely, one can show the following: Let U, V be connected components of Ω_* having the property that if w is Lipschitz continuous on Ω and $Dw(x) = 0$ for a.e. $x \in \Omega_*$, then w is constant on $U \cup V$. Then $U \sim V$, i.e., U, V are subsets of an L-connected component G_λ. In light of these observations, we have chosen the term "Lipschitz-connected" (L-connected).

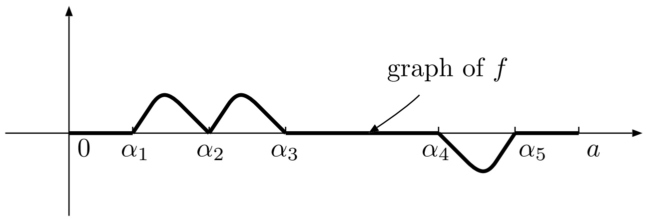


FIG. 1.

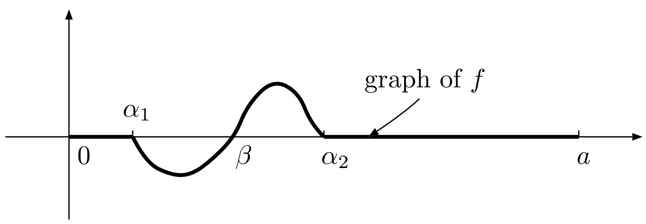


FIG. 2.

Our assumption on (f,Ω) is as follows:

$$\text{For any nonempty } \Gamma \subset \Lambda \setminus \Lambda_0 \text{ and } \omega := \bigcup\{G_\lambda \mid \lambda \in \Gamma\}, \tag{2.4}$$

$$\int_\omega f(x)\,\mathrm{d}x \neq 0.$$

We call this condition that of *no balanced family* (of L-connected components).

Figure 1 gives pictorially an example of a function f which satisfies condition (2.4). Here $\int_{\alpha_1}^{\alpha_2} f(x)\,\mathrm{d}x = \int_{\alpha_2}^{\alpha_3} f(x)\,\mathrm{d}x = -\int_{\alpha_4}^{\alpha_5} f(x)\,\mathrm{d}x$ is assumed. In this example, $\Omega_+ = (\alpha_1,\alpha_2)\cup(\alpha_2,\alpha_3)$, $\Omega_- = (\alpha_4,\alpha_5)$, and the L-connected components are Ω_+ and Ω_-. The integral of f over $\omega = \Omega_+$, Ω_-, or $\Omega_+ \cup \Omega_-$ does not vanish. On the other hand, the connected components of Ω_* are (α_1,α_2), (α_2,α_3), and (α_4,α_5), and the integral of f over $\omega = (\alpha_1,\alpha_2)\cup(\alpha_4,\alpha_5)$ vanishes. For this f, the condition similar to (2.4) but with the usual notion of connectedness in place of that of L-connectedness does not hold.

Next, we examine the function f given pictorially by Figure 2, where $\int_{\alpha_1}^{\alpha_2} f(x)\,\mathrm{d}x = 0$ is assumed. For this function f, $\Omega_* = (\alpha_1,\alpha_2)$ is the only L-connected component of Ω_*, and condition (2.4) does not hold. This function f will appear in Example 3.2 in section 3.

THEOREM 2.3. *Under the assumption* (2.4)*,* (C) *holds.*

Regarding the cases when (2.4) is violated, we restrict ourselves to the case where

$$\int_\Omega f(x)\,\mathrm{d}x = 0 \quad \text{and} \quad f \neq 0, \tag{2.5}$$

and

$$\Omega_+ \text{ and } \Omega_- \text{ are connected.} \tag{2.6}$$

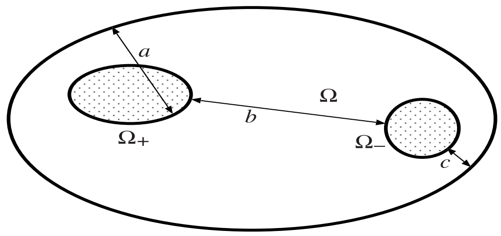


FIG. 3. *A case where* $b \geq a + c$.

We give two sufficient conditions for (C) to hold. One is a symmetry requirement on (f, Ω). That is, we assume that

$$\Omega \text{ is symmetric with respect to the origin, i.e., } -\Omega = \Omega, \tag{2.7}$$

and

$$f \text{ is an odd function, i.e., } f(-x) = -f(x) \quad \text{for all } x \in \Omega. \tag{2.8}$$

The asymptotic problem, as $p \to \infty$, for (1.1) has applications to the Monge–Kantorovich mass transfer problem. In the mass transfer problem, the condition of vanishing integral, (2.5), is a natural compatibility condition, which means conservation of the total mass in the process of mass transfer.

The second one is the condition that

$$\min\left\{ \inf_{x \in \Omega_+} \sup_{y \in \Omega_-} [d(x) + d(y) - |x - y|],\ \inf_{y \in \Omega_-} \sup_{x \in \Omega_+} [d(x) + d(y) - |x - y|] \right\} \leq 0 \tag{2.9}$$

holds. Here and henceforth $d(x)$ denotes the distance between x and $\partial\Omega$, i.e., $d(x) = \operatorname{dist}(x, \partial\Omega)$. See Figure 3.

THEOREM 2.4. *Under the assumptions* (2.5) *and* (2.6), *if either* (2.7) *and* (2.8) *or* (2.9) *are satisfied, then* (C) *holds.*

3. One-dimensional case. In this section we prove Theorem 2.1.

Let $\Omega = (0, a)$, where $a > 0$ is a constant, and $f \in C([0, a])$. Fix $p > 1$ and consider the $(p+1)$-Laplace equation with the inhomogeneous term f,

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(|u'(x)|^{p-1}u'(x)\right) = -f(x) \quad \text{in } \Omega, \tag{3.1}$$

together with the Dirichlet condition

$$u(0) = u(a) = 0. \tag{3.2}$$

Here u' denotes the derivative of u. The unique solution in $W_0^{1,p+1}(\Omega)$ of (3.1) (i.e., the solution of (3.1)–(3.2)) is denoted by u_{p+1} as in the previous sections.

We seek an explicit formula for u_{p+1}. For this, noting that $u := u_{p+1} \in C^1([0, a])$ and integrating both sides of (3.1), we get

$$|u'(x)|^{p-1}u'(x) = |u'(0)|^{p-1}u'(0) - F(x) \quad \text{for } x \in \Omega,$$

where $F(x) := \int_0^x f(t)\,\mathrm{d}t$.

Let ψ_p be the inverse function of $r \mapsto |r|^{p-1}r$. That is, $\psi_p(s) = |s|^{\frac{1}{p}-1}s$ for $s \in \mathbf{R}$. Note that as $p \to \infty$,

$$\psi_p(r) \to \begin{cases} 1 & \text{for } r > 0, \\ -1 & \text{for } r < 0. \end{cases}$$

Observe, moreover, that for any $\varepsilon \in (0,1)$, the above convergence is uniform for $|r| \in [\varepsilon, \varepsilon^{-1}]$.

Writing $\beta = |u'(0)|^{p-1}u'(0)$ and integrating the equality $u'(x) = \psi_p(\beta - F(x))$, we get

$$u(x) = \int_0^x \psi_p(\beta - F(t))\,\mathrm{d}t \quad \text{for } x \in \overline{\Omega}. \tag{3.3}$$

Conversely, if we can choose $\beta \in \mathbf{R}$ so that

$$\int_0^a \psi_p(\beta - F(t))\,\mathrm{d}t = 0,$$

then the function u defined by (3.3) is in $C^1([0,\,a])$ and the unique solution of (3.1)–(3.2).

We show directly that there is a unique constant $\beta_p \in \mathbf{R}$ such that

$$\int_0^a \psi_p(\beta_p - F(t))\,\mathrm{d}t = 0, \tag{3.4}$$

although this can be deduced from the general existence and uniqueness result for solutions of (1.1).

Set

$$G_p(r) = \int_0^a \psi_p(r - F(t))\,\mathrm{d}t$$

for $r \in \mathbf{R}$. Since the function $\psi_p(r)$ is strictly increasing, the function G_p is strictly increasing on $\mathbf{R}$. In view of the monotone convergence theorem, we see that the function G_p is continuous on $\mathbf{R}$. If $f = 0$, then it is clear that $\beta_p = 0$ gives the unique solution of (3.4).

We may thus assume in what follows that $f \neq 0$. We set

$$F_- = \min_{[0,a]} F, \quad F_+ = \max_{[0,a]} F, \quad \delta(F) = F_+ - F_-. \tag{3.5}$$

Note that $F_- \le 0 \le F_+$ and $\delta(F) > 0$. Since $F_- - F(x) \le 0$ for all $x \in \Omega$ and $\delta(F) > 0$, we have $G_p(F_-) < 0$. Similarly, we have $G_p(F_+) > 0$. Thus we see that there is a unique constant $\beta_p \in (F_-, F_+)$ such that $G_p(\beta_p) = 0$, and we find an explicit formula

$$u_{p+1}(x) = \int_0^x \psi_p(\beta_p - F(t))\,\mathrm{d}t \quad \text{for } x \in \overline{\Omega}. \tag{3.6}$$

Next, we study the asymptotic behavior of the function u_{p+1} given by (3.6) as $p \to \infty$. Recall that

$$h(r) = \mu(\{x \in \Omega \mid F(x) < r\}), \qquad \beta^* = \sup\left\{r \in \mathbf{R} \mid h(r) \le \frac{a}{2}\right\},$$
$$O_- = \{x \in \Omega \mid F(x) < \beta^*\},\ \ O_+ = \{x \in \Omega \mid F(x) > \beta^*\},\ \ O_0 = \{x \in \Omega \mid F(x) = \beta^*\}.$$

We define open sets $O(r) \subset \mathbf{R}$ for $r \in \mathbf{R}$ by

$$O(r) = \{x \in \Omega \mid F(x) < r\}.$$

We have (i) for $r \le s$, $O(r) \subset O(s)$; (ii) $O(F_-) = \emptyset$; (iii) $O(r) = (0, a)$ for all $r > F_+$; and (iv)

$$\bigcup_{t<r} O(t) = O(r), \qquad \bigcap_{t>r} O(t) = \{x \in \Omega \mid F(x) \le r\}.$$

Consequently, we have (i) h is nondecreasing in $\mathbf{R}$; (ii) $h(r)=\mu(\emptyset)=0$ for $r\in(-\infty, F_-]$; (iii) $h(r) = \mu(\Omega) = a$ for $r \in (F_+, \infty)$; and (iv)

$$\lim_{t \nearrow r} h(t) = h(r) \le \mu(\{x \in \Omega \mid F(x) \le r\}) = \lim_{t \searrow r} h(t).$$

Now, property (iv) for h and the definition of β^* implies that

$$h(\beta^*) \le \frac{a}{2} \le h(\beta^* + 0) := \lim_{t \searrow \beta^*} h(t).$$

A key step in the proof of Theorem 2.1 is the following lemma.

LEMMA 3.1. *We have*

$$\lim_{p \to \infty} \beta_p = \beta^*.$$

We use three lemmas to prepare for the proof of Lemma 3.1.

LEMMA 3.2. *Let $r \in (F_-, F_+)$. Then, if $s > r$ (resp., $s < r$), then $h(s) > h(r)$ (resp., $h(s) < h(r)$).*

Proof. We consider the case when $s > r$. We may assume that $s \in (F_-, F_+)$. By the intermediate value theorem, we have

$$F(y) = \frac{s + r}{2}$$

for some $y \in \overline{\Omega}$. By the continuity of F, we can choose $\delta > 0$ so that

$$F(x) \in (r, s) \quad \text{ for all } x \in \omega := (y - \delta, y + \delta) \cap \Omega.$$

It is clear that $\omega \subset O(s)$, $\mu(\omega) > 0$, and $\omega \cap O(r) = \emptyset$. Hence we have

$$h(s) = \mu(O(s)) = \mu(O(r)) + \mu(O(s) \setminus O(r)) \ge \mu(O(r)) + \mu(\omega) > \mu(O(r)) = h(r).$$

The proof for the case when $s < r$ is similar and will be omitted. □

LEMMA 3.3. *Let $\beta \in [F_-, F_+]$. We have*

$$|\psi_p(\beta - F(x))| \le \psi_p(\max\{\delta(F), 1\}) \le \psi_1(\max\{\delta(F), 1\}) \quad \textit{for } x \in \Omega.$$

Proof. For $x \in \Omega$ we have

$$-\delta(F) = F_- - F_+ \le \beta - F(x) \le F_+ - F_- = \delta(F),$$

and hence $|\psi_p(\beta - F(x))| \le \psi_p(\max\{\delta(F), 1\}) \le \psi_1(\max\{\delta(F), 1\})$. □

Lemma 3.4. *Let $\{\alpha_j\} \subset [F_-, F_+]$ be a sequence converging to some $r \in \mathbf{R}$, let $\{p_j\} \subset (1,\infty)$ be a sequence such that $p_j \to \infty$ as $j \to \infty$, and let $\phi \in L^1(\Omega)$. Set*

$$O_-(r) = \{x \in \Omega \mid F(x) < r\} \quad \text{and} \quad O_+(r) = \{x \in \Omega \mid F(x) > r\}.$$

Then

$$\int_{O_-(r)} \psi_{p_j}(\alpha_j - F(x))\phi(x)\,\mathrm{d}x \to \int_{O_-(r)} \phi(x)\,\mathrm{d}x,$$
$$\int_{O_+(r)} \psi_{p_j}(\alpha_j - F(x))\phi(x)\,\mathrm{d}x \to -\int_{O_+(r)} \phi(x)\,\mathrm{d}x.$$

Proof. Fix $x \in O_-(r)$. Since $r - F(x) > \delta$ for some constant $\delta > 0$, there is a $J \in \mathbf{N}$ such that for all $j \geq J$,

$$\delta < \alpha_j - F(x) \leq \delta(F),$$

which implies that

$$\lim_{j\to\infty} \psi_{p_j}(\alpha_j - F(x)) = 1.$$

Now, in view of Lemma 3.3 and the Lebesgue convergence theorem, we conclude that

$$\lim_{j\to\infty} \int_{O_-(r)} \psi_{p_j}(\alpha_j - F(x))\phi(x)\,\mathrm{d}x = \int_{O_-(r)} \phi(x)\,\mathrm{d}x.$$

In the same way we see that

$$\lim_{j\to\infty} \int_{O_+(r)} \psi_{p_j}(\alpha_j - F(x))\phi(x)\,\mathrm{d}x = -\int_{O_+(r)} \phi(x)\,\mathrm{d}x. \qquad \square$$

Proof of Lemma 3.1. First of all we show that $\liminf_{p\to\infty} \beta_p \geq \beta^*$. For this, we argue by contradiction and thus suppose that $r := \liminf_{p\to\infty} \beta_p < \beta^*$. There is a sequence $\{p_j\} \subset (1,\infty)$ such that $\lim_{j\to\infty} p_j = \infty$ and $\lim_{j\to\infty} \beta_{p_j} = r$. By Lemma 3.2, we see that $h(r+0) < h(\beta^*) \leq a/2$. Therefore, setting $A = \{x \in \Omega \mid F(x) \leq r\}$ and $B = \{x \in \Omega \mid F(x) > r\}$, we have

$$h(r+0) = \mu(A) < \frac{a}{2}, \qquad \mu(B) = a - \mu(A) > \frac{a}{2}.$$

Since $|\psi_p(\beta_p - F(x))| \leq \psi_p(\max\{\delta(F), 1\})$ for $x \in \Omega$ by Lemma 3.3, we have

$$\begin{aligned}&\limsup_{j\to\infty} \left| \int_A \psi_{p_j}(\beta_{p_j} - F(x))\,\mathrm{d}x \right| \\ &\quad\leq \limsup_{j\to\infty} \int_A \psi_{p_j}(\max\{\delta(F), 1\})\,\mathrm{d}x = \int_A \mathrm{d}x = \mu(A) < \frac{a}{2}.\end{aligned}$$

Observe next by Lemma 3.4 that

$$\lim_{j\to\infty} \int_B \psi_{p_j}(\beta_{p_j} - F(x))\,\mathrm{d}x = -\mu(B) < -\frac{a}{2}.$$

Since

$$0 = \int_A \psi_{p_j}(\beta_{p_j} - F(x))\,\mathrm{d}x + \int_B \psi_{p_j}(\beta_{p_j} - F(x))\,\mathrm{d}x,$$

we find that $0 < \frac{a}{2} - \frac{a}{2} = 0$, which is a contradiction. This shows that $\liminf_{p\to\infty} \beta_p \geq \beta^*$.

An argument similar to the above shows that $\limsup_{p\to\infty} \beta_p \leq \beta^*$, and we conclude that $\lim_{p\to\infty} \beta_p = \beta^*$. ∎

Recall that $k = 0$ if $\mu(O_0) = 0$, and otherwise

$$k = \frac{\mu(O_+) - \mu(O_-)}{\mu(O_0)}.$$

LEMMA 3.5. *We have* $|k| \leq 1$.

Proof. Since

$$\mu(O_-) = h(\beta^*) \leq \frac{a}{2} \leq h(\beta^* + 0) = \mu(O_-) + \mu(O_0),$$

we have

$$\begin{aligned} &0 \geq 2\mu(O_-) - a = \mu(O_-) - \mu(O_+) - \mu(O_0), \\ &0 \leq 2\mu(O_-) + 2\mu(O_0) - a = \mu(O_-) + \mu(O_0) - \mu(O_+). \end{aligned}$$

Hence, if $k \neq 0$, then

$$-1 \leq k = \frac{\mu(O_+) - \mu(O_-)}{\mu(O_0)} \leq 1. \qquad ∎$$

Proof of Theorem 2.1. We show first that the family $\{u_p\}_{p>2}$ is uniformly bounded and equicontinuous on $\overline{\Omega}$.

To see this, fix $x \in \Omega$ and $p > 1$. By Lemma 3.3, we have

$$|u_{p+1}(x)| \leq \int_0^x \psi_1(\max\{\delta(F), 1\})\,\mathrm{d}x \leq a\psi_1(\max\{\delta(F), 1\})$$

and $|u'_{p+1}(x)| \leq \psi_1(\max\{\delta(F), 1\})$. These show that the family $\{u_p\}_{p>2}$ is uniformly bounded and equicontinuous on $\overline{\Omega}$.

Next we show that

$$\lim_{p\to\infty} \psi_p(\beta_p - \beta^*) = k \quad \text{ if } \mu(O_0) > 0. \tag{3.7}$$

In fact, we have

$$\begin{aligned} 0 = \int_0^a \psi_p(\beta_p - F(x))\mathrm{d}x &= \int_{O_-} \psi_p(\beta_p - F(x))\mathrm{d}x \\ &\quad + \psi_p(\beta_p - \beta^*)\mu(O_0) + \int_{O_+} \psi_p(\beta_p - F(x))\mathrm{d}x, \end{aligned}$$

and then Lemma 3.4 yields

$$0 = \mu(O_-) - \mu(O_+) + \lim_{p\to\infty} \psi_p(\beta_p - \beta^*)\mu(O_0),$$

which shows (3.7).

Since $\{u_p\}_{p>2}$ is precompact in $C([0, a])$, we need to show only that for each fixed $x \in \Omega$,

$$u_{p+1}(x) \to U(x) := \int_0^x (\mathbf{1}_{O_-}(t) - \mathbf{1}_{O_+}(t) + k\mathbf{1}_{O_0}(t))\,\mathrm{d}t \quad \text{ as } p \to \infty.$$

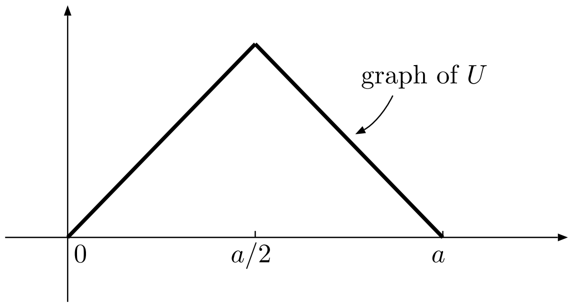


FIG. 4. *A case where* $f > 0$.

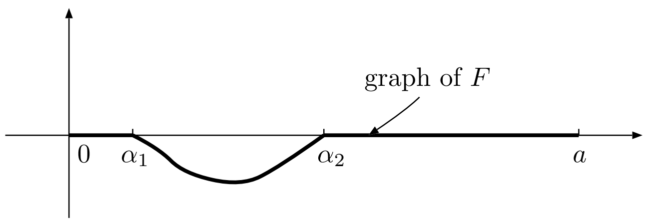


FIG. 5.

Fix $x \in \Omega$ and note that

$$u_{p+1}(x) = \int_{(0,x)\cap O_-} \psi_p(\beta_p - F(t))\,\mathrm{d}t + \int_{(0,x)\cap O_+} \psi_p(\beta_p - F(t))\,\mathrm{d}t + \psi_p(\beta_p - \beta^*) \int_{(0,x)\cap O_0} \mathrm{d}t.$$

Sending $p \to \infty$ and using Lemma 3.4, we get

$$\lim_{p\to\infty} u_p(x) = \int_{(0,x)\cap\Omega_-} \mathrm{d}t - \int_{(0,x)\cap\Omega_+} \mathrm{d}t + k \int_{(0,x)\cap\Omega_0} \mathrm{d}t = U(x). \qquad \square$$

Next we examine the limit function U in a few cases.

Example 3.1. We consider the case when $f(x) > 0$ for all $x \in \Omega$. Then the function F is strictly increasing in $\overline{\Omega}$. Therefore we have $O_- = (0, a/2)$ and $O_+ = (a/2, a)$, and hence

$$U(x) = \int_0^x \left(\mathbf{1}_{O_-}(t) - \mathbf{1}_{O_+}(t)\right) dt = \begin{cases} x & \text{for } 0 \le x \le a/2, \\ a - x & \text{for } a/2 \le x \le a. \end{cases}$$

This is the distance function from $\partial\Omega = \{0, a\}$ and, as is well known, it is the unique viscosity solution of $|U'(x)| = 1$ in Ω and $U(0) = U(a) = 0$. See Figure 4.

Example 3.2. Let $0 < \alpha_1 < \alpha_2 < a$ satisfy $\alpha_2 - \alpha_1 < \frac{a}{2}$. Let F satisfy $F(x) = 0$ for $x \in [0, \alpha_1] \cup [\alpha_2, a]$ and $F(x) < 0$ for $x \in (\alpha_1, \alpha_2)$. (See Figure 5.) Then we have $\beta^* = 0$, $O_- = (\alpha_1, \alpha_2)$, $O_+ = \emptyset$, and $O_0 = (0, \alpha_1) \cup (\alpha_2, a)$. Furthermore, we have $k = -(\alpha_2 - \alpha_1)/(a - (\alpha_2 - \alpha_1))$, and

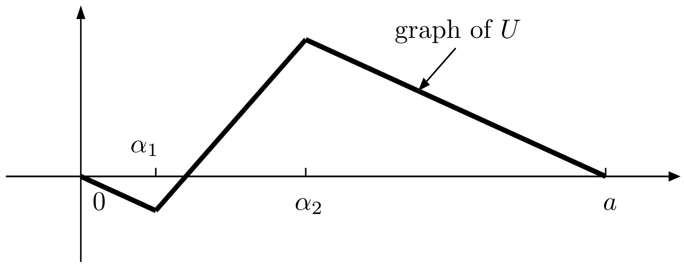


FIG. 6. *A case where* $|k| < 1$.

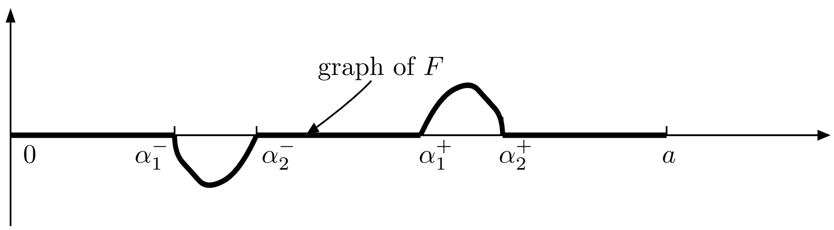


FIG. 7.

$$U(x) = \int_0^x \big(\mathbf{1}_{O_-}(t) + k\mathbf{1}_{O_0}(t)\big)\,dt = \begin{cases} kx & \text{for } 0 \le x \le \alpha_1, \\ k\alpha_1 + x - \alpha_1 & \text{for } \alpha_1 \le x \le \alpha_2, \\ k(x-a) & \text{for } \alpha_2 \le x \le a. \end{cases}$$

See Figure 6.

Example 3.3. Let $0 < \alpha_1^- < \alpha_2^- < \alpha_1^+ < \alpha_2^+ < a$ satisfy $\alpha_2^- - \alpha_1^- = \alpha_2^+ - \alpha_1^+ < \frac{a}{2}$. Let F satisfy the following: $F(x) = 0$ for $x \in [0, \alpha_1^-] \cup [\alpha_2^-, \alpha_1^+] \cup [\alpha_2^+, a]$, $F(x) < 0$ for $x \in (\alpha_1^-, \alpha_2^-)$, and $F(x) > 0$ for $x \in (\alpha_1^+, \alpha_2^+)$. (See Figure 7.) Then we have $\beta^* = 0$ and $k = (\alpha_2^+ - \alpha_1^+ - (\alpha_2^- - \alpha_1^-))/(a - (\alpha_2^+ - \alpha_1^+) - (\alpha_2^- - \alpha_1^-)) = 0$, and the limit function U is given by

$$U(x) = \begin{cases} 0 & \text{for } x \in [0, \alpha_1^-] \cup [\alpha_2^+, a], \\ x - \alpha_1^+ & \text{for } x \in (\alpha_1^-, \alpha_2^-), \\ \alpha_2^- - \alpha_1^- & \text{for } x \in [\alpha_2^-, \alpha_1^+], \\ -x + \alpha_2^+ & \text{for } x \in (\alpha_1^+, \alpha_2^+). \end{cases}$$

See Figure 8.

4. Radial case. In this section we give a proof of Theorem 2.2, which is rather close to that of Theorem 2.1 presented in the previous section.

Let $a > 0$ and $g \in C([0, a])$, and define $f \in C(B(0, a))$ by $f(x) = g(|x|)$. Set $\Omega = \operatorname{int} B(0, a)$.

We consider the Dirichlet problem for u_{p+1} as in the previous section:

$$\begin{cases} -\Delta_{p+1} u(x) = f(x) & \text{in } \Omega, \\ u(x) = 0 & \text{on } \partial\Omega. \end{cases} \tag{4.1}$$

By the uniqueness of the solution of (4.1), we see that the function u_{p+1} is a radial function, i.e., $u_{p+1}(x) = v_p(|x|)$ for some $v_p \in C([0, a])$. By the regularity results for

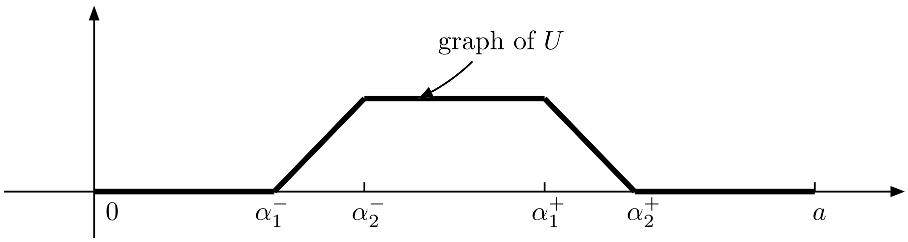


FIG. 8. *A case where* $k = 0$.

(4.1), we know that $u_{p+1} \in C^{1,\gamma}(\overline{\Omega})$ for some $\gamma \in (0, 1)$. In particular, we have

$$v_p \in C^1([0, a]), \quad v_p'(0) = 0.$$

The PDE (4.1) is now reduced to the following ODE for v_p:

$$\big(r^{n-1}|v'(r)|^{p-1}v'(r)\big)' = -r^{n-1}g(r) \quad \text{in } (0, a); \tag{4.2}$$

and the boundary condition for v_p is $v'(0) = v(a) = 0$. Integrating twice yields

$$v(r) = \alpha - \int_0^r \psi_p\Big(t^{1-n}\int_0^t s^{n-1}g(s)\,\mathrm{d}s\Big)\,\mathrm{d}t \quad \text{for all } r \in [0,\, a]$$

and for some $\alpha \in \mathbf{R}$, where $\psi_p \in C(\mathbf{R})$ is the function given by $\psi_p(s) = |s|^{\frac{1}{p}-1}s$ as in the previous section. Here the constant α for $v = v_p$ should be determined by

$$\alpha = \int_0^a \psi_p\Big(t^{1-n}\int_0^t s^{n-1}g(s)\,\mathrm{d}s\Big)\,\mathrm{d}t.$$

Setting

$$\alpha_p = \int_0^a \psi_p\Big(t^{1-n}\int_0^t s^{n-1}g(s)\,\mathrm{d}s\Big)\,\mathrm{d}t,$$

we have

$$v_p(r) = \alpha_p - \int_0^r \psi_p\Big(t^{1-n}\int_0^t s^{n-1}g(s)\,\mathrm{d}s\Big)\,\mathrm{d}t \quad \text{for } r \in [0, r].$$

At this point one can check directly and easily that $v_p \in C^1([0, a])$, and it satisfies (4.2) and the boundary condition $v_p'(0) = v_p(a) = 0$.

Completion of the proof of Theorem 2.2. It is easy to see that as $p \to \infty$,

$$\alpha_p \to \alpha^* := \int_0^a \big(\mathbf{1}_{O_+}(r) - \mathbf{1}_{O_-}(r)\big)\mathrm{d}r,$$

and

$$v_p(r) \to V(r) := \alpha^* + \int_0^r \big(\mathbf{1}_{O_-}(t) - \mathbf{1}_{O_+}(t)\big)\,\mathrm{d}t \quad \text{for each } r \in [0, a],$$

where

$$O_+ = \left\{ t \in (0,a) \,\middle|\, \int_0^t s^{n-1} g(s)\,\mathrm{d}s > 0 \right\} = \left\{ t \in (0,a) \,\middle|\, \int_{B(0,t)} f(x)\,\mathrm{d}x > 0 \right\},$$
$$O_- = \left\{ t \in (0,a) \,\middle|\, \int_0^t s^{n-1} g(s)\,\mathrm{d}s < 0 \right\} = \left\{ t \in (0,a) \,\middle|\, \int_{B(0,t)} f(x)\,\mathrm{d}x < 0 \right\}.$$

As in the previous section, it is easy to show that the collection of functions $v_p(|x|)$, with $p > 1$, is precompact in $C(\overline{\Omega})$. Thus the above pointwise convergence is enough for us to conclude that $u_p(x)$ converges to $U(x) := V(|x|)$ uniformly for $x \in \overline{\Omega}$ as $p \to \infty$. □

Remark. Contrary to the general one-dimensional case, the limit function V has the property that $V'(r) \in \{-1, 0, 1\}$ for all $r \in [0, a]$.

Remark. We also have a convergence result in the case when Ω is an annulus and f is radial. Indeed, let $0 < r_1 < r_2$, $\Omega = \{x \in \mathbf{R}^n \mid r_1 < |x| < r_2\}$, and $f(x) = g(|x|)$ for some $g \in C([r_1, r_2])$. Let u_p be the solution of (1.1). We define, for $r \in \mathbf{R}$,

$$G(r) = \int_{r_1}^{r} t^{n-1} g(t)\,\mathrm{d}t \quad \text{for } r \in [r_1, r_2], \quad h(r) = \mu_1(\{t \in (r_1,\, r_2) \mid G(t) < r\})$$

$$\beta^* = \sup\left\{ t \in (r_1,\, r_2) \middle| h(r) \le \frac{r_2 - r_1}{2} \right\},\ \ O_+ = \{r \in (r_1,\, r_2) \mid G(r) > \beta^*\},$$

$$O_- = \{r \in (r_1,\, r_2) \mid G(r) < \beta^*\},\ \ O_0 = \{r \in (r_1,\, r_2) \mid G(r) = \beta^*\},$$

$$k = \begin{cases} 0 & \text{for } \mu_1(O_0) = 0, \\ \frac{\mu_1(O_+) - \mu_1(O_-)}{\mu_1(O_0)} & \text{otherwise,} \end{cases}$$

$$U(x) = \int_{r_1}^{|x|} \big(\mathbf{1}_{O_-}(t) - \mathbf{1}_{O_+}(t) + k\mathbf{1}_{O_0}(t)\big)\,\mathrm{d}t \quad \text{for } x \in \overline{\Omega}.$$

Then we have

$$u_p \to U \quad \text{in } C(\overline{\Omega}) \quad \text{as } p \to \infty.$$

We do not give the proof of this result here since it is a simple combination of the proofs of Theorems 2.1 and 2.2.

5. General observations. Here we study a few general properties of the solution u_p of (1.1), the set $\mathcal{A}$ of the limits of u_p defined by (1.6), and the set $\mathcal{M}$ of the maximizers of the variational problem (1.4), i.e.,

$$\mathcal{M} = \Big\{ v \in X \mid I_\infty(v) = \sup_{u \in X} I_\infty(u) \Big\}.$$

We start by observing that the estimate

$$\|Du_p\|_{L^p(\Omega)} \le C \tag{5.1}$$

holds, where the constant C can be chosen independently of p for $p > 2$. Indeed, using the test function $u = u_p$ in the weak formulation of (1.1), we get

$$\int_\Omega |Du|^p \,\mathrm{d}x = \int_\Omega f u \,\mathrm{d}x,$$

and hence by the Poincaré inequality for functions in $W_0^{1,1}(\Omega)$,

$$\begin{aligned}\int_\Omega |Du|^p \,\mathrm{d}x &\le \|f\|_{L^\infty(\Omega)} \|u\|_{L^1(\Omega)} \le C_1 \|f\|_{L^\infty(\Omega)} \|Du\|_{L^1(\Omega)} \\ &\le C_1 \|f\|_{L^\infty(\Omega)} \mu(\Omega) \int_\Omega |Du| \frac{\mathrm{d}x}{\mu(\Omega)} \le C_1 \|f\|_{L^\infty(\Omega)} \mu(\Omega) \left(\int_\Omega |Du|^p \frac{\mathrm{d}x}{\mu(\Omega)} \right)^{\frac{1}{p}} \\ &\le C_1 \|f\|_{L^\infty(\Omega)} \mu(\Omega)^{1-\frac{1}{p}} \|Du\|_{L^p(\Omega)},\end{aligned}$$

where C_1 is a positive constant independent of p. Hence, we obtain

$$\|Du\|_{L^p(\Omega)} \le \left(C_1 \|f\|_{L^\infty(\Omega)} \right)^{\frac{1}{p-1}} \mu(\Omega)^{\frac{1}{p}},$$

which shows (5.1).

From the above estimate (5.1), we have the following well-known observations (see [BDM], for instance).

PROPOSITION 5.1. (i) *For any $q > n$, the collection $\{u_p\}_{p \ge q}$ is precompact in $C(\overline{\Omega})$. In particular, for any sequence $1 < p_k \to \infty$ there is a subsequence p_{k_j} such that $u_{p_{k_j}}(x) \to U(x)$ uniformly on $\overline{\Omega}$ for some $U \in C(\overline{\Omega})$.*

(ii) *Let $U \in C(\overline{\Omega})$ be as above. Then $U \in W^{1,\infty}(\Omega)$ and $|DU(x)| \le 1$ for a.e. $x \in \Omega$.*

Proof. We first show (i). For $p \ge q$, we have

$$\|Du_p\|_{L^q(\Omega)} \le \mu(\Omega)^{\frac{1}{q} - \frac{1}{p}} \|Du_p\|_{L^p(\Omega)}. \tag{5.2}$$

For $q > n$, by the Sobolev embedding theorem (see, e.g., [GT]), we have

$$\|u_p\|_{C^{0,\gamma}(\overline{\Omega})} \le C_q \|Du\|_{L^q(\Omega)}$$

for some constants $\gamma \in (0, 1)$ and $C_q > 0$. These together with (5.1) imply that for any $q > n$, the collection $\{u_p\}_{p \ge q}$ is precompact.

Next, we prove (ii). The estimates (5.1) and (5.2) and the weak compactness of the balls in $W_0^{1,q}(\Omega)$, with $1 < q < \infty$, guarantee that $U \in W_0^{1,q}(\Omega)$ for any $q \in (1, \infty)$. This weak compactness, (5.1), and (5.2) yield

$$\|DU\|_{L^q(\Omega)} \le \mu(\Omega)^{\frac{1}{q}} \quad \text{for any } q > 1,$$

which implies that $|DU(x)| \le 1$ almost everywhere in Ω. ∎

Recalling the definition (1.6) of the set $\mathcal{A}$, from Proposition 5.1 we immediately have the following proposition.

PROPOSITION 5.2. (i) *$\mathcal{A} \ne \emptyset$ and $\mathcal{A} \subset X$.* (ii) *$u_p \to U$ in $C(\overline{\Omega})$ as $p \to \infty$ if and only if $\mathcal{A} = \{U\}$.*

Next, we consider the functional $I_\infty(u)$ for $u \in X$ defined by (1.4) and study the set $\mathcal{M}$ of maximizers of this functional.

The following proposition states a basic relation between $\mathcal{A}$ and $\mathcal{M}$.

PROPOSITION 5.3. (i) $\mathcal{A} \subset \mathcal{M}$. (ii) *As* $p \to \infty$, $I_p(u_p) \to \sup_{u\in X} I_\infty(u)$.

Proof. Let $U \in \mathcal{A}$ and $p_j \to \infty$ be such that $u_{p_j} \to U$ in $C(\overline{\Omega})$ as $j \to \infty$. As $p = p_j \to \infty$, we have

$$I_p(u_p) = I_\infty(u_p) - \frac{1}{p}\int_\Omega |Du_p|^p \, dx \le I_\infty(u_p) \to I_\infty(U).$$

Fix any $V \in X$ and observe that as $p \to \infty$,

$$I_p(u_p) \ge I_p(V) = I_\infty(V) - \frac{1}{p}\int_\Omega |DV(x)|^p \, dx \to I_\infty(V).$$

Hence we get

$$I_\infty(U) \ge \limsup_{j\to\infty} I_{p_j}(u_{p_j}) \ge \liminf_{p\to\infty} I_p(u_p) \ge I_\infty(V).$$

Since $U \in X$ by Proposition 5.2, we thus conclude that

$$I_\infty(U) = \sup_{u\in X} I_\infty(u), \qquad \lim_{j\to\infty} I_{p_j}(u_{p_j}) = \sup_{u\in X} I_\infty(u),$$

and $\mathcal{A} \subset \mathcal{M}$. Using (i) of Proposition 5.1, we deduce that

$$I_p(u_p) \to \sup_{u\in X} I_\infty(u) \quad \text{ as } p \to \infty. \qquad \square$$

PROPOSITION 5.4. *If* $u \in \mathcal{A}$, *then* u *satisfies*

$$-\Delta_\infty u(x) \le 0 \quad \text{ in } \Omega \setminus \overline{\Omega}_+ \quad \text{ and } \quad -\Delta_\infty u(x) \ge 0 \quad \text{ in } \Omega \setminus \overline{\Omega}_-$$

in the viscosity sense.

Proof. We prove only the first inequality, as the proof of the other inequality is similar. We set $W = \Omega \setminus \overline{\Omega}_+$. Let $\varphi \in C^2(W)$ and $\hat{x} \in W$. We assume that $u - \varphi$ attains a strict maximum at $\hat{x}$ and will show that $-\Delta_\infty\varphi(\hat{x}) \le 0$. For this, we argue by contradiction, and hence we assume that $-\Delta_\infty\varphi(\hat{x}) > 0$. Here we may assume that $u(\hat{x}) = \varphi(\hat{x})$.

Since $u \in \mathcal{A}$, there is a sequence $1 < p_j \to \infty$ such that $u_{p_j} \to u$ in $C(\overline{\Omega})$ as $j \to \infty$. Fix an $r > 0$ so that $B(\hat{x}, r) \subset W$ and $\Delta_\infty\varphi(x) < 0$ for all $x \in B(\hat{x}, r)$. Since

$$\Delta_p\varphi(x) = |D\varphi(x)|^{p-4}(|D\varphi(x)|^2\Delta\varphi(x) + (p-2)\Delta_\infty\varphi(x)) \quad \text{ for } x \in B(\hat{x}, r),$$

and

$$\min_{B(\hat{x},\,r)} |D\varphi| > 0 \quad \text{ and } \quad \max_{B(\hat{x},\,r)} \Delta_\infty\varphi < 0,$$

we see that if p is large enough, then $\Delta_p\varphi(x) < 0$ for all $x \in B(\hat{x}, r)$.

Set $\omega = \operatorname{int} B(\hat{x}, r)$. Choose an $\varepsilon > 0$ so that $(u - \varphi)|_{\partial\omega} < -3\varepsilon$. If we choose $j \in \mathbf{N}$ large enough, then we have $(u_{p_j} - \varphi)|_{\partial\omega} < -2\varepsilon$ and $(u_{p_j} - \varphi)(\hat{x}) > -\varepsilon$. Fix such a j and set $v = u_{p_j} + \varepsilon$ and $q = p_j$ for notational simplicity. We may assume as well that $\Delta_q\varphi(x) < 0$ for all $x \in \omega$.

Since $f \le 0$ in ω and $(v-\varphi)^+ \in W_0^{1,q}(\omega)$, we have

$$\int_\omega |Dv|^{q-2}Dv \cdot D(v-\varphi)^+ \,\mathrm{d}x = \int_\omega f(v-\varphi)^+ \,\mathrm{d}x \le 0,$$
$$\int_\omega |D\varphi|^{q-2}D\varphi \cdot D(v-\varphi)^+ \,\mathrm{d}x = -\int_\omega \Delta_q\varphi(v-\varphi)^+ \,\mathrm{d}x > 0,$$

and hence

$$\int_{\omega_+} (|Dv|^{q-2}Dv - |D\varphi|^{q-2}D\varphi) \cdot D(v-\varphi) \,\mathrm{d}x < 0,$$

where $\omega_+ = \{x \in \omega \mid v(x) > \varphi(x)\}$. On the other hand, because of the convexity of the function, $\xi \mapsto |\xi|^q$, we know that

$$\int_{\omega_+} (|Dv|^{q-2}Dv - |D\varphi|^{q-2}D\varphi) \cdot D(v-\varphi) \,\mathrm{d}x \ge 0,$$

which contradicts the above inequality. ◻

Remark. Let $u \in \mathcal{A}$. By an argument similar to the above proof, we can prove that $\min\{|Du(x)| - 1, -\Delta_\infty u(x)\} \le 0$ in Ω in the viscosity sense. However, we have a stronger conclusion that

$$|Du(x)| \le 1 \quad \text{in } \Omega \text{ in the viscosity sense.} \tag{5.3}$$

Indeed, if $u \in \mathcal{A}$, then $u \in X$, which implies that u satisfies (5.3) (see, for instance, Proposition 3.4 in [Ln]).

DEFINITION. *Let $Y \subset X$. We call Y essentially single if for any u, $v \in Y$, $u = v$ on* spt f.

PROPOSITION 5.5. *Let $Y \subset X$ be such that $\mathcal{A} \subset Y$. If Y is essentially single, then $\mathcal{A}$ is a singleton. In particular, the whole family $\{u_p\}_{p>1}$ converges in $C(\overline{\Omega})$.*

The following proof has already been explained in the introduction.

Proof. Let u, $v \in \mathcal{A}$. By assumption, we have $u = v$ on spt f. By Proposition 5.4, we see that u and v are both viscosity solutions of

$$-\Delta_\infty w(x) = 0 \quad \text{in } \Omega \setminus \operatorname{spt} f.$$

By the uniqueness result for this PDE due to [J], we conclude that $u = v$ in $\Omega \setminus \operatorname{spt} f$, which guarantees that $u = v$ in $\overline{\Omega}$. ◻

6. Properties of the set $\mathcal{M}$. In this section we collect some properties of the set $\mathcal{M}$ of the maximizers of the functional I_∞.

PROPOSITION 6.1. *Let $u \in \mathcal{M}$. Then*

$$u(x) = \inf\{u(y) + |x-y| \mid y \in \Omega_- \cup \partial\Omega\} \quad \text{for all } x \in \Omega_+ \tag{6.1}$$

and

$$u(x) = \sup\{u(y) - |x-y| \mid y \in \Omega_+ \cup \partial\Omega\} \quad \text{for all } x \in \Omega_-. \tag{6.2}$$

A proposition similar to this can be found in [EG] (Lemma 3.1 of [EG]), the proof of which can be easily adapted to our case, but we give a proof here for completeness.

Proof. We prove only (6.1), since the proof of (6.2) is similar. Let $u \in X$. Then

$$|u(x) - u(y)| \leq |x - y| \quad \text{for all } x, y \in \overline{\Omega},$$

from which we have

$$u(x) \leq \inf\{u(y) + |x - y| \mid y \in A\}$$

for all $x \in \overline{\Omega}$ and any $A \subset \overline{\Omega}$. In particular, we have

$$u(x) \leq \inf\{u(y) + |x - y| \mid y \in \Omega_- \cup \partial\Omega\} \quad \text{for all } x \in \overline{\Omega}. \tag{6.3}$$

Now, let $u \in \mathcal{M}$. Since $\mathcal{M} \subset X$, inequality (6.3) holds with this u. Setting

$$v(x) = \inf\{u(y) + |x - y| \mid y \in \Omega_- \cup \partial\Omega\} \quad \text{for } x \in \overline{\Omega},$$

we see immediately from the definition of v that

$$v(x) - v(y) \leq |x - y| \quad \text{for all } x, y \in \overline{\Omega},$$

which implies that $|Dv(x)| \leq 1$ almost everywhere in Ω. Also, we have

$$u(x) \leq v(x) \quad \text{for all } x \in \overline{\Omega}, \text{ by (6.3)},$$

and

$$v(x) \leq u(x) \quad \text{for all } x \in \Omega_- \cup \partial\Omega, \quad \text{by the definition of } v.$$

Combining these we find that $u(x) = v(x)$ for all $x \in \Omega_- \cup \partial\Omega$. In particular, $v(x) = 0$ for all $x \in \partial\Omega$. Thus we see that $v \in X$.

Next note that $I_\infty(u) = \max_{w \in X} I_\infty(w) \geq I_\infty(v)$. On the other hand, since $u = v$ on Ω_- and $v \geq u$ on Ω_+, we get $I_\infty(u) \leq I_\infty(v)$. Hence, we see that $I_\infty(v) = I_\infty(u)$. Now, since

$$\int_{\Omega_+} f(x)u(x)\,\mathrm{d}x = \int_{\Omega_+} f(x)v(x)\,\mathrm{d}x$$

and $v \geq u$ on Ω_+, we conclude that $u = v$ on Ω_+, which completes the proof. ∎

Remark. As one can see from the above proof, the set $\Omega_- \cup \partial\Omega$ in (6.1) can be replaced by any set $A \subset \overline{\Omega}$ satisfying $\Omega_- \cup \partial\Omega \subset A$. Similarly, the set $\Omega_+ \cup \partial\Omega$ in (6.2) can be replaced by any set $A \subset \overline{\Omega}$ satisfying $\Omega_+ \cup \partial\Omega \subset A$.

PROPOSITION 6.2. *Let $u, v \in \mathcal{M}$ and $k \geq 0$. Then $u \wedge (v + k)$, $(u - k) \vee v \in \mathcal{M}$.*

Proof. It is easy to see that $u \wedge (v + k)$, $(u - k) \vee v \in X$. In particular, we have

$$\max\{I_\infty(u \wedge (v + k)),\ I_\infty((u - k) \vee v)\} \leq I_\infty(u) = I_\infty(v).$$

Noting that $u \wedge (v + k) = u - (u - v - k)^+$ and $(u - k) \vee v = v + (u - v - k)^+$, we see that

$$I_\infty(u \wedge (v+k)) = I_\infty(u) - I_\infty((u-v-k)^+), \quad I_\infty((u-k) \vee v) = I_\infty(v) + I_\infty((u-v-k)^+),$$

and hence

$$\begin{aligned} 0 &\leq I_\infty(u) - I_\infty(u \wedge (v + k)) \\ &= I_\infty((u - v - k)^+) = I_\infty((u - k) \vee v) - I_\infty(v) \leq 0. \end{aligned}$$

Consequently, we have

$$I_\infty(u \wedge (v+k)) = I_\infty((u-k) \vee v) = I_\infty(u),$$

Thus we conclude that $u \wedge (v+k),\ (u-k) \vee v \in \mathcal{M}$. $\square$

The following proposition establishes the existence of the maximal and minimal elements of $\mathcal{M}$.

PROPOSITION 6.3. *Define* $V, W : \overline{\Omega} \to \mathbf{R}$ *by*

$$V(x) = \sup\{v(x) \mid v \in \mathcal{M}\} \quad \textit{and} \quad W(x) = \inf\{v(x) \mid v \in \mathcal{M}\}.$$

Then $V, W \in \mathcal{M}$.

Proof. We prove only the identity for V, since the proof of the other is similar.

First of all, note that $V \in X$. Choose a dense subset $\{y_k\}_{k\in\mathbf{N}}$ of Ω. For each $k \in \mathbf{N}$ we choose a sequence $\{v_{kj}\}_{j\in\mathbf{N}} \subset \mathcal{M}$ such that $\lim_{j\to\infty} v_{kj}(y_k) = V(y_k)$. By the definition of V, we have $v(x) \le V(x)$ for all $x \in \Omega$ and $v \in \mathcal{M}$. Therefore, we find that

$$V(y_l) = \sup\{v_{kj}(y_l) \mid k, j \in \mathbf{N}\} \quad \text{for } l \in \mathbf{N}.$$

Define $w \in X$ by setting

$$w(x) = \sup\{v_{kj}(x) \mid k, j \in \mathbf{N}\} \quad \text{for } x \in \overline{\Omega}.$$

It is immediate to see that $V = w$ on $\overline{\Omega}$.

We intend to show that $V \in \mathcal{M}$. Relabeling the countable set $\{v_{kj}\}_{k,j\in\mathbf{N}}$, we find a sequence $\{v_m\}_{m\in\mathbf{N}} \subset \mathcal{M}$ such that

$$V(x) = \sup\{v_m(x) \mid m \in \mathbf{N}\} \quad \text{for all } x \in \overline{\Omega}.$$

We define the nondecreasing sequence $\{w_j\}_{j\in\mathbf{N}}$ by induction as follows:

$$w_1 = v_1, \quad w_{j+1} = w_j \vee v_{j+1} \ \text{ for } j \in \mathbf{N}.$$

By Proposition 6.2, we see that $w_j \in \mathcal{M}$ for all $j \in \mathbf{N}$. It is clear that

$$\lim_{j\to\infty} w_j(x) = V(x) \quad \text{for all } x \in \overline{\Omega}.$$

Therefore we see by the monotone convergence theorem that

$$I_\infty(V) = \lim_{j\to\infty} I_\infty(w_j) = \max_{v\in X} I_\infty(v)$$

and conclude that $V \in \mathcal{M}$. $\square$

PROPOSITION 6.4. *For any* $u, v \in \mathcal{M}$*, we have*

$$\sup_{\Omega_+}(u-v)^+ = \sup_{\Omega_-}(u-v)^+.$$

Proof. Set $k = \sup_{\Omega_-}(u-v)^+$ and observe that $u(y) \le v(y) + k$ for $y \in \Omega_- \cup \partial\Omega$. Using Proposition 6.1, we see that for $x \in \Omega_+$,

$$u(x) = \inf\{u(y) + |x-y| \mid y \in \Omega_- \cup \partial\Omega\} \le \inf\{v(y) + |x-y| \mid y \in \Omega_- \cup \partial\Omega\} + k.$$

Hence, we have $u(x) \le v(x) + k$ for all $x \in \Omega_+$, and therefore

$$\sup_{\Omega_+}(u-v)^+ \le \sup_{\Omega_-}(u-v)^+.$$

Exchanging the role of Ω_+ and Ω_- in the above argument, we get

$$\sup_{\Omega_-}(u-v)^+ \le \sup_{\Omega_+}(u-v)^+$$

and finish the proof. ∎

PROPOSITION 6.5. *If $u \in \mathcal{M}$, then in the viscosity sense u satisfies*

$$|Du(x)| = 1 \quad \text{in } \Omega_+ \quad \text{and} \quad -|Du(x)| = -1 \quad \text{in } \Omega_-. \tag{6.4}$$

This proposition is an easy consequence of Proposition 6.1. For completeness we give a proof here.

Proof. Fix $u \in \mathcal{M}$. Let $\varphi \in C^1(\Omega)$ and $\hat{x} \in \Omega_+$. Assume that $u - \varphi$ attains a maximum at $\hat{x}$. Then, since $|u(x) - u(\hat{x})| \le |x - \hat{x}|$ for all $x \in \Omega$, we have as $x \to \hat{x}$

$$-|x-\hat{x}| \le u(x) - u(\hat{x}) \le \varphi(x) - \varphi(\hat{x}) = D\varphi(\hat{x}) \cdot (x - \hat{x}) + o(|x-\hat{x}|).$$

Substituting $\hat{x} - tD\varphi(\hat{x})$ for x and sending $t \searrow 0$, we see that $|D\varphi(\hat{x})| \le 1$.

Now, we assume that $u - \varphi$ attains a minimum at $\hat{x}$. In view of Proposition 6.1, we choose a point $y \in \overline{\Omega}_- \cup \partial\Omega$ so that $u(\hat{x}) = u(y) + |\hat{x} - y|$ holds. As before, we have as $x \to \hat{x}$

$$|x-y| - |\hat{x}-y| \ge u(x) - u(\hat{x}) \ge \varphi(x) - \varphi(\hat{x}) \ge -|D\varphi(\hat{x})||x-\hat{x}| + o(|x-\hat{x}|).$$

Substituting $\hat{x} + t(y - \hat{x})$ for x and sending $t \searrow 0$, we see that $|D\varphi(\hat{x})| \ge 1$.

Thus we see that u is a viscosity solution of $|Du(x)| = 1$ in Ω_+. A parallel argument shows that u is a viscosity solution of $-|Du(x)| = -1$ in Ω_-. ∎

PROPOSITION 6.6. *$\mathcal{M}$ is a convex set as a subset of $C(\overline{\Omega})$.*

Proof. Note that $X \subset C(\overline{\Omega})$ is a convex set. Since I_∞ is a linear functional on $C(\overline{\Omega})$, we conclude that $\mathcal{M}$ is convex. ∎

PROPOSITION 6.7. *Let $u, v \in \mathcal{M}$. Then*

$$Du(x) = Dv(x) \quad \textit{for a.e. } x \in \Omega_*.$$

Proof. Let u and $v \in \mathcal{M}$. Define $w \in C(\overline{\Omega})$ by

$$w = \frac{1}{2}(u + v).$$

According to Proposition 6.6, we have $w \in \mathcal{M}$. By Rademacher's theorem, we see that functions u, v, w are almost everywhere differentiable in Ω. Now, Proposition 6.5 yields

$$|Du(x)| = |Dv(x)| = |Dw(x)| = 1 \quad \text{for a.e. } x \in \Omega_*,$$

and therefore the strict convexity of the Euclidean norm in $\mathbf{R}^n$ implies that

$$Du(x) = Dv(x) = Dw(x) \quad \text{for a.e. } x \in \Omega_*. \qquad ∎$$

7. Case of no balanced family. In this section we first prove Theorem 2.3 and then examine a case where the hypothesis (2.4) is satisfied.

We begin with a lemma. Let $\{\mathcal{O}_\lambda\}_{\lambda\in\Lambda}$ be the classification of $\mathcal{O}_*$,

$$G_\lambda := \bigcup\{U \mid U \in \mathcal{O}_\lambda\} \quad \text{for } \lambda \in \Lambda,$$

as in section 2. Also, let $\Lambda_0 \subset \Lambda$ be as in section 2.

LEMMA 7.1. *If u,$v \in \mathcal{M}$, then $u - v$ is constant on any G_λ, with $\lambda \in \Lambda$.*

Proof. Let u, $v \in \mathcal{M}$. First of all, we observe that for any A, $B \subset \Omega$,

$$\begin{aligned}\inf_{(x,y)\in A\times B} |(u-v)(x)-(u-v)(y)| &\le \inf_{(x,y)\in A\times B}(|u(x)-u(y)|+|v(x)-v(y)|) \\ &\le 2d(A,B).\end{aligned} \tag{7.1}$$

Fix $\lambda \in \Lambda$ and $U, V \in \mathcal{O}_\lambda$. By Proposition 6.7, we have

$$(u-v)(x) = \begin{cases} k_U & \text{for } x \in U, \\ k_V & \text{for } x \in V \end{cases}$$

for some constants k_U, k_V. Fix any $\varepsilon > 0$. Since $\rho(U, V) = 0$, there is a finite family $W_1, \dots, W_m \in \mathcal{O}_*$ such that $d(U, W_1) + d(W_1, W_2) + \cdots + d(W_m, V) < \varepsilon$. By Proposition 6.7, for each $i \in \{1, \dots, m\}$ there is a constant k_i such that $(u-v)(x) = k_i$ for all $x \in U_i$.

Now, using (7.1), we get

$$\begin{aligned}|k_U - k_V| &\le |k_U - k_1| + |k_1 - k_2| + \cdots + |k_m - k_V| \\ &\le 2(d(U, W_1) + d(W_1, W_2) + \cdots + d(W_m, V)) < 2\varepsilon.\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we conclude that $k_U = k_V$. This shows that $u - v$ is constant on G_λ. □

LEMMA 7.2. *Let u, $v \in \mathcal{M}$ and $\lambda \in \Lambda_0$. Then $u = v$ on G_λ.*

Proof. In view of Lemma 7.1, let $k \in \mathbf{R}$ be a constant such that $u = v + k$ on G_λ. Fix any $U \in \mathcal{O}_\lambda$ and $\varepsilon > 0$. There is a finite sequence $U_1, \dots, U_m \in \Omega$ such that $d(U, U_1) + d(U_1, U_2) + \cdots + d(U_m, \partial\Omega) < \varepsilon$. As in the proof of Lemma 7.1, since $u = v = 0$ on $\partial\Omega$, we find that $|k| \le 2(d(U, U_1) + d(U_1, U_2) + \cdots + d(U_m, \partial\Omega)) < 2\varepsilon$. This is enough for us to conclude that $u = v$ on G_λ. □

Proof of Theorem 2.3. In view of Proposition 5.5, it is enough to show that $\mathcal{M}$ is essentially single.

For this we argue by contradiction. Thus we let $u, v \in \mathcal{M}$ and assume that $u \ne v$ on $\operatorname{spt} f$. We may assume that u and v are, respectively, the maximal and minimal elements of $\mathcal{M}$, i.e.,

$$u(x) \ge w(x) \ge v(x) \quad \text{ for all } x \in \Omega \text{ and } w \in \mathcal{M}.$$

Fix any $k > 0$ so that $k < \sup_{\Omega_*}(u - v)$. For $t \in (0, k]$ we set $w_t = u \wedge (v + t)$. Note that for $x \in \Omega$ and $0 \le t < s \le k$,

$$v(x) \le w_t(x) \le w_s(x) \le w_k(x).$$

Also, since $0 < k < \sup_{\Omega_*}(u - v)$, we see that $w_k - v$ attains the maximum value k at some point of Ω_*.

By Proposition 6.2, we have $w_t \in \mathcal{M}$ for all $t \in (0,\, k]$. Hence we have $I_\infty(w_k) = I_\infty(w_t)$ for all $t \in (0, k)$, which reads

$$0 = \int_\Omega \frac{(w_k - w_t)(x)}{k - t} f(x)\,\mathrm{d}x \quad \text{ for all } t \in (0, k).$$

For $0 \le t < k$, we set

$$A_t = \{x \in \Omega_* \mid w_t(x) < w_k(x)\} \quad \text{ and } \quad B = \bigcap_{0<t<k} A_t.$$

Note that $A_t \supset A_s$ for $0 < t < s < k$.

We claim here that

$$B = \{x \in \Omega_* \mid (w_k - v)(x) = k\}.$$

To see this, we write C for the right-hand side of the above identity. Let $x \in B$. By definition, we have $w_t(x) < w_k(x)$ for all $t \in (0, k)$. This implies that $w_t(x) = v(x) + t$ for all $t \in (0, k)$, and hence that $v(x) + t < u(x)$ for all $t \in (0, k)$. Therefore, we have $v(x) + k \le u(x)$ and, moreover, $w_k(x) = v(x) + k$. Thus, we see that $B \subset C$.

Next, let $x \in C$. We then have $w_k(x) = v(x) + k$, which yields that $u(x) \ge v(x) + k > v(x) + t$ for all $t \in [0, k)$. Hence we have $w_t(x) < w_k(x)$ for all $t \in (0, k)$. That is, we have $x \in B$, which concludes that $C \subset B$ and, moreover, $B = C$.

Since $w_k - v$ takes the value k at some point of Ω_*, we have $B \neq \emptyset$.

Now we go back to the equality

$$0 = \int_\Omega \frac{(w_k - w_t)(x)}{k - t} f(x)\, \mathrm{d}x = \int_{A_t} \frac{(w_k - w_t)(x)}{k - t} f(x)\, \mathrm{d}x \quad \text{for } t \in (0, k).$$

We are going to apply the Lebesgue convergence theorem. Since $A_t \supset A_s$ for $0 < t < s < k$ and $\bigcap_{0<t<k} A_t = B$, we see that as $t \nearrow k$,

$$\mathbf{1}_{A_t}(x) \to \mathbf{1}_B(x) \quad \text{for all } x \in \Omega.$$

Note that $|(w_k - w_t)(x)| \le |k - t|$ for all $t \in (0, k)$ and $x \in A_t$, and hence

$$\mathbf{1}_{A_t}(x) \left| \frac{(w_k - w_t)(x)}{k - t} f(x) \right| \le |f(x)| \quad \text{for all } t \in (0, k) \text{ and } x \in \Omega.$$

For $x \in B$, we have $w_k(x) = v(x) + k$ and $w_t(x) = v(x) + t$, and therefore

$$\frac{(w_k - w_t)(x)}{k - t} = 1.$$

Therefore, as $t \nearrow k$,

$$\mathbf{1}_{A_t}(x) \frac{(w_k - w_t)(x)}{k - t} f(x) \to \mathbf{1}_B(x) f(x).$$

We apply the Lebesgue convergence theorem along any sequence $t_k \nearrow k$, to conclude that $\int_B f(x)\, \mathrm{d}x = 0$.

Finally, noting by Lemma 7.1 that the function $w_k - v$ is constant on any G_λ, with $\lambda \in \Lambda$, and setting

$$\Gamma = \{\lambda \in \Lambda \mid (w_k - v)(x) = k \ \text{ on } G_\lambda\},$$

we have

$$B = \bigcup \{G_\lambda \mid \lambda \in \Gamma\}.$$

We see from Lemma 7.2 that $\Gamma \subset \Lambda \setminus \Lambda_0$. Recalling that $B \neq \emptyset$, by the assumption (2.4) we have $\int_B f(x)\, \mathrm{d}x \neq 0$. This is a contradiction. □

Let us examine the case where

$$\mu(\{x \in \Omega \mid f(x) = 0\}) = 0. \tag{7.2}$$

We have the following theorem as a corollary of Proposition 6.7.

THEOREM 7.3. *Under the assumption* (7.2), $\mathcal{M}$ *is a singleton. As a consequence, the whole family* $\{u_p\}_{p>1}$ *converges in* $C(\overline{\Omega})$.

Proof. Let $u, v \in \mathcal{M}$. By Proposition 6.7, we have $Du(x) = Dv(x)$ almost everywhere in Ω_*. By (7.2), we have $Du(x) = Dv(x)$ almost everywhere in Ω. Hence, $u = v$ in Ω. □

We wish to explain here that the convergence result in Theorem 7.3 can be shown as a consequence of Theorem 2.3.

PROPOSITION 7.4. *If* (7.2) *holds, then* $\Lambda = \Lambda_0$ *and hence* (2.4) *is satisfied.*

Proof. We need to show only that $\rho(U, \partial\Omega) = 0$ for all $U \in \mathcal{O}_*$.

To do this, we fix $U \in \mathcal{O}_*$ and $x \in U$. Choose a closest point y in $\partial\Omega$ to the point x. Set $R = |y - x|$. Choose a constant $r \in (0, R)$ so that $B(x, r) \subset U$. Let H be the hyperplane normal to the vector $y - x$ and passing through the point x, i.e., $H = \{\xi \in \mathbf{R}^n \mid (\xi - x) \cdot (y - x) = 0\}$. Let C be the truncated open cone generated by the point y and the $(n-1)$-dimensional sphere $H \cap B(x, r)$. That is, we write

$$C = \{ty + (1-t)\xi \mid (t, \xi) \in (0, 1) \times (H \cap B(x, r))\}.$$

Note that $C \subset \operatorname{int} B(x, R) \subset \Omega$.

By the assumption (7.2), we have

$$\mu(C) = \mu(C \cap \Omega_*).$$

Using the Fubini theorem, from this we deduce that for μ_{n-1}-almost all $\xi \in H \cap B(x, r)$, we have $\mu_1(\{t \in (0, 1) \mid ty + (1-t)\xi \in \Omega_*\}) = 1$.

Fix a point $\xi \in H \cap B(x, r)$ so that $\mu_1(\{t \in (0, 1) \mid ty + (1-t)\xi \in \Omega_*\}) = 1$. Define $I \subset (0, 1)$ by setting $I = \{t \in (0, 1) \mid ty + (1-t)\xi \in \Omega_*\}$. Since I is an open subset of $(0, 1)$, there is a sequence $\{I_j\}_{j \in J}$, with $J \subset \mathbf{N}$, of nonempty open intervals $I_j \subset (0, 1)$ such that $I = \bigcup\{I_j \mid j \in J\}$. We may assume as well that if $i, j \in J$ and $i \neq j$, then $I_i \cap I_j = \emptyset$. Since $\mu_1(I) = 1$, we have $\sum_{j \in J} \mu_1(I_j) = 1$.

Fix any $\varepsilon > 0$. There is a finite subset $J_\varepsilon \subset J$ such that $\sum_{j \in J_\varepsilon} \mu_1(I_j) > 1 - \varepsilon$. We may assume that $J_\varepsilon = \{1, \dots, m\}$, where m is a positive integer which depends on ε. For each $j \in J_\varepsilon$, we choose $a_j, b_j \in [0, 1]$ so that $I_j = (a_j, b_j)$. We may further assume that $a_1 < b_1 \leq a_2 < b_2 \leq \cdots \leq a_m < b_m$. Then we have $\sum_{j \in J_\varepsilon} \mu_1(I_j) = \sum_{j \in J_\varepsilon}(b_j - a_j) > 1 - \varepsilon$.

For each $j \in J_\varepsilon$, since $ty + (1-t)\xi \in \Omega_*$ for $t \in I_j$, and the set $\{ty + (1-t)\xi \mid t \in I_j\}$ is connected, we see that there is a $U_j \in \mathcal{O}_*$ such that

$$ty + (1-t)\xi \in U_j \text{ for } t \in I_j.$$

Observe that

$$\begin{aligned}
d(U, U_1) &\leq |\xi - (a_1 y + (1 - a_1)\xi)| \leq a_1 |y - \xi|, \\
d(U_{j-1}, U_j) &\leq |(b_{j-1} y + (1 - b_{j-1})\xi) - (a_j y + (1 - a_j)\xi)| \\
&\leq (a_j - b_{j-1})|y - \xi| \quad \text{ for all } j \in \{2, \dots, m\}, \\
d(U_m, \partial\Omega) &\leq |y - (b_m y + (1 - b_m)\xi)| \leq (1 - b_m)|y - \xi|.
\end{aligned}$$

Adding all of these, we get

$$\begin{aligned}
d(U, U_1) + d(U_1, U_2) + \cdots + d(U_m, \partial\Omega) &\leq (a_1 + (a_2 - b_1) + \cdots + (1 - b_m))|y - \xi| \\
&= \left(1 - \sum_{j \in J_\varepsilon} (b_j - a_j)\right) |y - \xi| < \varepsilon |y - \xi|.
\end{aligned}$$

Thus we see that $\rho(U, \partial\Omega) = 0$ and finish the proof. $\square$

8. Cases with vanishing integral. In this section we prove Theorem 2.4.

Proof of Theorem 2.4. Assume that (2.5) and (2.6) are satisfied.

First, we assume that (2.7) and (2.8) are satisfied and show in view of Proposition 5.5 that $\mathcal{A}$ is essentially single.

We observe that every $u \in \mathcal{A}$ is an odd function. This follows from the uniqueness of solutions of the Dirichlet problem

$$\begin{cases} -\Delta_p u(x) = f(x) & \text{in } \Omega, \\ \quad u(x) = 0 & \text{on } \partial\Omega. \end{cases} \tag{8.1}$$

Indeed, if u is a solution of (8.1), then the function $-u(-x)$ is a solution of (8.1) as well and, by the uniqueness, $u(x) = -u(-x)$ for all $x \in \overline{\Omega}$. This shows that every function $u \in \mathcal{A}$ is an odd function.

Now, let $u, v \in \mathcal{A}$. Since Ω_+ and Ω_- are connected and $Du(x) = Dv(x)$ for a.e. $x \in \Omega_*$, there is a constant $k \in \mathbf{R}$ such that $u(x) = v(x) + k$ for all $x \in \Omega_+$. By symmetry in u and v, we may assume that $k \geq 0$.

Since u and v are odd functions, we have

$$-u(-x) = -v(-x) + k \quad \text{ for all } x \in \Omega_+.$$

That is, $u(x) = v(x) - k$ for all $x \in \Omega_-$. Using Proposition 6.4, we thus get

$$k = \max_{\Omega_+}(u - v)^+ = \max_{\Omega_-}(u - v)^+ = 0,$$

which shows that $u(x) = v(x)$ for all $x \in \Omega_*$ and hence $\mathcal{A}$ is essentially single.

Next we turn to the case where (2.9) is satisfied. It is enough to show that $\mathcal{M}$ is essentially single. Let $u, v \in \mathcal{M}$; we will show that $u = v$ on Ω_*.

By Proposition 6.4, we have

$$\sup_{\Omega_+}(u - v)^+ = \sup_{\Omega_-}(u - v)^+. \tag{8.2}$$

We argue by contradiction, and hence suppose that $u(z) \neq v(z)$ for some $z \in \Omega_*$. We may assume that $u(z) > v(z)$. In view of (8.2), there is a constant $k > 0$ such that $u(x) - v(x) = k$ for all $x \in \Omega_*$.

From (2.9) we have either

$$\inf_{x\in\Omega_+} \sup_{y\in\Omega_-} [d(x) + d(y) - |x - y|] \leq 0 \tag{8.3}$$

or

$$\inf_{y\in\Omega_-} \sup_{x\in\Omega_+} [d(x) + d(y) - |x - y|] \leq 0. \tag{8.4}$$

We consider only the case where (8.3) holds, since the other case can be treated similarly.

For each $\varepsilon > 0$ there exists a point $x_\varepsilon \in \Omega_+$ such that $d(x_\varepsilon) + d(y) - |x_\varepsilon - y| \leq \varepsilon$ for all $y \in \Omega_-$. Since $u, v \in X$, we have $|u(x)| \vee |v(x)| \leq d(x)$ for all $x \in \Omega$. From these we get

$$d(x_\varepsilon) \leq \inf\{v(y) + |x_\varepsilon - y| \mid y \in \Omega_-\} + \varepsilon.$$

Since $v = 0$ on $\partial\Omega$, we have

$$d(x_\varepsilon) = \inf\{v(y) + |x_\varepsilon - y| \mid y \in \partial\Omega\}.$$

Therefore, using Proposition 6.1, we have

$$d(x_\varepsilon) \le \inf\{v(y) + |x_\varepsilon - y| \mid y \in \Omega_- \cup \partial\Omega\} + \varepsilon = v(x_\varepsilon) + \varepsilon.$$

Thus we obtain $u(x_\varepsilon) \le d(x_\varepsilon) \le v(x_\varepsilon) + \varepsilon$, which yields a contradiction by choosing $\varepsilon \in (0, k)$. This completes the proof. $\square$

REFERENCES

[ACBBV] L. Ambrosio, L. A. Caffarelli, Y. Brenier, G. Buttazzo, and C. Villani, *Optimal Transportation and Applications*, L. A. Caffarelli and S. Salsa, eds., Lecture Notes in Math. 1813, Springer-Verlag, Berlin, Centro Internazionale Matematico Estivo, Florence, 2003.

[A1] G. Aronsson, *Extension of functions satisfying Lipschitz conditions*, Ark. Mat., 6 (1967), pp. 551–561.

[A2] G. Aronsson, *On the partial differential equation* $u_x^2 u_{xx} + 2u_x u_y u_{xy} + u_y^2 u_{yy} = 0$, Ark. Mat., 7 (1968), pp. 395–425.

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

[BB] G. Barles and J. Busca, *Existence and comparison results for fully nonlinear degenerate elliptic equations without zeroth-order term*, Comm. Partial Differential Equations, 26 (2001), pp. 2323–2337.

[BBD] G. Bouchitté, G. Buttazzo, and L. De Pascale, *A p-Laplacian approximation of some mass optimization problems*, J. Optim. Theory Appl., 118 (2003), pp. 1–25.

[BDM] T. Bhattacharya, E. DiBenedetto, and J. Manfredi, *Limits as* $p \to \infty$ *of* $\Delta_p u_p = f$ *and related extremal problems*, in Some Topics in Nonlinear PDEs (Turin, 1989), Rend. Sem. Mat. Univ. Politec. Torino, 1991, pp. 15–68.

[BK] M. Belloni and B. Kawohl, *The pseudo-p-Laplace eigenvalue problem and viscosity solutions as* $p \to \infty$, ESAIM Control Optim. Calc. Var., 10 (2004), pp. 28–52.

[CEG] M. G. Crandall, L. C. Evans, and R. F. Gariepy, *Optimal Lipschitz extensions and the infinity Laplacian*, Calc. Var. Partial Differential Equations, 13 (2001), pp. 123–139.

[CIL] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[D] E. DiBenedetto, $C^{1+\alpha}$ *local regularity of weak solutions of degenerate elliptic equations*, Nonlinear Anal., 7 (1983), pp. 827–850.

[EG] L. C. Evans and W. Gangbo, *Differential equations methods for the Monge-Kantorovich mass transfer problem*, Mem. Amer. Math. Soc., 137 (653) (1999).

[FIN] N. Fukagai, M. Ito, and K. Narukawa, *Limit as* $p \to \infty$ *of p-Laplace eigenvalue problems and* L^∞*-inequality of the Poincaré type*, Differential Integral Equations, 12 (1999), pp. 183–206.

[GT] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Grundlehren Math. Wiss. 224, Springer-Verlag, Berlin, 1983.

[IK] T. Ishibashi and S. Koike, *On fully nonlinear PDEs derived from variational problems of* L^p *norms*, SIAM J. Math. Anal., 33 (2001), pp. 545–569.

[IL1] H. Ishii and P. Loreti, *On relaxation in an* L^∞ *optimization problem*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 599–615.

[IL2] H. Ishii and P. Loreti, *Relaxation of Hamilton-Jacobi equations*, Arch. Rational Mech. Anal., 169 (2003), pp. 265–304.

[IL3] H. Ishii and P. Loreti, *Asymptotic Behavior of the Solution of the p-Laplace Equation on a Finite Interval*, unpublished.

[IL4] H. Ishii and P. Loreti, *Asymptotic behavior of solutions of p-Laplace type equations as p goes to infinity*, in Proceedings of the 12th Tokyo Conference on Nonlinear PDE 2003, Tokyo, Japan, 2004, pp. 32–50.

[J] R. Jensen, *Uniqueness of Lipschitz extensions: Minimizing the sup norm of the gradient*, Arch. Ration. Mech. Anal., 123 (1993), pp. 51–74.

[JLM] P. Juutinen, P. Lindqvist, and J. J. Manfredi, *The* ∞*-eigenvalue problem*, Arch. Ration. Mech. Anal., 148 (1999), pp. 89–105.

[K] B. Kawohl, *On a family of torsional creep problems*, J. Reine Angew. Math., 410 (1990), pp. 1–22.

[L] J.-L. Lions, *Quelques méhodes de réolution des problèmes aux limites non linéires*, Dunod, Gauthier-Villars, Paris, 1969.

[Lb] G. M. Lieberman, *Boundary regularity for solutions of degenerate elliptic equations*, Nonlinear Anal., 12 (1988), pp. 1203–1219.

[Ln] P.-L. Lions, *Generalized solutions of Hamilton-Jacobi Equations*, Res. Notes in Math. 69, Pitman, Boston, London, 1982.

[PP] L. E. Payne and G. A. Philippin, *Some applications of the maximum principle in the problem of torsional creep*, SIAM J. Math. Anal., 33 (1977), pp. 446–455.

[T] P. Tolksdorf, *Regularity for a more general class of quasilinear elliptic equations*, J. Differential Equations, 51 (1984), pp. 126–150.

[U] K. Uhlenbeck, *Regularity for a class of non-linear elliptic systems*, Acta Math., 138 (1977), pp. 219–240.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 438–454

A GENERALIZED ENERGY FUNCTIONAL FOR PLANE COUETTE FLOW*

R. KAISER†, A. TILGNER‡, AND W. VON WAHL†

Abstract. We present a generalized energy functional $\mathcal{E}$ for plane Couette flow providing conditional nonlinear stability for Reynolds numbers Re below $\mathrm{Re}_{\mathcal{E}} := 177.2$, which is larger than the ordinary energy stability limit. The method allows the explicit calculation of so-called stability balls in the $\mathcal{E}^{1/2}$-norm; i.e., the system is stable with respect to any perturbation with $\mathcal{E}^{1/2}$-norm in this ball.

Key words. hydrodynamical stability, plane Couette flow

AMS subject classifications. 35Q30, 76E05, 76E30

DOI. 10.1137/S0036141004442604

1. Introduction. Plane Couette flow is a paradigm with a long history of scientific investigation for a whole class of hydrodynamic stability problems, viz. plane parallel shear flows (cf. [DR, SH]). For these flows there is no obvious physical mechanism which triggers instability as, e.g., in circular Couette flow or in convection problems. Instead, viscous stresses seem to play the dominant role at the onset of instability. So, despite the striking simplicity of the set-up of these systems the onset of instability is up to the present insufficiently understood and its nature is the subject of ongoing research [TTRD, Gr].

The classical methods which yield rigorous stability results are the method of linearized stability and the energy method. The former method provides the critical value Re_c of the Reynolds number Re, below which the system is conditionally stable and above which it is unstable. In the case of Couette flow[1] it turns out that $\mathrm{Re}_c = \infty$, i.e., the system is linearly completely stable [Ro]. The second method provides global asymptotic stability below some value Re_E. For Couette flow $\mathrm{Re}_E = 82.6$ if Re is defined with the separation of the walls and their velocity difference [Jo]. This has to be compared with the experimentally observed onset of instability, which occurs at $\mathrm{Re} \approx 1300$ [Gr, DD]. Thus, none of the classical methods describes the instability behavior of Couette flow satisfactorily.

A more recent method which has successfully been applied to a couple of hydrodynamic stability problems uses generalized energy functionals which are better adjusted to the specific problems under consideration [J1, J2, GP, St]. A generalized energy functional $\mathcal{E}$ is a bilinear form of the dynamic variables of the problem. In comparison with the ordinary energy these variables are, however, differently weighted by additional coupling parameters and appear possibly in the form of higher derivatives. A first part $\mathcal{E}_1$ of the functional determines (analogously to the energy method) via a variational problem the stability boundary $\mathrm{Re}_{\mathcal{E}}$. The coupling parameters are chosen such that $\mathrm{Re}_{\mathcal{E}}$ becomes as large as possible—in particular, larger than Re_E.

*Received by the editors March 30, 2004; accepted for publication (in revised form) January 4, 2005; published electronically October 14, 2005.
http://www.siam.org/journals/sima/37-2/44260.html

†Fakultät für Mathematik und Physik, Universität Bayreuth, D-95440 Bayreuth, Germany (ralf.kaiser@uni-bayreuth.de, wolf.vonwahl@uni-bayreuth.de).
‡Institut für Geophysik, Universität Göttingen, D-37075 Göttingen, Germany (andreas.tilgner@geo.physik.uni-goettingen.de).

[1]Henceforth Couette flow always means *plane* Couette flow.

Contrary to the energy balance the nonlinear terms in general do not drop from the generalized energy balance. Therefore, a second part $\mathcal{E}_2$ involving higher derivatives of the dynamic variables is needed in $\mathcal{E}$ in order to dominate these terms. If the method works one obtains conditional stability for all Reynolds numbers below $\mathrm{Re}_{\mathcal{E}}$ together with explicit stability balls in the $\mathcal{E}^{1/2}$-norm. This is different from the method of linearized stability which gives no estimates for stability balls.

Generalized energy functionals have already been applied to plane parallel shear flows, however, under the assumption of stress-free boundary conditions for the perturbations [RM]. This assumption clearly overestimates the stability of these flows since wall induced stresses are neglected. In fact, the authors find conditional stability for all Reynolds numbers not only for Couette flow but also for Poiseuille flow, a system with finite critical Reynolds number if rigid boundary conditions are used.

If rigid boundary conditions are used no generalized energy functionals have been found so far, neither in plane parallel shear flows nor in any other hydrodynamical system with nontrivial basic flow and unrestricted (three-dimensional) perturbations. Moreover, in the case of Couette flow it has been argued that the generalized energy method as applied to systems with stress-free boundary conditions is incompatible with rigid boundary conditions [KT].

We present in this paper a generalized energy functional $\mathcal{E}$ for Couette flow (with correct rigid boundary conditions), which provides conditional stability for all Reynolds numbers below $\mathrm{Re}_{\mathcal{E}} = 177.2$. This number is still far below what is desired. However, there is now hope that still more appropriate functionals can be found which cover a larger stability region. The crucial point which allows the treatment of rigid boundary conditions is a more refined calculus inequality that takes advantage of the special geometry of the system.

The following point is of some historical interest: For Couette flow $\mathrm{Re}_{\mathcal{E}} = 177.2$ is just the two-dimensional energy stability limit, where perturbations are not allowed to vary in the spanwise direction. Following Orr [Or], early researchers in the field took this number for the correct energy stability limit. It came as a surprise when Joseph [Jo] showed that the complementary two-dimensional problem provided a considerably lower limit. Busse proved subsequently that the latter limit is in fact the correct energy stability limit [Bu]. Thus, our result may be viewed as a late justification of (a weakened version of) Orr's original claim.

The paper is organized as follows: Section 2 sets the mathematical framework for the subsequent analysis. In particular, we introduce the so-called poloidal-toroidal decomposition of divergence-free vector fields. This decomposition eliminates the divergence constraint and provides appropriate building blocks for generalized energy functionals. In section 3 a linear auxiliary problem is solved, viz. the variational problem associated with $\mathcal{E}_1$ which determines the stability limit $\mathrm{Re}_{\mathcal{E}}$. Section 4 provides estimates of the remaining terms in the energy balance of $\mathcal{E}_1 + \mathcal{E}_2$, the nonlinear terms in particular, and it formulates the basic stability result. Some well-known inequalities as well as the refined calculus inequality are collected in Appendix A, and the results of a numerical computation related to the variational problem are contained in Appendix B.

2. Mathematical setting. The Couette system is appropriately modeled by an infinite layer $\mathbb{R} \times (-\frac{1}{2}, \frac{1}{2})$ of thickness 1 with horizontal coordinates x, y and vertical coordinate z. The basic flow in this system takes the dimensionless form

$$\mathbf{U}_0 = \mathbf{U}_0(z) = \mathrm{Re} \begin{pmatrix} -z \\ 0 \\ 0 \end{pmatrix} \tag{2.1}$$

with Re being the Reynolds number based on the distance between bottom and top boundaries of the layer and their velocity difference. In order to investigate the stability of $\mathbf{U}_0$ we impose perturbations $\mathbf{u} = (u_x, u_y, u_z)$. These are governed by the system

$$\begin{aligned} \partial_t \mathbf{u} - \Delta \mathbf{u} - \mathrm{Re}(z\,\partial_x \mathbf{u} + u_z \mathbf{e}_x) + \mathbf{u} \cdot \nabla \mathbf{u} + \nabla p &= 0, \\ \nabla \cdot \mathbf{u} &= 0 \end{aligned} \tag{2.2}$$

in $\mathbb{R}^2 \times (-\frac{1}{2}, \frac{1}{2}) \times (0, T)$, $T > 0$, and satisfy the boundary conditions

$$\mathbf{u}(x, y, z, t) = 0 \qquad \text{for} \quad (x, y, z) \in \mathbb{R}^2 \times \left\{-\frac{1}{2}, \frac{1}{2}\right\},\ t > 0. \tag{2.3}$$

Here $\mathbf{e}_x = (1, 0, 0)^{\mathrm{T}}$. The initial value $\mathbf{u}(\cdot, \cdot, \cdot, 0) = \mathbf{u}_0$ is assumed to be given (and of course solenoidal). $\mathbf{u}$ corresponds to the velocity field of the perturbation and p denotes the pressure perturbation. Both $\mathbf{u}$ and ∇p are assumed to be x, y-periodic with respect to a rectangle $\mathcal{P} = (-\frac{\pi}{\alpha}, \frac{\pi}{\alpha}) \times (-\frac{\pi}{\beta}, \frac{\pi}{\beta})$ with wave numbers $(\alpha, \beta) \in \mathbb{R}^2_+$. In the following it suffices, therefore, to consider functions over the box

$$\Omega = \mathcal{P} \times \left(-\frac{1}{2}, \frac{1}{2}\right) = \left(-\frac{\pi}{\alpha}, \frac{\pi}{\alpha}\right) \times \left(-\frac{\pi}{\beta}, \frac{\pi}{\beta}\right) \times \left(-\frac{1}{2}, \frac{1}{2}\right).$$

As basic function space we take $L^2(\Omega)$. In what follows, $\|\cdot\|$ is always the norm in $L^2(\Omega)$ except in the case when applied to a function defined on $(-\frac{1}{2}, \frac{1}{2})$. Then, $\|\cdot\|$ means the norm in $L^2(-\frac{1}{2}, \frac{1}{2})$; the correct notion should be clear from the context. $(\cdot, \cdot)$ always denotes the scalar product associated with $\|\cdot\|$.

In order to cope with the divergence constraint $(2.2)_2$ we make use of the poloidal-toroidal decomposition [SW]:

$$\begin{aligned} \mathbf{u} &= \nabla \times (\nabla \times (\varphi\, \mathbf{e}_z)) + \nabla \times (\psi\, \mathbf{e}_z) + \mathbf{F} \\ &=: \ \ \varphi + \ \ \psi + \mathbf{F}. \end{aligned} \tag{2.4}$$

Here $\mathbf{e}_z = (0, 0, 1)^{\mathrm{T}}$. The functions φ and ψ are determined uniquely if one requires them to be periodic with respect to $\mathcal{P}$ and to fulfill $\int_{\mathcal{P}} \varphi(x, y, z)\, dx\, dy = \int_{\mathcal{P}} \psi(x, y, z)\, dx\, dy = 0$ for every $z \in (-\frac{1}{2}, \frac{1}{2})$. The first part in (2.4) is called the poloidal part of $\mathbf{u}$ and the second part the toroidal one. The third part, the mean flow, depends only on z and has a constant third component. These three parts are mutually orthogonal in $L^2(\Omega)^3$. The vector operators and have the form

$$\varphi = \begin{pmatrix} \partial_x \partial_z \varphi \\ \partial_y \partial_z \varphi \\ (-\Delta_2)\varphi \end{pmatrix}, \qquad \psi = \begin{pmatrix} \partial_y \psi \\ -\partial_x \psi \\ 0 \end{pmatrix},$$

where $\Delta_2 = \partial_x^2 + \partial_y^2$ is the horizontal Laplacian. The boundary conditions (2.3) for $\mathbf{u}$ transform into

$$\varphi = \partial_z \varphi = 0, \quad \psi = 0, \quad F_x = F_y = 0 \qquad \text{for} \quad z = \pm\frac{1}{2}, \tag{2.5}$$

and $F_z(z) \equiv 0$. Applying the operators and to $(2.2)_1$ as well as taking the mean with respect to $\mathcal{P}$, the system (2.2) can equivalently be formulated in terms of the new variables $(\varphi, \psi, F_x, F_y)$:

(2.6)
$$
\begin{aligned}
(-\Delta)(-\Delta_2)\partial_t\varphi + \Delta^2(-\Delta_2)\varphi - \operatorname{Re} z\,(-\Delta)(-\Delta_2)\partial_x\varphi + \quad \cdot (\mathbf{u} \cdot \nabla\mathbf{u}) &= 0,\\
(-\Delta_2)\partial_t\psi + (-\Delta)(-\Delta_2)\psi - \operatorname{Re} z\,(-\Delta_2)\partial_x\psi + \operatorname{Re}(-\Delta_2)\partial_y\varphi - \quad \cdot (\mathbf{u} \cdot \nabla\mathbf{u}) &= 0,\\
\partial_t F_x + (-\partial_z^2)F_x + \tfrac{1}{|\mathcal{P}|}\textstyle\int_{\mathcal{P}} \tilde{\mathbf{u}} \cdot \nabla\tilde{u}_x \, dxdy &= 0,\\
\partial_t F_y + (-\partial_z^2)F_y + \tfrac{1}{|\mathcal{P}|}\textstyle\int_{\mathcal{P}} \tilde{\mathbf{u}} \cdot \nabla\tilde{u}_y \, dxdy &= 0.
\end{aligned}
$$

$\tilde{\mathbf{u}} := \quad \varphi + \quad \psi$ is that part of $\mathbf{u}$ which has vanishing mean value over $\mathcal{P}$, and $|\mathcal{P}| := \frac{4\pi^2}{\alpha\beta}$ denotes the volume of $\mathcal{P}$.

With $\Phi := (\varphi, \psi, F_x, F_y)^{\mathrm{T}}$ a neat matrix notation can be used for system (2.6):

$$
\mathcal{B}\,\partial_t\Phi + \mathcal{A}\Phi - \operatorname{Re}\mathcal{C}\Phi + \mathcal{M}(\Phi, \Phi) = 0. \tag{2.7}
$$

Here, $\mathcal{B}$ and $\mathcal{A}$ are diagonal matrix operators, $\mathcal{C}$ is a nonnormal interaction matrix, and $\mathcal{M}$ is a bilinear form. The operator $\mathcal{A}$, for example, has the form

$$
\mathcal{A} = \operatorname{diag}\big(\Delta^2(-\Delta_2), (-\Delta)(-\Delta_2), (-\partial_z^2), (-\partial_z^2)\big)
$$

acting in the Hilbert space

$$
\mathcal{H} := L^2_M(\Omega) \times L^2_M(\Omega) \times L^2\left(\left(-\frac{1}{2}, \frac{1}{2}\right)\right) \times L^2\left(\left(-\frac{1}{2}, \frac{1}{2}\right)\right),
$$

where $L^2_M(\Omega)$ denotes the space $\{f \in L^2(\Omega) \mid \int_{\mathcal{P}} f(x,y,z)\,dxdy = 0$ for a.e. $z \in (-\frac{1}{2}, \frac{1}{2})\}$. The domain $D(\mathcal{A})$ is most easily described in terms of a Fourier mode expansion for φ and ψ with respect to the horizontal variables x and y:

$$
\varphi(x,y,z) = \frac{1}{\sqrt{|\mathcal{P}|}} \sum_{\in\mathbb{Z}^2\setminus\{0\}} a \;\; (z) e^{i(\alpha\kappa_1 x + \beta\kappa_2 y)}, \tag{2.8}
$$

$$
\psi(x,y,z) = \frac{1}{\sqrt{|\mathcal{P}|}} \sum_{\in\mathbb{Z}^2\setminus\{0\}} b \;\; (z) e^{i(\alpha\kappa_1 x + \beta\kappa_2 y)}. \tag{2.9}
$$

We then define (cf. [KS, Wa])

$$
D(\mathcal{A}) = D(\Delta^2(-\Delta_2)) \times D((-\Delta)(-\Delta_2)) \times D(-\partial_z^2) \times D(-\partial_z^2),
$$

where

$$
\begin{aligned}
D(\Delta^2(-\Delta_2)) = \Bigg\{\varphi \;\Bigg|\; &\varphi \text{ expanded as in (2.8)},\\
&a \;\; \in H^4\left(\left(-\frac{1}{2}, \frac{1}{2}\right)\right), \quad a \;\; = \partial_z a \;\; = 0 \quad \text{at} \quad z = \pm\frac{1}{2},\\
&\sum_{\in\mathbb{Z}^2\setminus\{0\}} (\alpha^2\kappa_1^2 + \beta^2\kappa_2^2)^2 \int_{-1/2}^{1/2} |(-\partial_z^2 + \alpha^2\kappa_1^2 + \beta^2\kappa_2^2)^2 a \;\; (z)|^2 \, dz < \infty \Bigg\},
\end{aligned}
$$

$$D((-\Delta)(-\Delta_2)) = \Bigg\{ \psi \;\Bigg|\; \psi \text{ expanded as in (2.9)},$$

$$b \;\in H^2\left(\left(-\frac{1}{2}, \frac{1}{2}\right)\right), \quad b \;= 0 \quad \text{at} \quad z = \pm\frac{1}{2},$$

$$\sum_{\in \mathbb{Z}^2\setminus\{0\}} (\alpha^2\kappa_1^2 + \beta^2\kappa_2^2)^2 \int_{-1/2}^{1/2} |(-\partial_z^2 + \alpha^2\kappa_1^2 + \beta^2\kappa_2^2) b \;(z)|^2\, dz < \infty \Bigg\},$$

and

$$D(-\partial_z^2) = H^2\left(\left(-\frac{1}{2}, \frac{1}{2}\right)\right) \cap \mathring{H}^1\left(\left(-\frac{1}{2}, \frac{1}{2}\right)\right).$$

With these definitions $\mathcal{A}$ is a self-adjoint and strictly positive operator. Thus, fractional powers of $\mathcal{A}$ make sense and can analogously be explained in terms of the expansions (2.8) and (2.9). Similar definitions apply to the operators $\mathcal{B}$ and $\mathcal{C}$.

A natural class of vector fields within which (2.7) can locally be uniquely solved is given by (cf. [Wa])

$$\Phi \in L^2((0,T), \mathcal{D}(\mathcal{A})), \qquad \partial_t \Phi \in L^2((0,T), \mathcal{D}(\mathcal{B})), \tag{2.10}$$

and, as a consequence,

$$\Phi \in C^0([0,T], I)$$

with I being an appropriate interpolation space between $D(\mathcal{A})$ and $D(\mathcal{B})$. Going back to (2.2) we obtain from (2.10) at least a solution $(\mathbf{u}, p)$ with

(2.11)

$$\mathbf{u} \in L^2((0,T), D(-\Delta)) \cap C^0([0,T], D((-\Delta)^{1/2})), \qquad \partial_t \mathbf{u} \in L^2((0,T), L^2(\Omega)),$$

$$\nabla p \in L^2((0,T), L^2(\Omega)),$$

where $D(-\Delta) = \{\mathbf{u} \in (H^2(\Omega))^3 \mid \mathbf{u} \text{ periodic in } x \text{ and } y,\ \mathbf{u} = 0 \text{ at } z = \pm\frac{1}{2}\}$. This is the usual notion of a strong solution. On the other hand, strong solutions have further regularity properties. In particular, decomposing $\mathbf{u}$ from the class (2.11) in its poloidal and toroidal part and the mean flow, Φ can be shown to lie in the class (2.10). In the following we work with solutions within this class. All manipulations with $\mathbf{u}$ (or Φ) and its horizontal derivatives $\partial_x \mathbf{u}, \partial_y \mathbf{u}$ in the subsequent sections are then justified.

The energy of the system (in the volume Ω) becomes in the new variables[2]

$$E = \frac{1}{2}\,\|\mathbf{u}\|^2 = \frac{1}{2}\Big\{\| \;\varphi\|^2 + \| \;\psi\|^2 + |\mathcal{P}|\|\mathbf{F}\|^2\Big\}, \tag{2.12}$$

and the variational expression determining Re_E takes the form (cf. [KS])

$$\frac{|\Re(u_x, u_z)|}{\|\nabla \mathbf{u}\|^2} = \frac{|\Re((-\Delta_2)\varphi, \partial_x\partial_z\varphi + \partial_y\psi + F_x)|}{\|(-\Delta) \;\varphi\|^2 + \| \;\psi\|^2 + |\mathcal{P}|\|\partial_z \mathbf{F}\|^2}. \tag{2.13}$$

[2] We use the usual notation for L^2-scalar products of vector- or tensor-type quantities. Thus, there is, e.g., $\|\mathbf{u}\|^2 = (\mathbf{u}, \mathbf{u}) = \sum_{i=1}^3 (u_i, u_i)$ or $\|\nabla \mathbf{u}\|^2 = (\nabla \mathbf{u}, \nabla \mathbf{u}) = \sum_{i,j=1}^3 (\partial_i u_j, \partial_i u_j)$. Note that $\nabla \mathbf{u}$ is understood in the sense of a tensor product, whereas $\mathbf{u} \cdot \nabla = \sum_{i=1}^3 u_i \partial_i$ means the scalar product in $\mathbb{R}^3$.

For later convenience we admit here complex valued velocity fields. Thus, the real part (denoted by $\Re$) of the interaction term appears in the numerator of (2.13). Re_E is then given by

$$\mathrm{Re}_E^{-1} = \sup_{(\alpha,\beta)\in\mathbb{R}_+^2} \sup_{(\varphi,\psi)\in\mathcal{V}_{\alpha\beta}} \frac{|\Re((-\Delta_2)\varphi,(\partial_x\partial_z\varphi+\partial_y\psi))|}{\|(-\Delta)\ \ \varphi\|^2+\|\ \ \psi\|^2}. \tag{2.14}$$

Note that $\mathbf{F}$ does not depend on x or y and, therefore, drops from the numerator of (2.13). Thus, $\mathbf{F}$ does not contribute to the supremum of (2.13) and can be omitted altogether.

The variational class $\mathcal{V}_{\alpha\beta}$ should reflect the mean value condition, the boundary conditions, and the periodicity of the functions φ and ψ. Moreover, it should ensure that the supremum is in fact attained. A suitable choice is $\mathcal{V}_{\alpha\beta} = D(\tilde{\mathcal{A}}^{1/2}) \setminus \{(0,0)\}$, where $\tilde{\mathcal{A}}$ is that part of $\mathcal{A}$ that is operating on (φ,ψ) in the Hilbert space $\tilde{\mathcal{H}} := L_M^2(\Omega)\times L_M^2(\Omega)$.

If the class $\mathcal{V}_{\alpha\beta}$ of admissible functions is restricted to the class $\mathcal{V}_\alpha$ of functions depending only on x and z, or to the class $\mathcal{V}_\beta$ of functions depending only on y and z, the corresponding two-dimensional limits Re_E^x and Re_E^y are determined by the following simplified variational expressions:

$$\frac{1}{\mathrm{Re}_E^y} = \sup_{\beta\in\mathbb{R}_+} \sup_{(\varphi,\psi)\in\mathcal{V}_\beta} \frac{|\Re((-\Delta_2)\varphi,\partial_y\psi)|}{\|(-\Delta)\partial_y\varphi\|^2+\|\ \ \psi\|^2}, \tag{2.15}$$

$$\frac{1}{\mathrm{Re}_E^x} = \sup_{\alpha\in\mathbb{R}_+} \sup_{(\varphi,0)\in\mathcal{V}_\alpha} \frac{|\Re((-\Delta_2)\varphi,\partial_x\partial_z\varphi)|}{\|(-\Delta)\partial_x\varphi\|^2}. \tag{2.16}$$

It is well known that $\mathrm{Re}_E = \mathrm{Re}_E^y = 82.6\ldots$ and $\mathrm{Re}_E^x = 177.2\ldots$ (cf. [Or, Jo, Bu]).

Applying the matrix notation the variational expression (2.13) takes the form

$$\frac{|(\Phi,\hat{\mathcal{C}}\Phi)|}{\|\mathcal{A}^{1/2}\Phi\|^2}$$

with the symmetric lower order operator

$$\hat{\mathcal{C}} = \frac{1}{2}\begin{pmatrix} 2(-\Delta_2)\partial_x\partial_z & (-\Delta_2)\partial_y & (-\Delta_2) & 0 \\ -(-\Delta_2)\partial_y & 0 & 0 & 0 \\ (-\Delta_2) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Thus, $\mathcal{A}^{-1/2}\hat{\mathcal{C}}\mathcal{A}^{-1/2}$ is a self-adjoint and compact operator in $\mathcal{H}$ and the supremum (with respect to $(\varphi,\psi)\in D(\tilde{\mathcal{A}}^{1/2})$) in (2.14) is actually a maximum. This argument applies, of course, also to the suprema in (2.15) and (2.16).

3. Generalized functional and variational problem. The usual method to proceed from the energy functional to a generalized one is to introduce additional coupling parameters and possibly additional derivatives in order to weigh the dynamic variables in an optimal way. For this purpose the generalized energy balance is considered and (analogously to the energy method) the ratio of the interaction term over the dissipative term is maximized with respect to the admissible functions. This maximum still depends on the coupling parameters and possibly discrete parameters counting the additional derivatives. Minimizing with respect to these parameters furnishes optimal (generalized) energy limits. Therefore, the first problem is to find functionals which furnish larger stability limits than those provided by the energy functional. Considering the functional (2.12) with $\mathbf{F} \equiv 0$ (as already noted, the mean flow does not contribute to the maximum in the variational problem) there is, however, not much freedom to introduce additional parameters. An obvious choice is the functional

$$\mathcal{E}_1[\varphi, \psi] := \frac{1}{2}\Big\{\| \ \ \varphi\|^2 + \lambda \, \| \ \ \psi\|^2\Big\} \tag{3.1}$$

with $0 < \lambda < \infty$. Taking the scalar product of $(2.6)_{1,2}$ with (φ, ψ) in $\tilde{\mathcal{H}}$ and using the boundary conditions (2.5) one obtains the generalized energy balance

$$\partial_t \mathcal{E}_1 = -\mathcal{D}_1 + \mathrm{Re}\, \mathcal{I}_1 + \mathcal{N}_1 \tag{3.2}$$

with

$$\begin{aligned} \mathcal{D}_1[\varphi, \psi] &:= \|(-\Delta) \ \ \varphi\|^2 + \lambda \| \ \ \psi\|^2, \\ \mathcal{I}_1[\varphi, \psi] &:= \Re\big((-\Delta_2)\varphi, \partial_x \partial_z \varphi\big) + \lambda \, \Re\big((-\Delta_2)\varphi, \partial_y \psi\big), \\ \mathcal{N}_1[\varphi, \psi, \mathbf{F}] &:= -\Re\big((\mathbf{u} \cdot \nabla \mathbf{u}), \ \ \varphi + \lambda \ \ \psi\big). \end{aligned} \tag{3.3}$$

The generalized energy limit $\mathrm{Re}_{\mathcal{E}}$ is then determined by

$$\begin{aligned} {\mathrm{Re}_{\mathcal{E}}}^{-1} &= \sup_{(\alpha,\beta) \in \mathbb{R}^2_+} \ \sup_{(\varphi,\psi) \in \mathcal{V}_{\alpha\beta}} \frac{\mathcal{I}_1}{\mathcal{D}_1}[\varphi, \psi] \\ &= \sup_{(\alpha,\beta) \in \mathbb{R}^2_+} \ \sup_{(\varphi,\psi) \in \mathcal{V}_{\alpha\beta}} \frac{|\Re((-\Delta_2)\varphi, \partial_x \partial_z \varphi) + \lambda \, \Re((-\Delta_2)\varphi, \partial_y \psi)|}{\|(-\Delta) \ \ \varphi\|^2 + \lambda \, \| \ \ \psi\|^2}. \end{aligned} \tag{3.4}$$

Note that in (3.4) $\mathcal{I}_1$ can always be replaced by $|\mathcal{I}_1|$; as with $(\varphi(x, y, z), \psi(x, y, z)) \in \mathcal{V}_{\alpha\beta}$, $(\varphi(-x, -y, z), \psi(-x, -y, z))$ is also admissible. Thus, $\mathcal{I}_1$ can always be chosen positive without affecting $\mathcal{D}_1$.

A comparison of (3.4) with the two-dimensional variational expressions (2.15) and (2.16) already furnishes some bounds on $\mathrm{Re}_{\mathcal{E}}$: Setting $\varphi = \varphi(x, z)$, $\psi = 0$ in (3.4) reduces the variational expression to that in (2.16), which implies the bound $\mathrm{Re}_{\mathcal{E}} \le \mathrm{Re}_E^x = 177.2\ldots$ for all $0 < \lambda < \infty$. For $\lambda \ge 1$ the substitution $\tilde{\psi} := \lambda \psi$ allows the estimate

$$\begin{aligned} &\frac{|\Re((-\Delta_2)\varphi, \partial_x \partial_z \varphi) + \lambda \, \Re((-\Delta_2)\varphi, \partial_y \psi)|}{\|(-\Delta) \ \ \varphi\|^2 + \lambda \, \| \ \ \psi\|^2} = \frac{|\Re((-\Delta_2)\varphi, \partial_x \partial_z \varphi) + \Re((-\Delta_2)\varphi, \partial_y \tilde{\psi})|}{\|(-\Delta) \ \ \varphi\|^2 + \frac{1}{\lambda} \| \ \ \tilde{\psi}\|^2} \\ &\ge \frac{|\Re((-\Delta_2)\varphi, \partial_x \partial_z \varphi) + \Re((-\Delta_2)\varphi, \partial_y \tilde{\psi})|}{\|(-\Delta) \ \ \varphi\|^2 + \| \ \ \tilde{\psi}\|^2}, \end{aligned}$$

and restricting φ and $\tilde{\psi}$ to functions independent of x furnishes the bound $\mathrm{Re}_{\mathcal{E}} \le \mathrm{Re}_E^y = \mathrm{Re}_E = 82.6\ldots$ for $\lambda \ge 1$. Thus the question remains whether $\mathrm{Re}_{\mathcal{E}}$ does exceed Re_E for some $0 < \lambda < 1$.

A numerical computation indicates that $\mathrm{Re}_{\mathcal{E}}$ attains its upper bound Re_E^x for sufficiently small values of λ (cf. Appendix B). In order to prove this, consider the variational expression

$$\frac{\mathcal{I}_1}{\mathcal{D}_1}[\varphi, \hat{\psi}] = \frac{|\Re((-\Delta_2)\varphi, \partial_x \partial_z \varphi) + \sqrt{\lambda}\, \Re((-\Delta_2)\varphi, \partial_y \hat{\psi})|}{\|(-\Delta) \ \ \varphi\|^2 + \| \ \ \hat{\psi}\|^2}, \tag{3.5}$$

where $\hat{\psi} := \sqrt{\lambda}\,\psi$. Now, inserting the mode expansions (2.8) and (2.9) for φ and $\hat{\psi}$ in (3.5) observe that the maximum for a fixed periodicity cell $\mathcal{P}$ is attained by a single mode. This can be seen as follows: Assume the maximum is attained by a (possibly infinite) linear combination of modes. By inserting this combination into the variational expression (3.5), the numerator as well as the denominator decomposes into a sum of bilinear terms each containing a single mode. Without restriction the modes can be chosen such that the expansion of the numerator contains only nonnegative terms. Applying Lemma 1 (cf. Appendix A) we can select a single mode with maximal ratio, which at most increases the value of the variational expression. Let $\ \in \mathbb{Z}^2 \setminus \{0\}$ be this maximal mode. With the abbreviation $\tilde{\alpha} := \kappa_1\alpha$, $\tilde{\beta} := \kappa_2\beta$, $\tilde{a}(z) := a\ \ (z)$, $\tilde{b}(z) := b\ \ (z)$ we obtain

$$\frac{\mathcal{I}_1}{\mathcal{D}_1}[\tilde{a},\tilde{b},\tilde{\alpha},\tilde{\beta}] = \frac{\tilde{\alpha}\,|\Im(\tilde{a},\tilde{a}')| + \sqrt{\lambda}\,\tilde{\beta}\,|\Im(\tilde{a},\tilde{b})|}{\|(\tilde{\alpha}^2+\tilde{\beta}^2)\tilde{a}-\tilde{a}''\|^2 + (\tilde{\alpha}^2+\tilde{\beta}^2)\|\tilde{b}\|^2 + \|\tilde{b}'\|^2}$$

(3.6)

$$\leq \max\left\{ \frac{\tilde{\alpha}\,|\Im(\tilde{a},\tilde{\alpha}')|}{\tilde{\alpha}^4\|\tilde{a}\|^2 + 2\tilde{\alpha}^2\|\tilde{a}'\|^2 + \|\tilde{a}''\|^2},\ \sqrt{\lambda}\,\frac{\tilde{\beta}\,|\Im(\tilde{a},\tilde{b})|}{\tilde{\beta}^4\|\tilde{a}\|^2 + 2\tilde{\beta}^2\|\tilde{a}'\|^2 + \tilde{\beta}^2\|\tilde{b}\|^2 + \|\tilde{b}'\|^2} \right\},$$

where we used partial integration and Lemma 1 in the last line. Abbreviating the first term in (3.6) with $\mathcal{F}_1[\tilde{a},\tilde{\alpha}]$ and the second with $\mathcal{F}_2[\tilde{a},\tilde{b},\tilde{\alpha}]$, it follows from (3.4)–(3.6) that

$$\text{(3.7)}\qquad {\mathrm{Re}_{\mathcal{E}}}^{-1} = \max\left\{ \sup_{\tilde{\alpha}\in\mathbb{R}_+}\ \sup_{(\tilde{a},0)\in\mathcal{W}} \mathcal{F}_1[\tilde{a},\tilde{\alpha}],\ \sqrt{\lambda}\sup_{\tilde{\beta}\in\mathbb{R}_+}\ \sup_{(\tilde{a},\tilde{b})\in\mathcal{W}} \mathcal{F}_2[\tilde{a},\tilde{b},\tilde{\beta}] \right\},$$

with

$$\mathcal{W} = \left\{ (a,b) \in H^4\left(\left(-\frac{1}{2},\frac{1}{2}\right)\right) \times H^2\left(\left(-\frac{1}{2},\frac{1}{2}\right)\right) \,\middle|\, a = \partial_z a = b = 0 \text{ at } z = \pm\frac{1}{2} \right\} \setminus \{0\}.$$

Inserting the mode expansion (2.8) into (2.16), the first term in (3.7) turns out to be $\frac{1}{\mathrm{Re}_E^x}$, whereas $\mathcal{F}_2[\tilde{a},\tilde{b},\tilde{\beta}]$ is estimated with the help of inequality (A.2) as follows:

$$\begin{aligned}\mathcal{F}_2[\tilde{a},\tilde{b},\tilde{\beta}] &\leq \frac{\tilde{\beta}\,\|\tilde{a}\|\|\tilde{b}\|}{(\tilde{\beta}^4+2\tilde{\beta}^2\pi^2)\|\tilde{a}\|^2 + (\tilde{\beta}^2+\pi^2)\|\tilde{b}\|^2} \leq \frac{\tilde{\beta}}{2[(\tilde{\beta}^4+2\tilde{\beta}^2\pi^2)(\tilde{\beta}^2+\pi^2)]^{1/2}} \\ &\leq \frac{1}{2\sqrt{2}\,\pi^2}.\end{aligned}$$

Therefore, by choosing $\sqrt{\lambda} \leq \frac{2\sqrt{2}\,\pi^2}{\mathrm{Re}_E^x}$, (3.7) yields $\mathrm{Re}_{\mathcal{E}} \geq \mathrm{Re}_E^x$, hence

$$\mathrm{Re}_{\mathcal{E}} = \mathrm{Re}_E^x.$$

We formulate this result in the following proposition.

Proposition 1. *For* $0 < \lambda < \frac{8\pi^4}{\mathrm{Re}_E^x} \approx 0.025$, $0 < \mathrm{Re} < \mathrm{Re}_{\mathcal{E}}$ *with* $\mathrm{Re}_{\mathcal{E}} = \mathrm{Re}_E^x = 177.2\ldots$, $(\alpha,\beta) \in \mathbb{R}_+^2$, *and* $(\varphi,\psi) \in \mathcal{V}_{\alpha,\beta} = D(\tilde{\mathcal{A}}^{1/2}) \setminus \{(0,0)\}$ *with* $D(\tilde{\mathcal{A}}^{1/2})$, *as explained in section* 2, *we have the bound*

$$\text{(3.8)}\qquad \mathrm{Re}\,\frac{\mathcal{I}_1}{\mathcal{D}_1} \leq \frac{\mathrm{Re}}{\mathrm{Re}_{\mathcal{E}}} < 1,$$

where

$$\frac{\mathcal{I}_1}{\mathcal{D}_1} = \frac{|\Re((-\Delta_2)\varphi, \partial_x\partial_z\varphi) + \lambda\,\Re((-\Delta_2)\varphi, \partial_y\psi)|}{\|(-\Delta)\ \ \varphi\|^2 + \lambda\,\|\ \ \psi\|^2}.$$

Remarks. 1. The numerical computation in Appendix B indicates coincidence of $\mathrm{Re}_{\mathcal{E}}$ with Re_E^x for values of λ up to $\lambda \approx 0.042$.

2. Whether other functionals provide even larger stability limits is an open problem. Another candidate which failed to provide a larger stability limit has been discussed in [KT].

4. Nonlinear stability. For $\lambda \neq 1$ the nonlinear term $\mathcal{N}_1$ in (3.2) does not vanish. In order to dominate this term we introduce a second part $\mathcal{E}_2$ of the generalized energy functional $\mathcal{E}$,

$$\mathcal{E}_2[\mathbf{u}, \mathbf{F}] := \frac{1}{2}\{\sigma\|\ \ \mathbf{u}\|^2 + \rho\,|\mathcal{P}|\|\mathbf{F}\|^2\}, \tag{4.1}$$

with yet undetermined nonnegative coupling parameters σ and ρ.

By scalar multiplication of (2.2) with $\sigma\Delta_2\mathbf{u}$ and of $(2.6)_{3,4}$ with ρF_x, ρF_y and using (2.3) and (2.5), we arrive at

$$\partial_t\mathcal{E}_2 = -\mathcal{D}_2 + \mathrm{Re}\,\mathcal{I}_2 + \mathcal{N}_2, \tag{4.2}$$

where[3]

$$\begin{aligned} &\mathcal{D}_2[\mathbf{u}, \mathbf{F}] = \sigma\|\ \ \mathbf{u}\|^2 + \rho\,|\mathcal{P}|\|\mathbf{F}'\|^2, \\ &\mathcal{I}_2[\mathbf{u}, \mathbf{F}] = \sigma\,\Re(\ \ u_z,\ \ u_x), \\ &\mathcal{N}_2[\mathbf{u}, \mathbf{F}] = -\sigma\,\Re(\ \ \mathbf{u}\cdot\nabla\mathbf{u},\ \ \mathbf{u}) - \rho\,\Re(\tilde{\mathbf{u}}\cdot\nabla\tilde{\mathbf{u}}, \mathbf{F}). \end{aligned} \tag{4.3}$$

By defining

$$\Delta\mathrm{Re} := 1 - \frac{\mathrm{Re}}{\mathrm{Re}_{\mathcal{E}}} \tag{4.4}$$

and

$$\mathcal{D} := \Delta\mathrm{Re}\,\mathcal{D}_1 + \mathcal{D}_2, \tag{4.5}$$

the interaction term $\mathrm{Re}\,\mathcal{I}_2$ can be estimated in terms of $\mathcal{D}$:

$$\mathcal{I}_2 \leq \sigma|(\ \ u_z,\ \ u_x)| \leq \sigma^{1/2}\|\ \ (-\Delta_2)\varphi\|\,\sigma^{1/2}\|\ \ \mathbf{u}\| \leq \sigma^{1/2}\mathcal{D}_1^{1/2}(2\mathcal{E}_2)^{1/2}.$$

Using $2\mathcal{E}_2 \leq \frac{\mathcal{D}_2}{\pi^2}$, which follows with (A.2), and setting

$$\sigma := \frac{\pi^2\,\Delta\mathrm{Re}}{\mathrm{Re}^2}, \tag{4.6}$$

we obtain

$$\mathrm{Re}\,\mathcal{I}_2 \leq (\Delta\mathrm{Re})^{1/2}\mathcal{D}_1^{1/2}\mathcal{D}_2^{1/2} \leq \frac{1}{2}\,(\Delta\mathrm{Re}\,\mathcal{D}_1 + \mathcal{D}_2) = \frac{1}{2}\,\mathcal{D}. \tag{4.7}$$

[3]Note that no boundary terms arise in (4.2) since the terms in $\mathcal{E}_2$ differ from those in the ordinary energy at most by *horizontal* derivatives.

We estimate next the nonlinear parts in terms of $\mathcal{D}\mathcal{E}^{1/2}$, $\mathcal{E} := \mathcal{E}_1 + \mathcal{E}_2$, and begin with $\mathcal{N}_1$:

$$\mathcal{N}_1 \le |(\mathbf{u}\cdot\nabla\mathbf{u},\ \ \varphi + \lambda\ \ \psi)| \le \operatorname{ess\,sup}_\Omega |\mathbf{u}|\,\|\nabla\mathbf{u}\|(\|\ \ \varphi\| + \lambda\,\|\ \ \psi\|).$$

The three factors are estimated separately. With (A.8) we obtain for the first factor

$$\|\mathbf{u}\|_\infty \le \frac{C}{\sqrt{2}}\,\|\ \ \tilde{\mathbf{u}}\| + \sqrt{\frac{2}{\pi}}\|\mathbf{F}'\|,$$

where $C = 8(\frac{\sqrt{2}}{m})^{3/2}$, and under the condition

$$\rho \ge \frac{4\,\sigma}{\pi\, C^2|\mathcal{P}|} \tag{4.8}$$

we get further

$$\|\mathbf{u}\|_\infty \le \frac{C}{\sqrt{2\sigma}}\Big\{\sqrt{\sigma}\,\|\ \ \tilde{\mathbf{u}}\| + (\rho\,|\mathcal{P}|)^{1/2}\|\mathbf{F}'\|\Big\} \le \frac{C}{\sqrt{\sigma}}\mathcal{D}_2^{1/2} \le \frac{C}{\sqrt{\sigma}}\mathcal{D}^{1/2}. \tag{4.9}$$

With the conditions

$$\rho \ge \lambda\,\Delta\mathrm{Re}, \qquad 0 < \lambda < 1 \tag{4.10}$$

we obtain for the second factor

$$\|\nabla\mathbf{u}\|^2 = \|\nabla\tilde{\mathbf{u}}\|^2 + |\mathcal{P}|\|\mathbf{F}'\|^2 \le \frac{\mathcal{D}_1}{\lambda} + \frac{\mathcal{D}_2}{\rho} \le \frac{1}{\lambda\,\Delta\mathrm{Re}}\,\mathcal{D};$$

thus

$$\|\nabla\mathbf{u}\| \le \frac{1}{\sqrt{\lambda\,\Delta\mathrm{Re}}}\,\mathcal{D}^{1/2}. \tag{4.11}$$

Finally, we have

$$\|\ \ \varphi\| + \lambda\,\|\ \ \psi\| \le \sqrt{1+\lambda}\,(2\mathcal{E}_1)^{1/2} \le \sqrt{2}\,\sqrt{1+\lambda}\,\mathcal{E}^{1/2}. \tag{4.12}$$

The conditions (4.8) and (4.10) are satisfied for the choice

$$\rho := \Delta\mathrm{Re}\max\Big\{\lambda, \frac{\alpha\,\beta\,m^3}{\sqrt{2}\,2^7\pi\,\mathrm{Re}^2}\Big\}, \tag{4.13}$$

and by collecting the estimates (4.9), (4.11), and (4.12) we have

$$\mathcal{N}_1 \le \frac{\sqrt{2}\,C}{\sqrt{\sigma\,\Delta\mathrm{Re}}}\,\sqrt{1+1/\lambda}\,\mathcal{D}\,\mathcal{E}^{1/2}. \tag{4.14}$$

As to $\mathcal{N}_2$, we obtain by partial integration

$$\begin{aligned}\mathcal{N}_2 &= -\sigma\,\Re\sum_{i,j=1}^{3}\sum_{n=1}^{2}\int_\Omega \partial_n u_i \partial_i u_j \partial_n \overline{u}_j\,d\tau - \rho\,\Re\sum_{i,j=1}^{3}\int_\Omega \tilde{u}_i\partial_i\tilde{u}_j\overline{F}_j\,d\tau\\ &= \sigma\,\Re\sum_{i,j=1}^{3}\sum_{n=1}^{2}\int_\Omega \partial_n u_i \partial_n\partial_i\overline{u}_j u_j\,d\tau + \rho\,\Re\sum_{n=1}^{2}\int_\Omega \tilde{u}_z\tilde{u}_n\overline{F}'_n\,d\tau.\end{aligned}$$

Estimates analogous to those for $\mathcal{N}_1$, in particular (4.9), then yield

$$
\begin{aligned}
\mathcal{N}_2 &\leq \sigma \|\mathbf{u}\|_\infty \| \;\; \mathbf{u}\| \| \;\; \mathbf{u}\| + \rho\, \|\tilde{\mathbf{u}}\|_\infty\, |\mathcal{P}|^{1/2} \|\mathbf{F}'\|\, \|\tilde{u}_z\| \\
&\leq \frac{C}{\sqrt{\sigma}}\, \mathcal{D}^{1/2} \sqrt{\sigma}\, \| \;\; \mathbf{u}\|\, (2\mathcal{E}_2)^{1/2} + \frac{C}{\sqrt{\sigma}}\, \mathcal{D}^{1/2}\, \rho\, |\mathcal{P}|^{1/2} \|\mathbf{F}'\|\, (2\mathcal{E}_1)^{1/2} \\
&\leq \frac{C}{\sqrt{\sigma}}\, \mathcal{D}^{1/2} \sqrt{1+\rho}\, \mathcal{D}_2^{1/2} (2\mathcal{E})^{1/2} \leq \frac{\sqrt{2}\, C}{\sqrt{\sigma}}\, \sqrt{1+\rho}\, \mathcal{D}\, \mathcal{E}^{1/2}.
\end{aligned} \tag{4.15}
$$

Summarizing (4.14) and (4.15) we have

$$
\mathcal{N}_1 + \mathcal{N}_2 \leq \frac{1}{2}\, \mathcal{D} \left(\frac{\mathcal{E}}{\delta} \right)^{1/2} \tag{4.16}
$$

with

$$
\delta := \frac{\sigma}{8\, C^2} \left(\sqrt{1 + 1/\lambda}\, \frac{1}{\sqrt{\Delta \mathrm{Re}}} + \sqrt{1+\rho} \right)^{-2}.
$$

Observe that the estimate (4.15) is based on the estimate (4.9), which increases the number of z-derivatives by only one. Previously used estimates (cf. [GP] or [St]) increase this number by two and do not work in our situation. On the other hand, functionals involving more z-derivatives do not work either, since there is not enough information about boundary values which would allow the necessary partial integrations [KT].

Finally, we add up equations (3.2) and (4.2), apply Proposition 1, and use the estimates (4.7) and (4.16). This yields the following for $\mathcal{E} = \mathcal{E}_1 + \mathcal{E}_2$:

$$
\begin{aligned}
\partial_t \mathcal{E} &= -\left[\mathcal{D}_1 \left(1 - \frac{\mathrm{Re}\, \mathcal{I}_1}{\mathcal{D}_1} \right) + \mathcal{D}_2 \right] + \mathrm{Re}\, \mathcal{I}_2 + \mathcal{N}_1 + \mathcal{N}_2 \\
&\leq -\mathcal{D} + \frac{1}{2}\, \mathcal{D} + \frac{1}{2}\, \mathcal{D} \left(\frac{\mathcal{E}}{\delta} \right)^{1/2} \\
&\leq -\frac{1}{2}\, \mathcal{D} \left[1 - \left(\frac{\mathcal{E}}{\delta} \right)^{1/2} \right].
\end{aligned} \tag{4.17}
$$

Inequality (4.17) implies that $\mathcal{E}(t)$ is monotonically nonincreasing if $\mathcal{E}(0) < \delta$. With

$$
\frac{1}{2}\, \mathcal{D} = \frac{1}{2}\, (\Delta \mathrm{Re}\, \mathcal{D}_1 + \mathcal{D}_2) \geq \pi^2 (\Delta \mathrm{Re}\, \mathcal{E}_1 + \mathcal{E}_2) \geq \pi^2\, \Delta \mathrm{Re}\, \mathcal{E},
$$

which follows from (A.2), (A.3), and $0 < \Delta \mathrm{Re} < 1$, we therefore have

$$
\partial_t \mathcal{E} \leq -\frac{1}{2}\, \mathcal{D} \left[1 - \left(\frac{\mathcal{E}(0)}{\delta} \right)^{1/2} \right] \leq -\pi^2\, \Delta \mathrm{Re}\, \mathcal{E} \left[1 - \left(\frac{\mathcal{E}(0)}{\delta} \right)^{1/2} \right],
$$

and integration yields

$$
\mathcal{E}(t) \leq \mathcal{E}(0) \exp \left\{ -\pi^2\, \Delta \mathrm{Re} \left[1 - \left(\frac{\mathcal{E}(0)}{\delta} \right)^{1/2} \right] t \right\}. \tag{4.18}
$$

We formulate our stability result in the following theorem.

THEOREM 2. *Let us consider perturbations* $\Phi = (\varphi, \psi, \mathbf{F})^{\mathrm{T}}$ *of the basic flow* $\mathbf{U}_0 = \mathrm{Re}(-z, 0, 0)^{\mathrm{T}}$ *in the plane Couette system satisfying globally (in time) the system* (2.6) *as a strong solution (i.e., in the sense of* (2.10)*) under rigid boundary conditions* (2.5) *and being periodic in the horizontal variables* x, y *with wave numbers* $(\alpha, \beta) \in \mathbb{R}^2_+$. *Let* $0 < \mathrm{Re} < \mathrm{Re}_{\mathcal{E}} = 177.2\ldots$, $\Delta\mathrm{Re} = 1 - \frac{\mathrm{Re}}{\mathrm{Re}_{\mathcal{E}}}$, $C = 8(\frac{\sqrt{2}}{m})^{3/2}$, *and* $m = \min(\alpha, \beta)$. *Consider, furthermore, the generalized energy functional*

$$\mathcal{E}[\varphi, \psi, \mathbf{F}] = \frac{1}{2}\left\{\| \ \varphi\|^2 + \lambda \, \| \ \psi\|^2 + \sigma \, \| \ \mathbf{u}\|^2 + \rho \, \frac{4\pi^2}{\alpha\beta} \, \|\mathbf{F}\|^2\right\}$$

with coupling parameters $0 < \lambda < \frac{8\pi^4}{\mathrm{Re}_{\mathcal{E}}}$ *and*

$$\sigma = \frac{\pi^2 \, \Delta\mathrm{Re}}{\mathrm{Re}^2}, \qquad \rho = \Delta\mathrm{Re} \max\left\{\lambda, \frac{\alpha \, \beta \, m^3}{\sqrt{2} \, 2^7 \pi \, \mathrm{Re}^2}\right\}.$$

Then, the solution $(\varphi, \psi, \mathbf{F})$ *of* (2.5) *and* (2.6) *decays in the norm* $\mathcal{E}^{1/2}$ *exponentially to zero provided the initial value satisfies*

$$\mathcal{E}(0) < \delta = \frac{\sigma}{8\,C^2}\left(\sqrt{1 + 1/\lambda}\,\frac{1}{\sqrt{\Delta\mathrm{Re}}} + \sqrt{1+\rho}\right)^{-2}. \tag{4.19}$$

Remarks. 1. The functional $\mathcal{E}$ dominates the classical energy $E = \frac{1}{2}\,\|\mathbf{u}\|^2$. Therefore, $E(t)$ also decays to zero for $\mathrm{Re} < \mathrm{Re}_{\mathcal{E}}$. However, for $\mathrm{Re} > \mathrm{Re}_E$, $E(t)$ does not necessarily decay monotonically.

2. We did not try to obtain optimal (i.e., as large as possible) stability balls δ. Considering the restricted Reynolds number range the stability balls have not yet any importance for experiments. The emphasis of the present paper is on demonstrating that the method of generalized energy functionals also works for rigid boundary conditions.

3. The stability balls δ vanish in the limit $\Delta\mathrm{Re} \to 0$ or $m \to 0$. Asymptotically we have

$$\delta^{1/2} \sim \begin{cases} \Delta\mathrm{Re} & \text{in the limit} \quad \Delta\mathrm{Re} \to 0, \\ m^{3/2} & \text{in the limit} \quad m \to 0. \end{cases}$$

This behavior seems to be intrinsic to the functional method and it is independent of the choice of boundary conditions (cf. [RM]). The decay constant (in time) in (4.18) for a fixed value $\frac{\mathcal{E}(0)}{\delta} = const < 1$ decreases likewise with $\Delta\mathrm{Re}$ to zero, but it is independent of m. This is different from the case of free boundary conditions,[4] where arbitrarily slowly decaying modes always exist; e.g.,

$$\mathbf{u} = e^{-\alpha^2 t} \sin \alpha y \, \mathbf{e}_x, \qquad p \equiv 0$$

for any $\alpha = m > 0$.

4. There is another interesting approach, which is at least in parts rigorous and which aims at providing stability balls of power law type in the Reynolds number; they have the form $c\,\mathrm{Re}^{-\gamma}$, where c depends on the geometry but is independent of Re. The starting point of the method is a power law bound on the resolvent of the linearized

[4]Note that the Poincaré-type inequalities in [RM] have to be corrected; cf. [KX].

operator, which has been obtained so far only by numerical methods. In a second step the exponent γ can then rigorously be derived whereas c remains unknown. Such bounds, valid for all Reynolds numbers, have been obtained for Couette flow [KLH] and have recently been improved [LK].

5. More generally, Theorem 2 applies to (not necessarily global) strong solutions on their maximal intervals of existence. In particular, it provides an a priori bound on the horizontal derivatives of $\mathbf{u}$ in the $L^2(\Omega)$-norm under an explicit condition on its initial values. An interesting (but so far open) question is whether this condition, viz. (4.19), guarantees already global existence of the solution in time. The following is known in this respect [KW]: A strong solution which is conditionally stable in the energy norm on the maximal interval of existence exists globally in time (in the class (2.10)) provided its initial value is small in the norm of the interpolation space $\mathcal{I}$. This norm is, however, stronger than $\mathcal{E}^{1/2}$; in particular, it involves nontangential derivatives of $\mathbf{u}$, which are not controlled by $\mathcal{E}^{1/2}$. The required smallness depends on the steady flow to be perturbed and the stability behavior of the kinetic energy of the perturbation.

Appendix A. We collect in this appendix some more-or-less standard inequalities we made use of in the main text. Only Lemma 3, which presents a refined calculus inequality, is proved.

LEMMA 1. *Let $n \in \mathbb{N}$ and $a_\nu \geq 0$, $b_\nu > 0$ for $1 \leq \nu \leq n$. Then*

$$\frac{\sum_{\nu=1}^{n} a_\nu}{\sum_{\nu=1}^{n} b_\nu} \leq \max\left\{ \frac{a_\nu}{b_\nu} \,\middle|\, 1 \leq \nu \leq n \right\} =: M \tag{A.1}$$

and equality holds if and only if $a_\nu = M\, b_\nu$ for every ν.

Note that inequality (A.1) remains valid for $n \to \infty$.

Frequent use is made of the Poincaré-type inequalities

$$\|f\| \leq \frac{1}{\pi}\|\nabla f\|, \tag{A.2}$$

$$\|\nabla f\| \leq \frac{1}{\pi}\|\nabla\nabla f\| = \frac{1}{\pi}\|\Delta f\|, \tag{A.3}$$

which are valid for $\mathcal{P}$-periodic functions f decomposed according to

$$f(x,y,z) = \frac{1}{\sqrt{\mathcal{P}}} \sum_{\ \in \mathbb{Z}^2} f_{\ }(z) e^{i(\alpha\kappa_1 x + \beta\kappa_2 y)} \tag{A.4}$$

with (at least) $f_{\ } \in H^1((-\frac{1}{2},\frac{1}{2}))$ and (weakly) satisfying the boundary conditions $f_{\ }(\pm\frac{1}{2}) = 0$, $\ \in \mathbb{Z}^2$ (cf. Appendix A in [KX]). The inequalities (A.2) and (A.3) hold likewise for vector valued functions if each component satisfies such a decomposition.

The next two lemmata provide bounds on the sup-norm $\|\cdot\|_\infty = \operatorname{ess\,sup}|\cdot|$ in terms of the L_2-norm $\|\cdot\|_2 = \|\cdot\|$ in one and three dimensions.

LEMMA 2. *Let $f \in H^1((-\frac{1}{2},\frac{1}{2}))$ with (weakly) $f(-\frac{1}{2}) = 0$. Then*

$$\|f\|_\infty^2 \leq 2\,\|f\|\,\|f'\|. \tag{A.5}$$

LEMMA 3. *Let $f : \mathbb{R}^2 \times [-\frac{1}{2},\frac{1}{2}] \to \mathbb{R}$ be $\mathcal{P}$-periodic and decomposed according to* (A.4) *with $f_{\ } \in H^1((-\frac{1}{2},\frac{1}{2}))$ and weakly satisfying the boundary conditions $f_{\ }(\pm\frac{1}{2}) = 0$ for $\ \in \mathbb{Z}^2 \setminus \{0\}$, $f_0 = \frac{1}{\sqrt{|\mathcal{P}|}} \int_{\mathcal{P}} f(x,y,z)\, dxdy = 0$. Then*

$$\|f\|_\infty \leq C\,\|(-\Delta_2)^{1/2}\partial_z f\|^{1/2}\|(-\Delta_2) f\|^{1/2} \tag{A.6}$$

with $C := 8(\frac{\sqrt{2}}{m})^{3/2}$, $m := \min\{\alpha,\beta\}$.

Proof. With (A.4) and Lemma 2 one obtains

$$\begin{aligned}\operatorname{ess\,sup}_{\Omega}|f(x,y,z)| &\leq \operatorname{ess\,sup}_{[-1/2,1/2]} \sum_{\in\mathbb{Z}^2\setminus\{0\}} |f\ \ (z)| \\ &\leq \sqrt{2} \sum_{\in\mathbb{Z}^2\setminus\{0\}} \|f'\ \|^{1/2}\|f\ \ \|^{1/2}.\end{aligned}$$

Therefore, with Hölder's inequality

$$\begin{aligned}\|f\|_\infty &\leq \sqrt{2} \sum_{\in\mathbb{Z}^2\setminus\{0\}} \|f'\ \|^{1/2}(\alpha^2\kappa_1^2+\beta^2\kappa_2^2)^{1/4}\,\|f\ \ \|^{1/2}(\alpha^2\kappa_1^2+\beta^2\kappa_2^2)^{1/2} \\ &\qquad\qquad \times(\alpha^2\kappa_1^2+\beta^2\kappa_2^2)^{-3/4} \\ &\leq \sqrt{2}\left(\sum_{\in\mathbb{Z}^2\setminus\{0\}} \|f'\ \|^2(\alpha^2\kappa_1^2+\beta^2\kappa_2^2)\right)^{1/4}\left(\sum_{\in\mathbb{Z}^2\setminus\{0\}} \|f\ \ \|^2(\alpha^2\kappa_1^2+\beta^2\kappa_2^2)^2\right)^{1/4} \\ &\qquad \times\left(\sum_{\in\mathbb{Z}^2\setminus\{0\}}(\alpha^2\kappa_1^2+\beta^2\kappa_2^2)^{-3/2}\right)^{1/2} \\ &\leq C\,\|(-\Delta_2)^{1/2}\partial_z f\|^{1/2}\|(-\Delta_2)f\|^{1/2}.\end{aligned}$$

In the last line we used the estimate

$$\left(\sum_{\in\mathbb{Z}^2\setminus\{0\}}(\alpha^2\kappa_1^2+\beta^2\kappa_2^2)^{-3/2}\right)^{1/2} \leq 2^{5/2}\left(\frac{\sqrt{2}}{m}\right)^{3/2}$$

(cf. Lemma 4.1 in [BK]). ◻

A more convenient form of (A.6) is

$$\|f\|_\infty \leq \frac{C}{\sqrt{2}}\|\ \ f\|, \tag{A.7}$$

which follows from (A.6) by

$$\begin{aligned}\|f\|_\infty^2 &\leq \frac{C^2}{2}\left[\|(-\Delta_2)^{1/2}\partial_z f\|^2+\|(-\Delta_2)f\|^2\right] \\ &= \frac{C^2}{2}\left[(\Delta_2 f,\partial_z^2 f)+(\Delta_2 f,\Delta_2 f)\right] \\ &= \frac{C^2}{2}(\Delta_2 f,\Delta f) = \frac{C^2}{2}\|\ \ f\|^2.\end{aligned}$$

If f has a nonzero mean value f_0 the inequalities (A.2), (A.5), and (A.7) furnish

$$\|f\|_\infty \leq \|\tilde{f}\|_\infty + \|f_0\|_\infty \leq \frac{C}{\sqrt{2}}\|\ \ \tilde{f}\| + \sqrt{\frac{2}{\pi}}\,\|f_0'\|, \tag{A.8}$$

where $\tilde{f} = f - f_0$.

The inequalities (A.5)–(A.8) hold likewise for vector valued functions if each component satisfies the appropriate conditions.

Appendix B. In this appendix the variational problem (3.4) with $0 \le \lambda \le 1$ is treated on a numerical basis. We first solve the eigenvalue problem associated with the variational problem with fixed periodicity cell $\mathcal{P}$ and subsequently perform the variation with respect to $\mathcal{P}$.

The Euler–Lagrange equations with Lagrange parameter μ read

$$\begin{aligned} \Delta^2(-\Delta_2)\varphi - \frac{\mu}{2}\Big(2\,(-\Delta_2)\partial_x\partial_z\varphi + \lambda\,(-\Delta_2)\partial_y\psi\Big) &= 0, \\ \lambda\,(-\Delta)(-\Delta_2)\psi + \frac{\mu}{2}\,\lambda\,(-\Delta_2)\partial_y\varphi &= 0. \end{aligned} \tag{B.1}$$

By inserting the mode expansions (2.8) and (2.9) the system (B.1) becomes equivalent to

$$\left.\begin{aligned} D^2_{\kappa_1\alpha,\kappa_2\beta}\, a\ \ (z) - i\frac{\mu}{2}\Big(2\alpha\kappa_1\partial_z a\ \ (z) + \lambda\,\beta\kappa_2 b\ \ (z)\Big) &= 0, \\ D_{\kappa_1\alpha,\kappa_2\beta}\, b\ \ (z) + i\frac{\mu}{2}\beta\kappa_2 a\ \ (z) &= 0, \end{aligned}\right. \qquad \in \mathbb{Z}^2 \setminus \{0\} \tag{B.2}$$

with $D_{\tilde\alpha,\tilde\beta} := \tilde\alpha^2 + \tilde\beta^2 - \partial_z^2$. The system (B.2) has to be complemented with the boundary conditions

$$a\ \ = \partial_z a\ \ = b\ \ = 0 \quad \text{at} \quad z = \pm\frac{1}{2}, \qquad \in \mathbb{Z}^2 \setminus \{0\},$$

in order to have a well-posed eigenvalue problem. As explained in section 3, the maximum is attained by a single mode. Since we are ultimately interested in the maximum with respect to all periodicity cells, it is sufficient to consider the finite dimensional system

$$\begin{aligned} D^2_{\tilde\alpha,\tilde\beta}\,\tilde a(z) - i\frac{\mu}{2}\Big(2\tilde\alpha\partial_z\tilde a(z) + \lambda\,\tilde\beta\tilde b(z)\Big) &= 0, \\ D_{\tilde\alpha,\tilde\beta}\,\tilde b(z) + i\frac{\mu}{2}\tilde\beta\tilde a(z) &= 0 \end{aligned} \tag{B.3}$$

together with

$$\tilde a = \partial_z\tilde a = \tilde b = 0 \quad \text{at} \quad z = \pm\frac{1}{2}. \tag{B.4}$$

$\mathrm{Re}_{\mathcal{E}}$ is then given by

$$\mathrm{Re}_{\mathcal{E}} = \min_{(\tilde\alpha,\tilde\beta)\in\mathbb{R}^2} \mu_0(\tilde\alpha,\tilde\beta,\lambda),$$

with μ_0 being the smallest positive eigenvalue in (B.3) and (B.4). Applying a standard shooting method based on a fourth order Runge–Kutta integration, μ_0 is determined as a function of $\tilde\alpha$, $\tilde\beta$, and λ. Subsequent minimization with respect to $\tilde\alpha$ and $\tilde\beta$ furnishes $\mathrm{Re}_{\mathcal{E}}$ as a function of λ. The result is displayed in Figure 1: With decreasing λ the stability limit $\mathrm{Re}_{\mathcal{E}}$ increases from the ordinary energy limit $\mathrm{Re}_E = 82.6\ldots$ ($\lambda = 1$) up to the value $\mathrm{Re}_E^x = 177.2\ldots$ (Figure 1, left), and this value is, in fact, attained for finite λ ($\lambda \approx 0.042$; see Figure 1, right).

Acknowledgment. The authors would like to thank an anonymous referee whose comments led to an improvement of the manuscript.

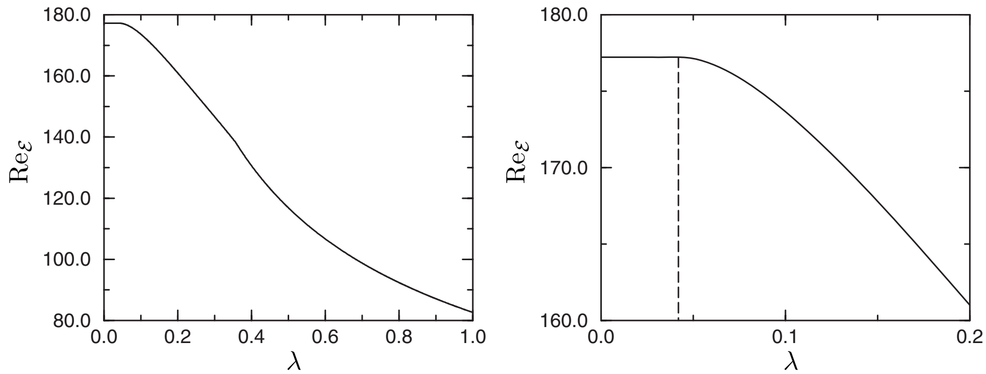


FIG. 1. *The generalized energy limit* $\mathrm{Re}_{\mathcal{E}}$ *versus coupling parameter* λ *with* $\mathcal{E}_1$ *given in* (3.1). *In the left graph,* λ *covers the range between* 0 *and* 1 *(*$\lambda = 1$ *corresponds to the ordinary energy); the right graph magnifies the region close to* $\lambda = 0$.

REFERENCES

[BK] M. BIRNER AND R. KAISER, *Stability balls in the rotating Bénard system*, J. Math. Fluid Mech., 6 (2004), pp. 311–364.

[Bu] F. H. BUSSE, *A property of the energy stability limit for plane parallel shear flow*, Arch. Rational Mech. Anal., 47 (1972), pp. 28–35.

[DD] O. DAUCHOT AND F. DAVIAUD, *Finite-amplitude perturbation in plane Couette flow*, Europhys. Lett., 28 (1994), pp. 225–230.

[DR] P. G. DRAZIN AND W. H. REID, *Hydrodynamic Stability*, Cambridge University Press, Cambridge, UK, 1981.

[GP] G. P. GALDI AND M. PADULA, *A new approach to the energy theory in the stability of fluid motion*, Arch. Rational Mech. Anal., 110 (1990), pp. 187–286.

[Gr] S. GROSSMANN, *The onset of shear flow turbulence*, Rev. Modern Phys., 72 (2000), pp. 603–618.

[Jo] D. D. JOSEPH, *Nonlinear stability of the Boussinesq equations by the method of energy*, Arch. Rational Mech. Anal., 22 (1966), pp. 163–184.

[J1] D. D. JOSEPH, *Stability of Fluid Motions*, Vol. I, Springer Tracts Nat. Philos. 27, Springer-Verlag, Berlin, 1976.

[J2] D. D. JOSEPH, *Stability of Fluid Motions*, Vol. II, Springer Tracts Nat. Philos. 28, Springer-Verlag, Berlin, 1976.

[KS] R. KAISER AND B. SCHMITT, *Bounds on the energy stability limit of plane parallel shear flows*, Z. Angew. Math. Phys., 52 (2001), pp. 573–596.

[KT] R. KAISER AND A. TILGNER, *On the generalized energy method for channel flows*, in Proceedings of the 11th Conference on Waves and Stability in Continuous Media (Porto Ercole, 2001), R. Monaco, M. P. Bianchi, and S. Rionero, eds., World Scientific, River Edge, NJ, 2002, pp. 259–270.

[KW] Y. KAGEI AND W. WAHL, *Stability of higher norms in terms of energy–stability for the Boussinesq–equations: Remarks on the asymptotic behaviour of convection–roll–type solutions*, Differential Integral Equations, 7 (1994), pp. 921–948.

[KX] R. KAISER AND L. X. XU, *Nonlinear stability of the rotating Bénard problem, the case* $Pr = 1$, NoDEA Nonlinear Differential Equations Appl., 5 (1998), pp. 283–307.

[KLH] G. KREISS, A. LUNDBLADH, AND D. S. HENNINGSON, *Bounds for threshold amplitudes in subcritical shear flow*, J. Fluid Mech., 270 (1994), pp. 175–198.

[LK] M. LIEFVENDAHL AND G. KREISS, *Bounds for the threshold amplitude for plane Couette flow*, J. Nonlinear Math. Phys., 9 (2002), pp. 311–324.

[Or] W. MCF. ORR, *The stability or instability of steady motions of a liquid, Part* II: *A viscous liquid*, Proc. Roy. Irish Acad. Sect. A, 27 (1907), pp. 69–138.

[RM] S. RIONERO AND G. MULONE, *On the nonlinear stability of parallel shear flows*, Contin. Mech. Thermodyn., 3 (1991), pp. 1–11.

[Ro] V. A. ROMANOV, *Stability of plane-parallel Couette flow*, Funct. Anal. Appl., 7 (1973), pp. 137–146.

[SH] P. J. Schmid and D. S. Henningson, *Stability and Transition in Shear Flows*, Springer-Verlag, New York, 2001.

[St] B. Straughan, *The Energy Method, Stability and Nonlinear Convection*, 2nd ed., Springer-Verlag, New York, 2004.

[SW] B. J. Schmitt and W. von Wahl, *Decomposition of solenoidal fields into poloidal fields, toroidal fields and the mean flow. Applications to the Boussinesq-equations*, The Navier-Stokes Equations II—Theory and Numerical Methods, Lecture Notes in Math. 1530, J. G. Heywood, K. Masuda, R. Rautmann, and S. A. Solonnikov, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1992, pp. 291–305.

[TTRD] L. N. Trefethen, A. E. Trefethen, S. C. Reddy, and T. A. Driscoll, *Hydrodynamic stability without eigenvalues*, Science, 261 (1993), pp. 578–584.

[Wa] W. von Wahl, *Necessary and sufficient conditions for the stability of flows of incompressible viscous fluids*, Arch. Rational Mech. Anal., 126 (1994), pp. 103–129.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 455–489

WRIGGLED LAMELLAR SOLUTIONS AND THEIR STABILITY IN THE DIBLOCK COPOLYMER PROBLEM*

XIAOFENG REN† AND JUNCHENG WEI‡

Abstract. In a diblock copolymer system the free energy field depends nonlocally on the monomer density field. In addition there are two positive parameters in the constitutive relation. One of them is small, with respect to which we do singular perturbation analysis. The second one is of order 1, with respect to which we do bifurcation analysis. Combining the two techniques we find wriggled lamellar solutions of the Euler–Lagrange equation of the total free energy under a hypothesis regarding the simplicity of the principal eigenvalue, which is generically satisfied. The wriggled lamellar solutions bifurcate from the perfect lamellar solutions. The stability of the wriggled lamellar solutions is reduced to a relatively simple finite dimensional problem, which may be solved accurately by a numerical method. Our tests show that most of them are stable. The existence of such stable wriggled lamellar solutions explains why in reality the lamellar phase is fragile and often exists in distorted forms.

Key words. distortion, bifurcation, singular perturbation, stability, wriggled lamellar solution, perfect lamellar solution, diblock copolymer

AMS subject classifications. 58E07, 35J55, 34D15, 45J05, 82D60

DOI. 10.1137/S0036141003433589

1. Introduction.

Symmetry breaking distortions often appear for intrinsic reasons in systems of condensed matters that exhibit self-organization and pattern formation. We study this phenomenon in diblock copolymers. A diblock copolymer is a soft material, characterized by fluidlike disorder on the molecular scale and a high degree of order at longer length scales. A molecule in a diblock copolymer melt is a linear subchain of A monomers grafted covalently to another subchain of B monomers. Because of the repulsion between the unlike monomers, the different type subchains tend to segregate below some critical temperature, but as they are chemically bonded in chain molecules, even a complete segregation of subchains cannot lead to a macroscopic phase separation. Only a local microphase separation occurs: microdomains rich in A and B are formed. These microdomains form morphology patterns/phases in a larger scale. The most commonly observed undistorted phases are the spherical, cylindrical, and lamellar, depicted in Figure 1. Here we seek distorted, defective lamellar patterns, where the interfaces separating the microdomains, unlike the ones in Figure 1(c), are wriggled (Figure 2(b)).

We consider a scenario that a diblock copolymer melt is placed in a domain D and maintained at fixed temperature. D is scaled to have unit volume in space. Let $a \in (0,1)$ be the relative number of the A monomers in a chain molecule, and $b = 1-a$ be the relative number of the B monomers in a chain. The relative A monomer density field u is an order parameter. $u \approx 1$ stands for high concentration of A monomers.

*Received by the editors August 25, 2003; accepted for publication (in revised form) January 12, 2005; published electronically October 14, 2005.
http://www.siam.org/journals/sima/37-2/43358.html

†Department of Mathematics and Statistics, Utah State University, Logan, UT 84322-3900 (ren@math.usu.edu).
‡Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong (wei@math.cuhk.edu.hk). The work of this author was supported in part by a Direct Grant from CUHK and an Earmarked Grant of RGC of Hong Kong.

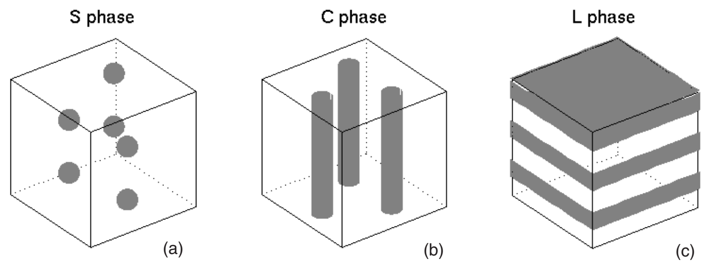


FIG. 1. *The* (a) *spherical,* (b) *cylindrical, and* (c) *lamellar morphology phases commonly observed in diblock copolymer melts. The dark color indicates the concentration of type A monomer, and the white color indicates the concentration of type B monomer.*

The melt is incompressible, and thus the relative B monomer density is $1-u$ and $u \approx 0$ stands for high concentration of B monomers.

Ohta and Kawasaki [11] introduced an equilibrium theory in which the free energy of the system is a functional of the relative A monomer density,

$$I(u) = \int_D \left\{ \frac{\epsilon^2}{2}|\nabla u|^2 + \frac{\epsilon\gamma}{2}|(-\Delta)^{-1/2}(u-a)|^2 + W(u) \right\}, \tag{1.1}$$

defined in $X_a = \{u \in W^{1,2}(D) : \overline{u} = a\}$, where $\overline{u} := \int_D u$ is the average of u on D. The original formula in [11] is given on the entire $\mathbf{R}^3$. The expression here on a bounded domain D first appeared in Nishiura and Ohnishi [9]. A mathematically more rigorous derivation is in Choksi and Ren [3]. The local function W is smooth and has the shape of a double well. It has the global minimum value 0 at two numbers: 0 and 1. To avoid unnecessary technical difficulties we assume that $W(p) = W(1-p)$. The two global minimum points are nondegenerate: $W''(0) = W''(1) \neq 0$.

The most unusual feature in (1.1) is the nonlocal expression $(-\Delta)^{-1/2}(u-a)$. It reflects the connectivity of polymer chains. $(-\Delta)^{-1/2}$ is the square root of the positive operator $(-\Delta)^{-1}$ from $\{w \in L^2(D) : \overline{w} = 0\}$ to itself. The integral of the nonlocal part in (1.1) may be rewritten as

$$\int_D |(-\Delta)^{-1/2}(u-a)|^2 = \int_D \int_D G_D(x,y)(u(x)-a)(u(y)-a)\,dxdy.$$

G_D is the Green function of $-\Delta$ with the Neumann boundary condition. It splits to a fundamental solution part and a regular part. The fundamental solution in $\mathbf{R}^3$ is

$$\frac{1}{4\pi|x-y|},$$

a long range Coulomb-type interaction, which is common in many important physical systems (see Muratov [7]).

The dimensionless parameters ϵ and γ are positive and depend on various physical quantities [3]. In the strong segregation region where morphology patterns form, ϵ is very small. The second parameter γ is of order 1 when we choose the size of the sample to be comparable to the size of the microdomains [3]. We develop a particular two-parameter perturbation method. We do singular perturbation analysis

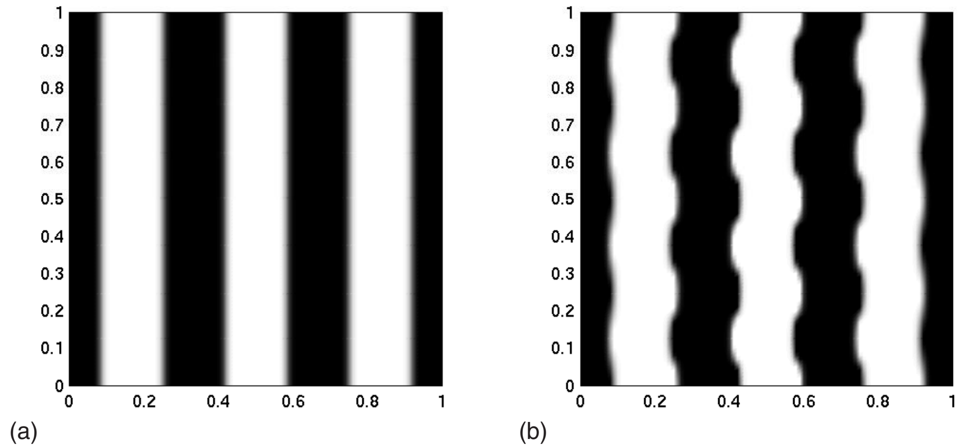


FIG. 2. *A perfect lamellar solution* (a) *and a wriggled lamellar solution* (b)*. In the dark regions the solutions are close to* 1 *and in the light regions the solutions are close to* 0*.*

with respect to ϵ and bifurcation analysis with respect to γ. The challenge is to combine these two techniques to derive fine analytical results. Even though this mathematical method is tailored for the diblock copolymer problem, we believe that it may be applied to other ones with multiple parameters. Examples include the Seul–Andelman membrane problem [25, 19], charged Langmuir monolayers [1, 16], and smectic films [24].

The Euler–Lagrange equation of I is

$$-\epsilon^2\Delta u + f(u) - \overline{f(u)} + \epsilon\gamma(-\Delta)^{-1}(u-a) = 0, \quad \partial_\nu u = 0 \text{ on } \partial D. \tag{1.2}$$

The function f is the derivative of W. The term $\overline{f(u)}$ is equal to the Lagrange multiplier corresponding to the constraint $\overline{u} = a$. Equation (1.2) may also be written as an elliptic system:

$$\begin{cases} -\epsilon^2\Delta u + f(u) + \epsilon\gamma v = \text{const}, \\ -\Delta v = u - a, \\ \partial_\nu u = \partial_\nu v = 0 \text{ on } \partial D, \\ \overline{u} - a = \overline{v} = 0. \end{cases} \tag{1.3}$$

Here const is the Lagrange multiplier.

In Ren and Wei [12] a family of lamellar solutions is found. When $D = (0,1)$, for each positive integer K there exists a 1-dimensional (1-D) local minimizer of I having K sharp interfaces if ϵ is sufficiently small (see Theorem 2.1). This 1-D local minimizer may be extended trivially to a three-dimensional (3-D) solution of (1.2) on a box. Such a solution, illustrated in Figure 2(a), models the lamellar phase, Figure 1(c), only if it is stable in the sense that it is a local minimizer of I in three dimensions. A local minimizer in three dimensions is called a *metastable* state of the physical system. It survives mild thermal fluctuation.

However, in Ren and Wei [15] it is shown that such 1-D local minimizers are not necessarily 3-D local minimizers after trivial extension. Detailed spectral information at each 1-D solution is found (see Theorem 3.1; a similar analysis for the FitzHugh–Nagumo system was carried out by Nishiura [8]). In summary, these 1-D local minimizers are not 3-D local minimizers unless K is sufficiently large or γ is sufficiently small. Moreover, the 1-D global minimizer, which is one of the 1-D local minimizers with the optimal number of interfaces

$$K_{opt} \approx \left(\frac{a^2 b^2 \gamma}{3\tau}\right)^{1/3},$$

where τ is defined in (2.2), has a delicate stability property. It actually lies near the *borderline* that separates the stable 1-D solutions from the unstable 1-D solutions.

All this suggests that the lamellar phase is only a metastable, transient state of the material. Thermal fluctuation will eventually destroy this phase. In reality one often observes the lamellar phase in distorted forms. We predict based on the model (1.1) that a defective, wriggled lamellar pattern (Figure 2(b)) exists in diblock copolymers. We point out that the wriggled lamellar pattern is typically observed in systems with competing interactions [25].

To simplify the presentation of our results, we consider wriggled lamellar solutions in two dimensions. We take $D = (0,1) \times (0,1)$ to be a 2-D square instead of a 3-D box. The 1-D local minimizers are trivially extended to D. Generalization of our results to 3-D is straightforward.

The existence of wriggled lamellar solutions is shown by a bifurcation analysis. Each perfect lamellar solution u_γ with K interfaces is stable when γ is sufficiently small. The spectrum of the second variation of I at u_γ, which consists of real eigenvalues only, lies to the right of 0. If we increase γ, the spectrum moves to the left. When γ reaches a critical value γ_{B}, the principal (smallest) eigenvalue in the spectrum becomes 0. Under Hypothesis 4.1, which is generically satisfied, the principal eigenvalue is simple. Then a new solution branch bifurcates out of $u_{\gamma_{\mathrm{B}}}$. This is a wriggled lamellar solution (Figure 2(b)). If we further increase γ, then another eigenvalue of u_γ, which is not the principal eigenvalue, may become 0, and another new solution also of a wriggled lamellar pattern bifurcates from u_γ. However, wriggled lamellar solutions that bifurcate from larger eigenvalues are unstable and physically less interesting.

Whether the wriggled lamellar solution associated with the principal eigenvalue of $u_{\gamma_{\mathrm{B}}}$ is stable is a subtle question. It is relatively easy to see that the bifurcation diagram has the shape of a pitchfork (Figure 3). The stability of the wriggled solution depends on the direction of the fork. Here we face a formidable problem. The direction is determined by the sign of a number which turns out to be terribly small (of order ϵ^5; see Lemma 5.2). To find this number we have to expand the "trivial solution" $u_{\gamma_{\mathrm{B}}}$, its principal eigenfunction corresponding to the 0 eigenvalue, and the third function $g'(0)$ defined in (5.4), with respect to ϵ. As we prove Lemma 5.2, these expansions have to be carefully managed. All the lower order terms up to ϵ^4 vanish. In the end we arrive at a quantity $S(a,K)$ that depends on a and K only. The bifurcating solution is stable if $S(a,K) > 0$ and unstable if $S(a,k) < 0$. The quantity $S(a,K)$ may be accurately calculated by a simple numerical method. Our tests, reported in section 5, show that for most values of a and K the wriggled lamellar solution bifurcating out of the principal eigenvalue is stable.

Our study of defects in diblock copolymers is partially motivated by the work of Tsori, Andelman, and Schick [27]. They considered two tilt lamellar patterns: chevron morphology and omega morphology. They used a model different from (1.1), which agrees well with experiments in the so-called weak segregation regime. In this regime interfaces separating microdomains are not sharp compared to the size of the microdomains. In our work we deal with the strong segregation regime using (1.1). Here microdomains are separated by sharp interfaces. The calculations in [27] are restricted to an ansatz without solving the Euler–Lagrange equation of their free energy functional. The wriggled lamellar patterns we find are solutions of the Euler–Lagrange equation (1.2). Many of them are proved to be stable. They are therefore genuine metastable states of the system.

Moreover, we are able to answer the question whether a 1-D global minimizer, mentioned earlier, is stable in 2-D. We will find cases where the 1-D global minimizer is stable and also cases where the 1-D global minimizer is unstable.

The paper is organized as follows. In section 2 we recall some properties of the perfect lamellar solutions u_γ. Section 3 contains some spectral information of the second variation of I at u_γ. Hypothesis 4.1, which guarantees the simplicity of the principal eigenvalue, is given in section 4. Under this condition the existence of the wriggled lamellar solutions is given in Theorem 4.2. The reduction of their stability to the positivity of $S(a,K)$ culminates in Theorem 5.4. The stability of the 1-D global minimizer is discussed after Theorem 5.4. The lengthy calculations that prove Lemma 5.2 are in Appendices B and C.

To avoid clumsy notations a quantity's dependence on ϵ is usually suppressed. For example, we write u, the lamellar solution, instead of u_ϵ. On the other hand we often emphasize a quantity's independence of ϵ with a superscript 0. For example, the limit of a lamellar solution u as $\epsilon \to 0$ is denoted by u^0. The parameter γ may vary in a compact interval of positive numbers. It is understood in this paper that all estimates are uniform in such γ. In estimates C is always a positive constant independent of ϵ and γ. Its value may vary from line to line. The L^2 inner product is denoted by $\langle \cdot, \cdot \rangle$ and the L^p norm by $\| \cdot \|_p$.

References on the mathematical aspects of the block copolymer theory include, in addition to the ones cited already, Ohnishi et al. [10], Choksi [2], Fife and Hilhorst [5], Henry [6], and Ren and Wei [14, 13, 20, 21, 22] on diblock copolymers, and Ren and Wei [17, 18] on triblock copolymers.

2. The perfect lamellar solution. The perfect lamellar solutions that serve as "trivial solutions" in the bifurcation theory are constructed in [12] by the Γ-limit theory. The findings there are summarized in the following theorem.

THEOREM 2.1 (see Ren and Wei [12]). *In 1-D for each positive integer K the functional*

$$I_1(u) = \int_0^1 \left\{ \frac{\epsilon^2}{2} \left(\frac{du}{dx} \right)^2 + \frac{\epsilon\gamma}{2} \left| \left(-\frac{d^2}{dx^2} \right)^{-1/2} (u-a) \right|^2 + W(u) \right\} dx,$$

in $\{u \in W^{1,2}(0,1) : \overline{u} = a\}$, has a local minimizer u near u^0, under the L^2 norm, when ϵ is sufficiently small. It satisfies the Euler–Lagrange equation

$$-\epsilon^2 u'' + f(u) - \overline{f(u)} + \epsilon\gamma G_0[u-a] = 0, \ u'(0) = u'(1) = 0$$

and has the properties

$$\lim_{\epsilon\to 0} \|u - u^0\|_2 = 0 \text{ and } \lim_{\epsilon \to 0} \epsilon^{-1} I_1(u) = \tau K + \frac{\gamma}{2} \int_0^1 \left| \left(-\frac{d^2}{dx^2} \right)^{-1/2} (u^0 - a) \right|^2 dx.$$

Let H be the solution of

$$-H'' + f(H) = 0 \text{ in } \mathbf{R}, \ H(-\infty) = 0, \ H(\infty) = 1, \ H(0) = \frac{1}{2}. \tag{2.1}$$

The constant τ in the theorem is defined by

$$\tau := \int_{\mathbf{R}} (H'(t))^2\, dt. \tag{2.2}$$

τ is often called the surface tension in the literature. u^0 is a step function with K jump discontinuity points, defined to be

$$u^0(x) = 1 \text{ on } (0, x_1^0),\ 0 \text{ on } (x_1^0, x_2^0),\ 1 \text{ on } (x_2^0, x_3^0),\ 0 \text{ on } (x_3^0, x_4^0),\ 1 \text{ on } (x_4^0, x_5^0), \dots$$

with (recall $b = 1 - a$)

$$x_1^0 = \frac{a}{K},\ x_2^0 = \frac{1+b}{K},\ x_3^0 = \frac{2+a}{K},\ x_4^0 = \frac{3+b}{K},\ x_5^0 = \frac{4+a}{K}, \dots. \tag{2.3}$$

G_0 is the solution operator of $-v'' = g$, $v'(0) = v'(1) = \overline{v} = 0$; i.e., $v = G_0[g] = (-\frac{d^2}{dx^2})^{-1} g$.

There is another K-interface 1-D local minimizer whose limiting value as $\epsilon \to 0$ is 0 instead of 1 on the first interval $(0, b/K)$. It is just $1 - \tilde{u}$, where $\tilde{u}$ is a solution constructed in Theorem 2.1, but with $\overline{\tilde{u}} = b$ instead. $1 - \tilde{u}$ has the same properties as u does, so we focus on u. u is periodic in the following sense.

Theorem 2.2 (see Ren and Wei [15]). *Let u be a 1-D local minimizer constructed in Theorem 2.1. When ϵ is small, for every $x \in (0, 1/K)$,*

$$\begin{aligned} u(x) &= u\left(\frac{2}{K} - x\right) = u\left(x + \frac{2}{K}\right) = u\left(\frac{4}{K} - x\right) \\ &= u\left(x + \frac{4}{K}\right) = \cdots = \begin{cases} u(1-x) & \text{if } K \text{ is even,} \\ u(x + \frac{K-1}{K}) & \text{if } K \text{ is odd.} \end{cases} \end{aligned}$$

Moreover, when ϵ is small, u is the unique local minimizer of I_1 in an L^2 neighborhood of u^0. If u on $((j-1)/K, j/K)$ for some $j = 1, 2, \dots, K$ is scaled to a function on $(0,1)$, then it is exactly a one-layer local minimizer of I_1 with ϵ and γ replaced by $\tilde{\epsilon} = \epsilon K$ and $\tilde{\gamma} = \gamma/K^3$.

Let us denote this u of K interfaces by u_γ, to emphasize its dependence on γ. We need asymptotic expansions of u_γ in terms of ϵ. According to [15, Lemma A.1] there exist exactly K points x_j, $j = 1, 2, \dots, K$ in $(0,1)$ so that $u(x_j) = 1/2$. These K points identify the interfaces of u. Theorem 2.2 implies that $x_2 = \frac{2}{K} - x_1$, $x_3 = \frac{4}{K} - x_2$, $x_4 = \frac{6}{K} - x_3$, etc. The zeroth order approximation of u_γ is

$$\begin{aligned} s(x) &= H\left(-\frac{x - x_1}{\epsilon}\right) + H\left(\frac{x - x_2}{\epsilon}\right) \\ &\quad + H\left(-\frac{x - x_3}{\epsilon}\right) - 1 + \cdots + \begin{cases} H(\frac{x - x_K}{\epsilon}) & \text{if } K \text{ is even,} \\ H(-\frac{x - x_K}{\epsilon}) - 1 & \text{if } K \text{ is odd.} \end{cases} \end{aligned} \tag{2.4}$$

The ϵ order outer expansion term is z^0, defined to be

$$z^0(x) = -\frac{\gamma(v^0(x) - v^0(x_j^0))}{f'(0)},\quad v^0 = \left(-\frac{d^2}{dx^2}\right)^{-1} (u^0 - a), \tag{2.5}$$

and the ϵ order inner expansion term is 0. Because of the periodicity, $v^0(x_j^0)$ is independent of j and z^0 is well defined. The ϵ^2 order inner expansion term is P, where P is the solution of

$$-P'' + f'(H)P = -\gamma(v^0)'(x_j^0)t,\ P \perp H'. \tag{2.6}$$

There are two different P's depending on whether j is odd or even. But they just differ by a sign, and it is always easy to tell from the context which one is referred to.

LEMMA 2.3 (see Ren and Wei [15]). *Let z be defined by $u_\gamma = s + \epsilon z$.*

1. $\|z - z^0\|_\infty = O(\epsilon)$.
2. *There exists a constant $C > 0$ independent of ϵ so that $|\epsilon^{-1} z(x_j + \epsilon t)| \le C(1 + |t|)$ for all $t \in (-\frac{x_j}{\epsilon}, \frac{1-x_j}{\epsilon})$. $\epsilon^{-1} z(x_j + \epsilon\cdot)$ converges to P in $C^2_{loc}(\mathbf{R})$.*

Proof. See [15, Lemmas 2.4 and 2.5]. ∎

It is proved in [15] that u_γ is a nondegenerate 1-D local minimizer in the sense that the 1-D spectrum at u_γ lies to the right of the origin (see part 2 of Theorem 3.1). This allows us to apply the implicit function theorem to conclude that u_γ depends on γ smoothly.

Next we estimate $du_\gamma/d\gamma$. The inner approximation is

$$q_j(x) = cH'\left(\frac{x - x_j}{\epsilon}\right) + \epsilon^2 R\left(\frac{x - x_j}{\epsilon}\right) \tag{2.7}$$

near each x_j. The constant c is chosen so that

$$\frac{du_\gamma}{d\gamma}(x_j) = cH'(0). \tag{2.8}$$

We will show in Appendix A that

$$\lim_{\epsilon\to 0} c = c_0 := -\frac{v^0(x^0_j)}{Kf'(0)}. \tag{2.9}$$

The function R is the solution of

$$-R'' + f'(H)R + c_0 f''(H)H'P + tv'(x_j) = \zeta,\ R(0) = 0. \tag{2.10}$$

The constant ζ is chosen so that

$$\int_{\mathbf{R}} (c_0 f''(H)H'P + v(x_j) - \zeta)H'\,dt = 0 \tag{2.11}$$

for solvability. Because of the periodicity of u_γ in Theorem 2.2, neither c nor R depends on j. However, they do depend on ϵ. Our argument here is a bit different from the formal matched asymptotics method. If we compare (2.10) to (2.6) we find that

$$\lim_{\epsilon\to 0} R = \frac{1}{\gamma}P + \tilde{R}, \tag{2.12}$$

where $\tilde{R}$ is the solution of

$$-\tilde{R}'' + f'(H)\tilde{R} + c_0 f''(H)H'P = \zeta_0,\ \tilde{R}(0) = 0, \tag{2.13}$$

where ζ_0 satisfies

$$\int_{\mathbf{R}} (c_0 f''(H)H'P + v^0(x^0_j) - \zeta)H'\,dt = 0. \tag{2.14}$$

The inner approximation (2.7) is used in the interval $(x_j - \epsilon^\alpha, x_j + \epsilon^\alpha)$, where α satisfies

$$\alpha \in \left(\frac{1}{2}, 1\right). \tag{2.15}$$

The outer approximation is denoted by q_o:

$$q_o = \frac{\eta - \epsilon\gamma r - \epsilon v}{f'(u_\gamma)}, \tag{2.16}$$

where

$$r = G_0\left[\frac{du_\gamma}{d\gamma}\right], \; v = G_0[u_\gamma - a]. \tag{2.17}$$

The outer approximation is used in

$$[0,1]\backslash(\cup_{j=1}^K (x_j - 2\epsilon^\alpha, x_j + 2\epsilon^\alpha)). \tag{2.18}$$

The inner approximation is matched to the outer approximation in the matching intervals $(x_j - 2\epsilon^\alpha, x_j - \epsilon^\alpha)$ and $(x_j + \epsilon^\alpha, x_j + 2\epsilon^\alpha)$, $j = 1, 2, \ldots, K$. Let χ_j be smooth cut-off functions so that

$$\chi_j(x) = \begin{cases} 0 & \text{if } x \notin (x_j - 2\epsilon^\alpha, x_j + 2\epsilon^\alpha), \\ 1 & \text{if } x \in (x_j - \epsilon^\alpha, x_j + \epsilon^\alpha); \end{cases}$$

moreover, $\chi_j' = O(\epsilon^{-\alpha})$ and $\chi_j'' = O(\epsilon^{-2\alpha})$ in $(x_j - 2\epsilon^\alpha, x_j - \epsilon^\alpha)$ and $(x_j + \epsilon^\alpha, x_j + 2\epsilon^\alpha)$. We glue the two approximations to form a uniform approximation

$$q = \sum_{j=1}^K \chi_j q_j + \left(1 - \sum_{j=1}^K \chi_j\right) q_o. \tag{2.19}$$

Lemma 2.4. $q - u = o(\epsilon^2)$.

The proof of this lemma is technical. We include it in Appendix A.

3. The 2-D spectrum at . The 1-D local minimizer u_γ of I_1 is now viewed as a function on D, through extension to the second dimension trivially, so $u_\gamma(x, y) = u_\gamma(x)$. It is a solution of (1.2) and $I_1(u_\gamma) = I(u_\gamma)$. In two dimensions it has straight interfaces. We call it a perfect lamellar solution of (1.2).

The linearized operator of (1.2) at u_γ is

$$L_\gamma\varphi := -\epsilon^2\Delta\varphi + f'(u_\gamma)\varphi - \overline{f'(u_\gamma)\varphi} + \epsilon\gamma(-\Delta)^{-1}\varphi, \;\; \varphi \in W^{2,2}(D), \;\; \partial_\nu\varphi = 0 \text{ on } \partial D, \; \overline{\varphi} = 0. \tag{3.1}$$

This is an unbounded self-adjoint operator defined densely on $\{\phi \in L^2(D) : \overline{\phi} = 0\}$ whose spectrum consists of real eigenvalues only.

For an eigenpair (λ, φ) of L_γ, separation of variables shows that $\varphi(x, y) = \phi_m(x)\cos(m\pi y)$, where m, a nonnegative integer, is called the mode of φ. We denote a λ that is associated with m by λ_m. We have the following reduced eigenvalue problems for (λ_m, ϕ_m).

1. When $m = 0$,

$$-\epsilon^2\phi_0'' + f'(u_\gamma)\phi_0 - \overline{f'(u_\gamma)\phi_0} + \epsilon\gamma G_0[\phi_0] = \lambda_0\phi_0, \;\; \phi_0'(0) = \phi_0'(1) = \overline{\phi_0} = 0. \tag{3.2}$$

2. When $m \neq 0$,

$$-\epsilon^2(\phi_m'' - m^2\pi^2\phi_m) + f'(u_\gamma)\phi_m + \epsilon\gamma G_m[\phi_m] = \lambda_m\phi_m, \;\; \phi_m'(0) = \phi_m'(1) = 0. \tag{3.3}$$

Here G_m are the solution operators of the differential equations

$$-X'' = \phi_0, \;\; X'(0) = X'(1) = 0, \; \overline{X} = 0 \text{ if } m = 0, \tag{3.4}$$

$$-X'' + m^2\pi^2 X = \phi_m, \quad X'(0) = X'(1) = 0 \text{ if } m \neq 0; \tag{3.5}$$

i.e., $G_m[\phi_m] = X$. We often identify the operators G_m with the Green functions of (3.4) and (3.5).

THEOREM 3.1 (see Ren and Wei [15]). *The following three statements hold when ϵ is sufficiently small.*

1. *There exists $M(K)$ depending on K but not ϵ so that when $m \geq M(K)$, $\lambda_m \geq C\epsilon^2$ for some $C > 0$ independent of ϵ.*
2. *When $m = 0$, there are K small positive λ_0's. One of them is of order ϵ. The other $K-1$ λ_0's are of order ϵ^2. The remaining λ_0's are positive and bounded below by a positive constant independent of ϵ.*
3. *When $1 \leq m < M(K)$, there are K λ_m's of order ϵ^2, which are not necessarily positive. The remaining λ_m's are positive and bounded below by a positive constant independent of ϵ. For every $\gamma > 0$ there exist $K_0 > 0$ and $\epsilon_0 > 0$ such that for $\epsilon < \epsilon_0$ and $K > K_0$ all the eigenvalues of L are positive and the solution u, of K interfaces, is stable. For every positive integer K there exist $\gamma_0 > 0$ and $\epsilon_0 > 0$ so that for $\epsilon < \epsilon_0$ and $\gamma < \gamma_0$ all the eigenvalues of L are positive and u is stable.*

Actually, [15, Theorem 1.1] is formulated for a 3-D box. Here we have stated the simpler 2-D version.

The eigenvalues λ_0 in part 2 of Theorem 3.1 are just the 1-D eigenvalues of u_γ. The asymptotic expansions of the λ_0's are given in [15, sections 4 and 5]. One of the small λ_0's satisfies

$$\lim_{\epsilon \to 0} \epsilon^{-1}\lambda_0 > 0.$$

The other $K-1$ small λ_0's satisfy

$$\lim_{\epsilon \to 0} \epsilon^{-2}\lambda_0 > 0.$$

That these eigenvalues are positive is consistent with the fact that u_γ is a local minimizer of I_1.

Bifurcation occurs at a zero eigenvalue, and thus we focus on the λ_m's of part 3, the proof of which is in [15, sections 6 and 7]. There we obtained asymptotic expansions of the K eigenpairs (λ_m, ϕ_m) of (3.3) satisfying $\lambda_m \to 0$ as $\epsilon \to 0$. Define

$$h_j(x) = H'\left(\frac{x - x_j}{\epsilon}\right)\chi\left(\frac{x - x_j}{\sqrt{\epsilon}}\right) = H'\left(\frac{x - x_j}{\epsilon}\right) + e.s., \tag{3.6}$$

where χ is a smooth, even cut-off function

$$\chi(s) = \begin{cases} 1 & \text{if } |s| \leq 1, \\ 0 & \text{if } |s| \geq 2. \end{cases}$$

Here $e.s.$ is an exponentially small quantity with respect to ϵ because of the exponentially fast decay rate of H': $H'(t) \leq C_1 e^{-C_2|t|}$. Therefore $h_j(0) = h_j(1) = h_j'(0) = h_j'(1) = 0$, $\|h_j' - \epsilon^{-1}H''(\frac{\cdot - x_j}{\epsilon})\|_\infty = O(\epsilon^{-C/\epsilon})$, and $\|h_j'' - \epsilon^{-2}H'''(\frac{\cdot - x_j}{\epsilon})\|_\infty = O(\epsilon^{-C/\epsilon})$. When $m \geq 1$,

$$\lambda_m = \epsilon^2\left(\frac{\gamma}{\tau}\left(\Lambda - \frac{ab}{K}\right) + m^2\pi^2\right) + o(\epsilon^2), \ \ \phi_m = \sum_j c_j h_j + \epsilon^2 \phi_m^\perp. \tag{3.7}$$

As mentioned in the introduction, the estimate of λ_m in (3.7) is uniform in γ when γ varies in a compact interval of positive numbers. Here the Λ's are the K eigenvalues of the K by K matrix $[G_m(x_j^0, x_k^0)]$. $[G_m(x_j^0, x_k^0)]$ is diagonalized in [15, section 7]. When $K = 1$, it has, for each $m \geq 1$, one eigenpair (Λ, c^0):

$$\Lambda = \frac{1}{m\pi(\tanh(m\pi a) + \tanh(m\pi b))}, \ c^0 \propto 1. \tag{3.8}$$

When $K = 2$, there are two eigenpairs (Λ, c^0):

$$\Lambda = \frac{1}{m\pi(\coth(m\pi a) + \cot(m\pi b) - \operatorname{csch}(m\pi a) + \operatorname{csch}(m\pi b))}, \ \ c^0 \propto (-1, 1),$$

$$\Lambda = \frac{1}{m\pi(\coth(m\pi a) + \cot(m\pi b) - \operatorname{csch}(m\pi a) - \operatorname{csch}(m\pi b))}, \ \ c^0 \propto (1, 1). \tag{3.9}$$

When $K \geq 3$, there are K eigenpairs

$$\Lambda = \frac{1}{d - q}, \ c^0. \tag{3.10}$$

Here q is one of the K eigenvalues of the trigonal matrix

$$Q = \begin{bmatrix} \alpha & \beta & & & \\ \beta & 0 & \alpha & & \\ & \alpha & 0 & \beta & \\ & & \beta & 0 & \alpha \\ & & & \cdots & \end{bmatrix}, \tag{3.11}$$

where

$$\alpha = m\pi \operatorname{csch} \frac{2m\pi a}{K}, \ \beta = m\pi \operatorname{csch} \frac{2m\pi b}{K}, \ d = m\pi \left(\coth \frac{2m\pi a}{K} + \coth \frac{2m\pi b}{K} \right),$$

and c^0 is a corresponding eigenvector of Q.

In (3.7) ϕ_m is decomposed to $\sum_j c_j h_j$ in the subspace spanned by h_j, $j = 1, 2, \ldots, K$, and $\epsilon^2 \phi_m^\perp$ in the orthogonal complement of the subspace. Moreover, $\|\phi_m^\perp\|_2 = O(|c|)$ (see [15, Formula (6.4)]). As $\epsilon \to 0$, $c_j \to c_j^0$.

Within each mode m the K eigenvalues λ_m described in (3.7) are all different when ϵ is sufficiently small because the K eigenvalues Λ of $[G_m(x_j^0, x_k^0)]$ are all different [15, sections 5 and 7]. However, this does not imply that each λ_m is a simple eigenvalue. It is possible that there exists a different mode m' such that a $\lambda_{m'}$ happens to equal λ_m. This problem will appear when we consider the simplicity of the principal eigenvalue in the next section.

In this section we improve the estimate $\|\phi_m^\perp\|_2 = O(|c|)$ to $\|\phi_m^\perp\|_\infty = O(|c|)$ and find the limiting behavior of $\phi_m^\perp$ near each x_j. Define Π to be the solution of

$$-\Pi'' + f'(H)\Pi = \frac{\gamma}{\tau} \left(\Lambda - \frac{ab}{K} \right) H' + \text{const}, \quad \Pi \perp H' \tag{3.12}$$

in $\mathbf{R}$. Recall P from (2.6).

LEMMA 3.2.

1. $\|\phi_m^\perp\|_\infty = O(|c|)$.
2. *At each* x_j, $\phi_m^\perp(x_j+\epsilon\cdot)$ *converges in* $C^2_{loc}(\mathbf{R})$ *to* $c_j^0(P'+\Pi)$.

Proof. We define an operator L_m so that the left side of (3.3) is $L_m\phi_m$. Note that this L_m differs from the one in [15] slightly. The function $\phi_m^\perp$ satisfies the equation

$$L_m\phi_m^\perp - \lambda_m\phi_m^\perp = \sum_j c_j\left\{-m^2\pi^2 h_j - \frac{1}{\epsilon^2}(f'(u_\gamma)-f'(H))h_j - \gamma G_m\left[\frac{h_j}{\epsilon}\right] + \frac{\lambda_m}{\epsilon^2}h_j\right\}. \tag{3.13}$$

We claim that the right side of (3.13) is $O(|c|)$. The first term inside {} on the right side is obviously $O(1)$. The last term is $O(1)$ by (3.7). The third term is $O(1)$ because $G_m[\frac{h_j}{\epsilon}] \to G_m(x,x_j^0)$ as $\epsilon\to 0$. The least obvious is the second term. It is $O(1)$ by Lemma 2.3.

Suppose that part 1 of Lemma 3.2 is false. Let

$$\psi = \frac{\phi_m^\perp}{\|\phi_m^\perp\|_\infty}.$$

Then ψ satisfies

$$L_m\psi = o(1). \tag{3.14}$$

There exists $x_*\in[0,1]$ so that $|\psi(x_*)| = \max|\psi| = 1$. Without the loss of generality we assume $\psi(x_*)=1$. Otherwise just change ψ to $-\psi$. Then $x_*-x_j = O(\epsilon)$ for some x_j. If this is not the case, (3.14) cannot be satisfied at x_* since, by a maximum principle argument,

$$L_m\psi(x_*) = -\epsilon^2(\psi''(x_*)-m^2\pi^2\psi(x_*))+f'(u_\gamma(x_*))\psi(x_*)+\epsilon\gamma G_m[\psi](x_*) \ge f'(0)+o(1).$$

Here $\psi''(x_*)\le 0$ because x_* is a maximum, and because the Neumann boundary condition in case x_* is on the boundary. Define $\Psi(t) = \psi(x_j+\epsilon t)$. On the rescaled interval $(x_j/\epsilon,(1-x_j)/\epsilon)$ Ψ satisfies

$$-\Psi_{tt}+\epsilon^2m^2\pi^2\Psi+f'(u_\gamma)\Psi+\epsilon\gamma G_m[\psi](x_j+\epsilon t) = o(1),\quad \Psi'\left(-\frac{x_j}{\epsilon}\right) = \Psi'\left(\frac{1-x_j}{\epsilon}\right) = 0.$$

On every bounded subinterval of $\mathbf{R}$, because Ψ is bounded and the terms $\epsilon^2m^2\pi^2\Psi$, $f'(u_\gamma)\Psi$, and $\epsilon\gamma G_m[\psi](x_j+\epsilon t)$ are bounded, by the elliptic regularity theory Ψ is bounded in $W^{2,p}$ for any $p>1$. A bootstrapping argument shows that Ψ is bounded in $C^{2,\alpha}$ on the subinterval.

As $\epsilon\to 0$, Ψ converges in $C^2_{loc}(\mathbf{R})$, along a subsequence of ϵ, to a bounded nonzero solution Ψ_∞ of

$$-\Psi_\infty'' + f'(H)\Psi_\infty = 0.$$

There are infinitely many solutions to the last equation. However, only H' and its scalar multiples are bounded on $\mathbf{R}$. Therefore $\Psi_\infty \propto H'$. But $\psi\perp h_j$ implies that $\Psi_\infty \perp H'$. Hence $\Psi_\infty = 0$, a contradiction.

To prove part 2 we let $\Phi^\perp(t) = \phi_m^\perp(x_j+\epsilon t)$. By part 1 we can pass the limit in (3.13) to find that $\Phi^\perp\to\Phi_\infty^\perp$ in $C^2_{loc}(\mathbf{R})$, which is a solution of

$$-(\Phi_\infty^\perp)'' + f'(H)\Phi_\infty^\perp = c_j^0\left(-f''(H)H'P + \frac{\gamma}{\tau}\left(\Lambda - \frac{ab}{K}\right)H' + \text{const}\right). \tag{3.15}$$

By differentiating the equation for P we find

$$-(P')'' + f'(H)P' = -f''(H)H'P - \frac{\gamma ab}{K}.$$

Thus $\Phi_\infty^\perp$ and $c_j^0(P'+\Pi)$ satisfy the same equation (3.15). Moreover, $\phi_m^\perp \perp h_j$ implies $\Phi_\infty^\perp \perp H'$. Hence $\Phi_\infty^\perp = c_j^0(P'+\Pi)$. □

4. Bifurcation at (B B). We use γ as a bifurcation parameter. Since we are mainly interested in stable wriggled solutions, we study bifurcation from the principal eigenvalue, i.e., the smallest eigenvalue, which we denote by $\lambda(\gamma)$. We must find $\gamma_{\rm B}$ so that $\lambda(\gamma_{\rm B}) = 0$ and make sure that $\lambda(\gamma_{\rm B})$ is a simple eigenvalue. The following procedure easily finds the critical $\gamma_{\rm B}$.

1. According to [15, section 7], for each m the eigenvalues Λ of the matrix $[G_m(x_j^0, x_k^0)]$ are all different and the smallest one is

$$\Lambda_m = \frac{1}{m\pi(\tanh(m\pi a) + \tanh(m\pi b))} \quad \text{if } K = 1,$$

$$\Lambda_m = \frac{1}{m\pi(\coth(m\pi a) + \cot(m\pi b) - \operatorname{csch}(m\pi a) + \operatorname{csch}(m\pi b))} \quad \text{if } K = 2, \tag{4.1}$$

$$\Lambda_m = \frac{1}{d + \sqrt{\alpha^2 + \beta^2 + 2\alpha\beta\cos\theta}}, \quad \theta = \frac{2\pi}{K} \quad \text{if } K \geq 3.$$

   Recall that α, β, and d are defined after (3.11).

2. For each $m \geq 1$ we define γ_m^0 to be the solution of

$$\frac{\gamma_m^0}{\tau}\left(\Lambda_m - \frac{ab}{K}\right) + m^2\pi^2 = 0. \tag{4.2}$$

   According to (3.7) at $\gamma = \gamma_m^0$, $\lambda_m = o(\epsilon^2)$ where

$$\lambda_m = \epsilon^2\left(\frac{\gamma_m^0}{\tau}\left(\Lambda_m - \frac{ab}{K}\right) + m^2\pi^2\right) + o(\epsilon^2)$$

   is the smallest eigenvalue of the mode m. The solution γ_m^0 may or may not be positive. If $\gamma_m^0 \leq 0$, this means that the mode m does not yield a zero eigenvalue for any $\gamma > 0$. We discard nonpositive γ_m^0.

3. We minimize the positive γ_m^0's with respect to m. Because $\lim_{m\to\infty} \gamma_m^0 = \infty$, the minimum is achieved and is denoted by ${\gamma_{\rm B}}^0$. If m is the mode where $\gamma_m^0 = {\gamma_{\rm B}}^0$ and m' is another mode such that $\gamma_{m'}^0 > {\gamma_{\rm B}}^0$, then, since for m'

$$\Lambda_{m'} - \frac{ab}{K} < 0,$$

   a consequence of $\gamma_{m'}^0$ being positive, we find that

$$\frac{{\gamma_{\rm B}}^0}{\tau}\left(\Lambda_{m'} - \frac{ab}{K}\right) + (m')^2\pi^2 > \frac{\gamma_{m'}^0}{\tau}\left(\Lambda_{m'} - \frac{ab}{K}\right) + (m')^2\pi^2 = 0.$$

   Hence at $\gamma = {\gamma_{\rm B}}^0$ the smallest eigenvalue associated with a mode m of ${\gamma_{\rm B}}^0$ vanishes up to order ϵ^2 while the greater eigenvalues associated with the mode m and the eigenvalues associated with the modes m' satisfying $\gamma_{m'}^0 > {\gamma_{\rm B}}^0$ are positive and of order ϵ^2. Then there exists $\gamma_{\rm B}$ such that $\lim_{\epsilon\to 0}\gamma_{\rm B} = {\gamma_{\rm B}}^0$ and at $\gamma = \gamma_{\rm B}$ the principal eigenvalue $\lambda(\gamma_{\rm B})$ is zero.

The simplicity of $\lambda(\gamma_{\rm B})$ is more complex. For most a and τ, the minimum ${\gamma_{\rm B}}^0$ of γ_m^0 is achieved at one m. Then, for sufficiently small ϵ, the mode of $\lambda(\gamma_{\rm B})$ is the unique m and $\lambda(\gamma_{\rm B})$ is a simple eigenvalue.

However, it is possible, for some particular a and τ, that two modes, m and $m+1$, both minimize γ_m^0, and thus ${\gamma_{\rm B}}^0$ is achieved by two modes. This does not necessarily mean that the principal eigenvalue $\lambda(\gamma_{\rm B})$ has multiplicity two because there is the effect of the $o(\epsilon^2)$ order term in (3.7). Nevertheless we cannot exclude the possibility. Throughout the rest of the paper we impose the following condition which can be easily tested numerically for given a and τ.

HYPOTHESIS 4.1. *The positive γ_m^0's defined in* (4.2) *are minimized at a unique m.*

Being a simple eigenvalue now, $\lambda(\gamma_{\rm B})$ is continued smoothly to a curve of simple eigenvalues $\lambda(\gamma)$ of L_γ as γ varies. $\lambda(\gamma)$, which is estimated in (3.7), is valid uniformly in γ when γ varies in a neighborhood (such as $[\gamma_{\rm B}/2, 2\gamma_{\rm B}]$) of $\gamma_{\rm B}$. Denote the eigenfunction associated with $\lambda(\gamma_{\rm B})$ by $\varphi_{\rm B}(x,y) = \phi_{\rm B}(x)\cos(m\pi y)$, where m from now on is the unique mode in Hypothesis 4.1. We write $u_{\rm B} := u_{\gamma_{\rm B}}$ and $L_{\rm B} := L_{\gamma_{\rm B}}$ for simplicity. Let

$$(4.3)\quad X := \{w \in W^{2,2}(D) :\ \partial_\nu w = 0 \text{ on } \partial D,\ \overline{w} = 0\},\quad Y := \{z \in L^2(D) :\ \overline{z} = 0\}.$$

Here X is a dense subspace of Y equipped with the $W^{2,2}$ norm; Y is a Hilbert space with the usual inner product $\langle \cdot, \cdot \rangle$ inherited from $L^2(D)$.

A nonlinear map $F : (0,\infty) \times X \to Y$ is defined by

$$(4.4)\ F(\gamma, w) := -\epsilon^2\Delta(u_\gamma + w) + f(u_\gamma + w) - \overline{f(u_\gamma + w)} + \epsilon\gamma(-\Delta)^{-1}(u_\gamma + w - a).$$

Obviously the trivial branch $(\gamma, 0)$ is a solution branch of $F(\gamma, w) = 0$. It corresponds to the K-interface, perfect lamellar solution u_γ of (1.2), parameterized by γ. We look for another solution branch, a bifurcating branch, $(\gamma(s), w(s))$ of F. It gives another solution $u_{\gamma(s)} + w(s)$ of (1.2).

THEOREM 4.2. *Under Hypothesis* 4.1*, when ϵ is sufficiently small, at $\gamma = \gamma_{\rm B}$ another solution branch $(\gamma(s), w(s))$ bifurcates from the trivial branch $(\gamma, 0)$. Here $w(s) = s\varphi_{\rm B} + sg(s)$, where the parameter s is in a neighborhood of* 0 *with $\gamma(0) = \gamma_{\rm B}$ and $w(0) = 0$. Moreover, $\gamma : s \to \gamma(s) \in \mathbf{R}$ and $g : s \to g(s) \in X$ are both continuously differentiable, g satisfies $g(s) \perp \varphi_{\rm B}$, and $g(0) = 0$.*

Note that $u_{\gamma(s)} + w(s)$ is approximately $u_{\gamma(s)}(x) + s\phi_{\rm B}(x)\cos(m\pi y)$ since $g(s)$ is a smaller term compared to $\varphi_{\rm B}(x,y) = \phi_{\rm B}(x)\cos(m\pi y)$. Figure 2(b) is made based on this observation.

Proof. We appeal to the standard bifurcation from the simple eigenvalue theorem (see [26, Theorem 13.5, page 173]). Denote the Fréchet derivatives of F with respect to γ by D_1 and with respect to w by D_2. We need to verify the following three properties.

1. $D_2F(\gamma_{\rm B}, 0)$, which is just $L_{\rm B} : X \to Y$, has a 1-D kernel spanned by $\varphi_{\rm B}$.
2. $\mathcal{R}(D_2F(\gamma_{\rm B}, 0))$, the range of $D_2F(\gamma_{\rm B}, 0)$, has codimension 1.
3. $D_1D_2F(\gamma_{\rm B}, 0)\varphi_{\rm B}$ is not in $\mathcal{R}(D_2F(\gamma_{\rm B}, 0))$.

Property 1 follows from the simplicity of $\lambda(\gamma)$. To prove 2 we claim that there exists a positive constant $c(\epsilon, \gamma_{\rm B})$ depending on ϵ and $\gamma_{\rm B}$ so that

$$(4.5)\qquad \|\psi\|_2 \le c(\epsilon, \gamma_{\rm B})\|L_{\rm B}\psi\|_2 \text{ for all } \psi \perp \varphi_{\rm B},\ \psi \in X.$$

Suppose (4.5) is false. There would exist a sequence $\psi_n \in X$, $\psi_n \perp \varphi_{\mathrm{B}}$, $\|\psi_n\|_2 = 1$ so that $\|L_{\mathrm{B}}\psi_n\|_2 \to 0$. Let $\psi_n \to \psi_*$ weakly in $L^2(D)$. Then $\psi_* \perp \varphi_{\mathrm{B}}$. For every $\omega \in X$,

$$\langle \psi_*, L_{\mathrm{B}}\omega\rangle = \lim_{n\to\infty} \langle \psi_n, L_{\mathrm{B}}\omega\rangle = \lim_{n\to\infty}\langle L_{\mathrm{B}}\psi_n, \omega\rangle = 0.$$

By the self-adjointness of L_{B}, $\psi_* \in X$ and $L_{\mathrm{B}}\psi_* = 0$. Hence $\psi_* = 0$ from property 1. Rewrite $L_{\mathrm{B}}\psi_n$ as

$$-\epsilon^2\Delta\psi_n = -f'(u_{\mathrm{B}})\psi_n + \overline{f'(u_{\mathrm{B}})\psi_n} - \epsilon\gamma_{\mathrm{B}}(-\Delta)^{-1}\psi_n + L_{\mathrm{B}}\psi_n.$$

Since $\|\psi_n\|_2 = 1$ and $\|L_{\mathrm{B}}\psi_n\|_2 \to 0$, the right side is bounded in $L^2(D)$. The elliptic regularity theory asserts that ψ_n is precompact in $L^2(D)$. Hence $\psi_n \to 0$ in $L^2(D)$. This is inconsistent with the fact $\|\psi_n\|_2 = 1$. Hence (4.5) holds.

We now prove 2. by showing $\mathcal{R}(L_{\mathrm{B}}) = \{\varphi_{\mathrm{B}}\}^\perp$. The self-adjointness of L_{B} and 1. imply that every $\psi \in \mathcal{R}(L_{\mathrm{B}})^\perp$ is φ_{B} multiplied by a constant. It suffices to show that $\mathcal{R}(L_{\mathrm{B}})$ is closed. Take $\omega_n \in \mathcal{R}(L_{\mathrm{B}})$ so that $\omega_n \to \omega_*$ in $L^2(D)$. Let $\psi_n \in X$, $\psi_n \perp \varphi_{\mathrm{B}}$ such that $L_{\mathrm{B}}\psi_n = \omega_n$. Since ω_n is a Cauchy sequence, by (4.5) ψ_n is also a Cauchy sequence. Let $\psi_n \to \psi_*$ in $L^2(D)$. Note that L_{B} is a closed operator since it is self-adjoint. Hence $\psi_* \in X$ and $L_{\mathrm{B}}\psi_* = \omega_*$. This proves property 2.

To prove 3, note that the linear map $D_1D_2F(\gamma_{\mathrm{B}}, 0) : X \to Y$ is

$$\psi \to f''(u_{\mathrm{B}})\frac{du_\gamma}{d\gamma}|_{\gamma=\gamma_{\mathrm{B}}}\psi - \overline{f''(u_{\mathrm{B}})\frac{du_\gamma}{d\gamma}|_{\gamma=\gamma_{\mathrm{B}}}\psi} + \epsilon(-\Delta)^{-1}\psi. \tag{4.6}$$

Since $\mathcal{R}(L_{\mathrm{B}}) = \{\varphi_{\mathrm{B}}\}^\perp$, it suffices to show

$$\langle D_1D_2F(\gamma_{\mathrm{B}},0)\varphi_{\mathrm{B}}, \varphi_{\mathrm{B}}\rangle \neq 0, \text{ i.e., } \int_D \left\{ f''(u_{\mathrm{B}})\frac{du_\gamma}{d\gamma}|_{\gamma=\gamma_{\mathrm{B}}}\varphi_{\mathrm{B}}^2 + \epsilon\varphi_{\mathrm{B}}(-\Delta)^{-1}\varphi_{\mathrm{B}}\right\} \neq 0. \tag{4.7}$$

This fact is established in the next lemma. ∎

Lemma 4.3. *When ϵ is sufficiently small,*

$$\int_D \left\{ f''(u_{\mathrm{B}})\frac{du_\gamma}{d\gamma}|_{\gamma=\gamma_{\mathrm{B}}}\varphi_{\mathrm{B}}^2 + \epsilon\varphi_{\mathrm{B}}(-\Delta)^{-1}\varphi_{\mathrm{B}}\right\} = -\frac{\epsilon^3|c^0|^2\tau m^2\pi^2}{2\gamma_{\mathrm{B}}} + o(\epsilon^3|c^0|^2) < 0.$$

Here τ is given in (2.2), *and c^0 is in* (3.8)–(3.10), *a nonzero vector.*

Proof. Note that $\varphi_{\mathrm{B}}(x,y) = \phi_{\mathrm{B}}(x)\cos(m\pi y)$ and $(-\Delta)^{-1}\varphi_{\mathrm{B}}(x,y) = G_m[\phi_{\mathrm{B}}](x)\cos(m\pi y)$. Hence after integrating out the y variable we deduce

$$\begin{aligned}&\int_D \left\{ f''(u_{\mathrm{B}})\frac{du_\gamma}{d\gamma}|_{\gamma=\gamma_{\mathrm{B}}}\varphi_{\mathrm{B}}^2 + \epsilon\varphi_{\mathrm{B}}(-\Delta)^{-1}\varphi_{\mathrm{B}}\right\}\\ &= \int_0^1 \left\{ \frac{1}{2}f''(u_{\mathrm{B}})\frac{du_\gamma}{d\gamma}|_{\gamma=\gamma_{\mathrm{B}}}\phi_{\mathrm{B}}^2 + \frac{\epsilon}{2}G_m[\phi_{\mathrm{B}}]\phi_{\mathrm{B}}\right\} dx.\end{aligned} \tag{4.8}$$

By Lemmas 2.4 and 3.2, we find

$$\begin{aligned}\int_0^1 f''(u_{\mathrm{B}})\frac{du_\gamma}{d\gamma}|_{\gamma=\gamma_{\mathrm{B}}}\phi_{\mathrm{B}}^2 &= \int_0^1 f''(u_{\mathrm{B}})\left(\sum_j (ch_j + \epsilon^2 R)\right)\phi_{\mathrm{B}}^2 + o(\epsilon^3)\\ &= \int_0^1 f''(u_{\mathrm{B}})\left(\sum_j (ch_j + \epsilon^2(\gamma^{-1}P + \tilde{R})\right)\phi_{\mathrm{B}}^2 + o(\epsilon^3|c|^2)\\ &= \epsilon^2\int_0^1 f''(u_{\mathrm{B}})(\gamma^{-1}P)\phi_{\mathrm{B}}^2 + o(\epsilon^3|c|^2).\end{aligned}$$

We have used the fact that P is odd, and $\tilde{R}$ and H_t are even. Hence we arrive at

$$\int_0^1 f''(u_{\mathrm{B}})\frac{du_\gamma}{d\gamma}|_{\gamma=\gamma_{\mathrm{B}}}\phi_{\mathrm{B}}^2 = \epsilon^3\int_{\mathbf{R}} f''(H)\gamma^{-1}P\left(\sum_j c_j^2\right)(H')^2\,dt + o(\epsilon^3|c|^2)$$
$$= -\frac{\epsilon^3|c^0|^2ab}{K} + o(\epsilon^3|c^0|^2),$$

where the last equation follows after we differentiate the equation for $\gamma^{-1}P$,

$$-(\gamma^{-1}P)''' + f'(H)(\gamma^{-1}P)' + f''(H)H'(\gamma^{-1}P) = -\frac{ab}{K},$$

multiply by H', and integrate: $\int_{\mathbf{R}} f''(H)\gamma^{-1}P(H')^2\,dt = -ab/K$. By Lemma 3.2 we obtain

$$\int_0^1 \epsilon G_m[\phi_{\mathrm{B}}]\phi_{\mathrm{B}} = \epsilon^3\int_0^1\left(\sum_j c_j G_m\left[\frac{h_j}{\epsilon}\right]\right)\left(\sum_k c_k\frac{h_k}{\epsilon}\right) + o(\epsilon^3|c|^2)$$
$$= \epsilon^3\sum_{j,k} c_j c_k G_m(x_k, x_j) + o(\epsilon^3|c|^2) \;=\; \epsilon^3\Lambda|c^0|^2 + o(\epsilon^3|c^0|^2).$$

Here Λ is an eigenvalue of the K by K matrix $G_m(x_k^0, x_j^0)$ and c^0 is a corresponding eigenvector, satisfying $\lim_{\epsilon\to 0} c_j = c_j^0$.

Hence the right side of (4.8) becomes

$$\frac{\epsilon^3|c^0|^2}{2}\left(\Lambda - \frac{ab}{K}\right) + o(\epsilon^3|c^0|^2).$$

To determine the sign of this quantity we recall (3.7):

$$\lambda(\gamma_{\mathrm{B}}) = \epsilon^2\left[\frac{\gamma_{\mathrm{B}}}{\tau}\left(\Lambda - \frac{ab}{K}\right) + m^2\pi^2\right] + o(\epsilon^2).$$

But here $\lambda(\gamma_{\mathrm{B}}) = 0$. Hence

$$\Lambda - \frac{ab}{K} = -\frac{\tau m^2\pi^2}{\gamma_{\mathrm{B}}} + o(1).$$

This proves the lemma. ∎

Remark 4.4. The proof of Theorem 4.2 does not use the fact that $\lambda(\gamma_{\mathrm{B}})$ is a principal eigenvalue. The theorem continues to hold if $\lambda(\gamma_{\mathrm{B}})$ is just a zero, simple eigenvalue, not necessarily principal. However, the bifurcating solution from a non-principal eigenvalue is unstable, and hence less interesting to us.

5. Stability of the bifurcating solutions. The eigenvalue $\lambda(\gamma)$ of the trivial branch u_γ corresponds to an eigenvalue $\lambda_*(s)$ of the bifurcating solution $u_{\gamma(s)} + w(s)$. The sign of $\lambda_*(s)$ may be determined from the shape of $\gamma(s)$. Thus we proceed to compute $\gamma'(0)$ and $\gamma''(0)$.

Place $w(s) = s\varphi_{\mathrm{B}} + sg(s)$ into $F(\gamma, w) = 0$ and divide by s:

$$-\epsilon^2\Delta\left(\frac{u_{\gamma(s)}}{s} + \varphi_{\mathrm{B}} + g(s)\right) + \frac{f(u_{\gamma(s)} + w(s))}{s}$$
$$+\,\epsilon\gamma(s)(-\Delta)^{-1}\left(\frac{u_{\gamma(s)} - a}{s} + \varphi_{\mathrm{B}} + g(s)\right) = \text{const}, \tag{5.1}$$

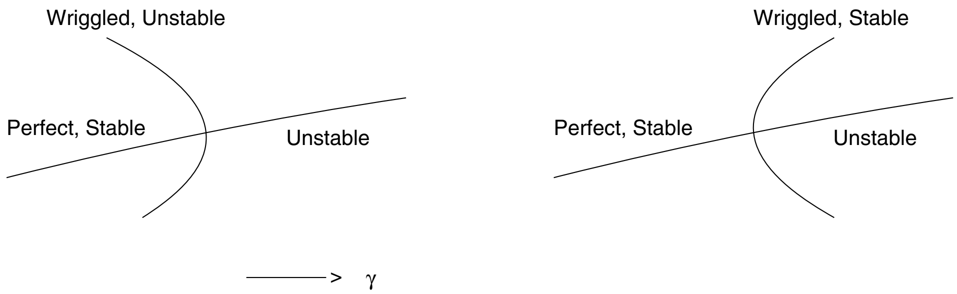


FIG. 3. *The two possible diagrams of wriggled lamellar solutions bifurcating out of perfect lamellar solutions. The bifurcating solutions are unstable in the first case (left) where $\gamma''(0) < 0$, and stable in the second case (right) where $\gamma''(0) > 0$.*

where const refers to the term coming from the average of f, which is independent of (x, y). Here we do not need its exact value. On the other hand, divide (1.2) of $u_{\gamma(s)}$ by s and subtract the result from (5.1):

(5.2)
$$-\epsilon^2\Delta(\varphi_{\rm B} + g(s)) + \frac{f(u_{\gamma(s)} + w(s)) - f(u_{\gamma(s)})}{s} + \epsilon\gamma(s)(-\Delta)^{-1}(\varphi_{\rm B} + g(s)) = \text{const.}$$

Differentiate (5.2) with respect to s and set $s = 0$ afterwards:

$$L_{\rm B}g'(0) + \gamma'(0)\left\{f''(u_{\rm B})\frac{du_\gamma}{d\gamma}|_{\gamma=\gamma_{\rm B}}\varphi_{\rm B} + \epsilon(-\Delta)^{-1}\varphi_{\rm B}\right\} + \frac{1}{2}f''(u_{\rm B})\varphi_{\rm B}^2 = \text{const.} \tag{5.3}$$

Then we multiply (5.3) by $\varphi_{\rm B}$ and integrate over D:

$$\gamma'(0)\int_D \left\{f''(u_{\rm B})\frac{du_\gamma}{d\gamma}|_{\gamma=\gamma_{\rm B}}\varphi_{\rm B}^2 + \epsilon\varphi_{\rm B}(-\Delta)^{-1}\varphi_{\rm B}\right\} = -\int_D \frac{1}{2}f''(u_{\rm B})\varphi_{\rm B}^3. \tag{5.4}$$

Clearly the right side of (5.4) is 0 since $\varphi_{\rm B}(x, y) = \phi_{\rm B}(x)\cos(m\pi y)$ and integration with respect to y yields 0. Lemma 4.3 then implies the following.

COROLLARY 5.1. $\gamma'(0) = 0$.

Consequently, (5.3) is simplified to

$$L_{\rm B}g'(0) = -\frac{1}{2}f''(u_{\rm B})\varphi_{\rm B}^2 + \text{const}, \quad g'(0) \perp \varphi_{\rm B}. \tag{5.5}$$

The right side of (5.5) is perpendicular to $\varphi_{\rm B}$ since the integration of the right side multiplied by $u_{\rm B}$ with respect to y yields 0, so there is a solution of $g'(0)$. $g'(0) \perp \varphi_{\rm B}$ follows from $g(s) \perp \varphi_{\rm B}$ in Theorem 4.2, so $g'(0)$ is uniquely determined.

Corollary 5.1 implies that the bifurcation diagram has the shape of a pitchfork. There are two possibilities illustrated in Figure 3. To determine which of the two cases occurs, we need to find $\gamma''(0)$. Differentiate (5.2) with respect to s twice and set $s = 0$ afterwards:

$$\begin{aligned} &L_{\rm B}g''(0) + \gamma''(0)\left\{f''(u_{\rm B})\frac{du_\gamma}{d\gamma}|_{\gamma=\gamma_{\rm B}}\varphi_{\rm B}^2 + \epsilon(-\Delta)^{-1}\varphi_{\rm B}\right\} \\ &+2f''(u_{\rm B})\varphi_{\rm B}g'(0) + \frac{1}{3}f'''(u_{\rm B})\varphi_{\rm B}^3 = \text{const.} \end{aligned} \tag{5.6}$$

We have used Corollary 5.1 in deriving (5.6). Again we multiply (5.6) by φ_{B} and integrate:

$$\begin{aligned}&\gamma''(0)\int_D\left\{f''(u_{\mathrm{B}})\frac{du_\gamma}{d\gamma}|_{\gamma=\gamma_{\mathrm{B}}}\varphi_{\mathrm{B}}^2+\epsilon\varphi_{\mathrm{B}}(-\Delta)^{-1}\varphi_{\mathrm{B}}\right\}\\ &\quad=-\int_D\left\{2f''(u_{\mathrm{B}})\varphi_{\mathrm{B}}^2g'(0)+\frac{1}{3}f'''(u_{\mathrm{B}})\varphi_{\mathrm{B}}^4\right\}.\end{aligned}\tag{5.7}$$

The integral on the left side of (5.7) has been calculated in Lemma 4.3. We now need to know the right side.

Lemma 5.2.

$$\begin{aligned}&-\int_D\left\{2f''(u_{\mathrm{B}})\varphi_{\mathrm{B}}^2g'(0)+\frac{1}{3}f'''(u_{\mathrm{B}})\varphi_{\mathrm{B}}^4\right\}\\ &=-\epsilon^5 m\pi\gamma_{\mathrm{B}}\sum_{j=1}^K c_j^4\left[\frac{2+\cosh(2m\pi)}{8\sinh(2m\pi)}+\frac{\cosh(2m\pi(1-2x_j^0))}{8\sinh(2m\pi)}\right.\\ &\qquad\left.+\frac{\cosh(m\pi(1-2x_j^0))}{4\sinh(m\pi)}-\frac{3(m\pi)^3\tau}{8\gamma_{\mathrm{B}}}\right]+o(\epsilon^5|c|^4).\end{aligned}$$

The proof of Lemma 5.2 is formidable. We have to expand the quantity to the ϵ^5 order term, because all the lower order terms up to ϵ^4 vanish. Our main idea is to expand $u_{\mathrm{B}}, \phi_{\mathrm{B}}, 2g'(0)$ as $(\cdots)+\epsilon^2(\cdots)$ near each interface x_j and then show that the quantity in Lemma 5.2 depends "locally" on these expansions near x_j. This is a very long computation; we do not know if there is a simpler proof. We include the computation in Appendices B and C. The reader may skip it in a first reading. Combining Lemmas 4.3 and 5.2 we obtain the following.

Corollary 5.3. *As* $\epsilon\to 0$, $\epsilon^{-2}\gamma''(0)\to$

$$\begin{aligned}&\frac{2({\gamma_{\mathrm{B}}}^0)^2}{|c^0|^2 m\pi\tau}\sum_{j=1}^K(c_j^0)^4\left[\frac{2+\cosh(2m\pi)}{8\sinh(2m\pi)}+\frac{\cosh(2m\pi(1-2x_j^0))}{8\sinh(2m\pi)}\right.\\ &\qquad\left.+\frac{\cosh(m\pi(1-2x_j^0))}{4\sinh(m\pi)}-\frac{3(m\pi)^3\tau}{8{\gamma_{\mathrm{B}}}^0}\right],\end{aligned}$$

where ${\gamma_{\mathrm{B}}}^0=\lim_{\epsilon\to0}\gamma_{\mathrm{B}}$ *is given in section* 4.

Define

$$\begin{aligned}S(a,K):=&\sum_{j=1}^K\left(\frac{c_j^0}{|c^0|}\right)^4\left[\frac{2+\cosh(2m\pi)}{8\sinh(2m\pi)}+\frac{\cosh(2m\pi(1-2x_j^0))}{8\sinh(2m\pi)}\right.\\ &\qquad\left.+\frac{\cosh(m\pi(1-2x_j^0))}{4\sinh(m\pi)}-\frac{3(m\pi)^3\tau}{8{\gamma_{\mathrm{B}}}^0}\right].\end{aligned}\tag{5.8}$$

Note that $S(a,K)$ depends on a and K only. It does not depend on τ. Since τ depends on the shape of W, $S(a,K)$ is independent of the exact shape of W. Then Corollary 5.3 implies

$$\lim_{\epsilon\to0}\epsilon^{-2}\gamma''(0)=\frac{2({\gamma_{\mathrm{B}}}^0)^2|c^0|^2}{m\pi\tau}S(a,K).\tag{5.9}$$

THEOREM 5.4. *Under Hypothesis* 4.1 *when* ϵ *is sufficiently small, the bifurcating solution* $u_{\gamma(s)} + w(s)$ *of* K *wriggled interfaces is stable if* $S(a, K) > 0$ *and unstable if* $S(a, K) < 0$.

Proof. We first find $\lambda'(\gamma_{\mathrm{B}})$. Differentiate the equation $L_\gamma \varphi = \lambda\varphi$ with respect to γ:

$$-\epsilon^2\Delta\varphi_\gamma + f'(u_\gamma)\varphi_\gamma + \epsilon\gamma(-\Delta)^{-1}\varphi_\gamma + f''(u_\gamma)\frac{du_\gamma}{d\gamma}\varphi_\gamma + \epsilon(-\Delta)^{-1}\varphi = \lambda\varphi_\gamma + \lambda'(\gamma)\varphi + \text{const.}$$

Set $\gamma = \gamma_{\mathrm{B}}$ in the equation, multiply the equation by φ_{B}, and integrate over D.

$$\int_D \left\{ f''(u_{\mathrm{B}})\frac{du_\gamma}{d\gamma}|_{\gamma=\gamma_{\mathrm{B}}}\varphi_{\mathrm{B}}^2 + \epsilon\varphi_{\mathrm{B}}(-\Delta)^{-1}\varphi_{\mathrm{B}} \right\} = \lambda'(\gamma_{\mathrm{B}}) \int_D \varphi_{\mathrm{B}}^2.$$

The left side is calculated in Lemma 4.3. The integral on the right side is

$$\int_D \varphi_{\mathrm{B}}^2 = \int_D \left(\sum_j c_j h_j \right)^2 \cos^2(m\pi y)\, dxdy + o(\epsilon|c|^2) = \frac{\epsilon\tau}{2}\sum_j c_j^2 + o(\epsilon|c|^2).$$

Therefore

$$\lambda'(\gamma_{\mathrm{B}}) = -\frac{\epsilon^2 m^2\pi^2}{\gamma_{\mathrm{B}}} + o(\epsilon^2) < 0. \tag{5.10}$$

According to Crandall and Rabinowitz [4, Theorem 1.16], who generalize an earlier result of Sattinger [23], near $s = 0$, $\lambda_*(s)$ and $-s\gamma'(s)\lambda'(\gamma_{\mathrm{B}})$ have the same zeros, and

$$\lim_{s\to 0,\ \lambda_*(s)\neq 0} \frac{-s\gamma'(s)\lambda'(\gamma_{\mathrm{B}})}{\lambda_*(s)} = 1. \tag{5.11}$$

Here $\lambda_*(s)$ is the principal eigenvalue of the bifurcating solution $u_{\gamma(s)}+w(s)$. Whether the bifurcating solution is stable depends on whether $\lambda_*(s)$ is positive. The theorem follows from (5.9), (5.10), and (5.11). □

Let us use Theorem 5.4 to work out some examples. The quantity $S(a, K)$ may be accurately calculated following these numerical steps.

1. Follow the procedure in section 4 to find ${\gamma_{\mathrm{B}}}^0/\tau$ and m.
2. Make certain that Hypothesis 4.1 is satisfied.
3. Find c^0 from Q corresponding to Λ of (4.1).
4. Find $S(a, K)$ from (5.8).

Tables 1, 2, and 3 report our numerical calculations based on this method for the cases $a = 1/2$, $1/8$, and $7/8$. In each table the first column is the number of the interfaces in the perfect lamellar solution u_{B}. The second column gives the value of m associated with the principal eigenvalue 0 of u_{B}. Note that m does not increase as fast as K does. The third column has the value of ${\gamma_{\mathrm{B}}}^0/\tau$. We will explain the fourth in a moment. The fifth column has the value of $S(a, K)$. The last column indicates the stability of the bifurcating solution with K wriggled interfaces.

We have deliberately chosen $a = 1/8$ and $a = 7/8$ because they are somehow "symmetric." With the exception of $K = 2$, the ${\gamma_{\mathrm{B}}}^0/\tau$'s are identical in Tables 2 and 3 for the same value of K. Moreover, the $S(a, K)$ values are the same in the two tables when K is odd. All these symmetries and asymmetries can be explained

TABLE 1

The stability of the wriggled lamellar solutions that bifurcate from the principal eigenvalues of the perfect lamellar solutions, when $a = 1/2$.

K	m	${\gamma_{\rm B}}^0/\tau$	K_{opt}	$S(1/2, K)$	Stability
1	1	1.2906e+02	2	5.7961e-02	Stable
2	2	8.6349e+02	3	1.4167e-02	Stable
3	3	2.7193e+03	4	5.4073e-03	Stable
4	3	5.3823e+03	5	2.5346e-02	Stable
5	3	9.7086e+03	6	2.8045e-02	Stable
6	4	1.6165e+04	7	1.9801e-02	Stable
7	4	2.4091e+04	8	2.0216e-02	Stable
8	4	3.4492e+04	9	1.9435e-02	Stable
9	4	4.7728e+04	10	1.8273e-02	Stable
10	4	6.4156e+04	11	1.7045e-02	Stable

TABLE 2

The stability of the wriggled lamellar solutions that bifurcate from the principal eigenvalues of the perfect lamellar solutions, when $a = 1/8$.

K	m	${\gamma_{\rm B}}^0/\tau$	K_{opt}	$S(1/8, K)$	Stability
1	3	1.7317e+03	2	-3.1473e-02	Unstable
2	5	1.3418e+04	4	2.7118e-02	Stable
3	2	1.0218e+04	3	6.0136e-02	Stable
4	3	2.3798e+04	5	5.5916e-02	Stable
5	3	4.3553e+04	6	3.5615e-02	Stable
6	3	7.3607e+04	7	3.0307e-02	Stable
7	4	1.1373e+05	8	2.5439e-02	Stable
8	4	1.6489e+05	9	2.2622e-02	Stable
9	4	2.3061e+05	10	2.0337e-02	Stable
10	4	3.1284e+05	11	1.8426e-02	Stable

TABLE 3

The stability of the wriggled lamellar solutions that bifurcate from the principal eigenvalues of the perfect lamellar solutions, when $a = 7/8$.

K	m	${\gamma_{\rm B}}^0/\tau$	K_{opt}	$S(7/8, K)$	Stability
1	3	1.7317e+03	2	-3.1473e-02	Unstable
2	2	3.4949e+03	2	4.9812e-02	Stable
3	2	1.0218e+04	3	6.0136e-02	Stable
4	3	2.3798e+04	5	2.8615e-02	Stable
5	3	4.3553e+04	6	3.5615e-02	Stable
6	3	7.3607e+04	7	3.0508e-02	Stable
7	4	1.1373e+05	8	2.5439e-02	Stable
8	4	1.6489e+05	9	2.2655e-02	Stable
9	4	2.3061e+05	10	2.0337e-02	Stable
10	4	3.1284e+05	11	1.8433e-02	Stable

from (4.1) for Λ and the matrix (3.11) of Q. In summary, our problem is invariant under the reflection $u(x) \to u(1-x)$ with respect to $x = 1/2$, and the exchange $u(x) \to 1 - u(x)$ of A and B monomers. When K is odd the perfect lamellar solution u_γ studied in Table 2 becomes u_γ in Table 3, and vice versa, after we perform the last two transformations. However, this is not true when K is even.

There is an interesting relationship between the perfect lamellar solution $u_{\rm B}$ whose principal eigenvalue is 0 and the 1-D global minimizer. In [15, section 8] it is shown that the 1-D global minimizer (the global minimizer of I_1 in Theorem 2.1, also a perfect lamellar solution on D after trivial extension), which is one of the 1-D local minimizers, has the number of interfaces K_{opt} which minimizes (among positive integers N) $\tau N + \gamma a^2 b^2/(6N^2)$. If we take $\gamma = \gamma_{\rm B}$ so that the K-interface, perfect lamellar solution $u_{\rm B}$ has 0 principal eigenvalue, we find the 1-D global minimizer corresponding to $\gamma_{\rm B}$. The number of interfaces K_{opt} of this 1-D global minimizer is reported in the fourth columns in Tables 1, 2, and 3.

It is known [15] that for any γ the 1-D global minimizer sits near the 2-D stability borderline. But it has not been determined whether the 1-D global minimizer is stable in 2-D. Now we show that the 1-D global minimizer is stable in two dimensions for some γ and unstable for other γ.

First consider $\gamma = \gamma_{\rm B}$. Then $u_{\rm B}$ is the borderline of 2-D stability. By examining the formulae (3.7) and (4.1) we find that for each γ the principal eigenvalue increases as K increases. If $\gamma = \gamma_{\rm B}$ and K is the number of the interfaces of $u_{\rm B}$ whose principal eigenvalue is zero, then Tables 1–3 show that in most cases $K_{opt} = K + 1$, and hence the 1-D global minimizer is stable.

However, in some cases, such as the third row of Table 2, $K = K_{opt}$. Then $u_{\rm B}$ is the 1-D global minimizer. Since the principal eigenvalue of $u_{\rm B}$ is zero, the linear stability of $u_{\rm B}$ is undetermined. However, if we consider γ in a small neighborhood of $\gamma_{\rm B}$, then u_γ continues to be the 1-D global minimizer. But Figure 3 shows that when γ is slightly greater than $\gamma_{\rm B}$ the 1-D global minimizer is unstable in two dimensions.

Appendix A. Proof of Lemma 2.4.

We differentiate the 1-D Euler–Lagrange equation in Theorem 2.1 with respect to γ to deduce

$$-\epsilon^2 \Delta \left(\frac{du_\gamma}{d\gamma} \right) + f'(u_\gamma) \frac{du_\gamma}{d\gamma} - \eta + \epsilon \gamma G_0 \left[\frac{du_\gamma}{d\gamma} \right] + \epsilon v = 0, \tag{A.1}$$

where

$$\eta = \overline{f'(u_\gamma) \frac{du_\gamma}{d\gamma}}. \tag{A.2}$$

We first show that $du_\gamma/d\gamma = O(1)$ so that c defined in (2.8) is of order $O(1)$. By the periodicity of u_γ and $du_\gamma/d\gamma$ we consider (A.1) in one half period $(0, l)$ where $l = 1/K$. Decompose

$$\frac{du_\gamma}{d\gamma} = \beta \phi_0 + \phi_0^\perp,\ \phi_0 \perp \phi_0^\perp, \tag{A.3}$$

where ϕ_0 is the eigenfunction associated with the ϵ order eigenvalue λ_0 described in part 2 of Theorem 3.1. Note that λ_0 is the only small eigenvalue of mode zero now. In [15, section 4] it was shown that

$$\phi_0 = h_1 - \overline{h_1} + O(\epsilon). \tag{A.4}$$

Denote the operator on the left side of (A.1) by L_0. Then

$$L_0 \phi_0^\perp = -\beta L_0 \phi_0 - \epsilon v = -\beta \lambda_0 - \epsilon v = O(\epsilon |\beta|) + O(\epsilon).$$

This equation and the fact that $\phi_0 \perp \phi_0^\perp$ imply, as in the proof of Lemma 3.2, that

$$\phi_0^\perp = O(\epsilon |\beta|) + O(\epsilon). \tag{A.5}$$

We rewrite (A.1) as $\beta L_0\phi_0 + L_0\phi_0^\perp = -\epsilon v$, multiply by ϕ_0, and integrate to deduce

$$\beta\lambda_0\|\phi_0\|_2^2 = -\epsilon\int_0^l v\phi_0\,dx.$$

It follows from (A.4) that

$$\beta = O(1), \tag{A.6}$$

and (A.5) in turn becomes

$$\phi_0^\perp = O(\epsilon). \tag{A.7}$$

Therefore $du_\gamma/d\gamma = O(1)$. We can assume that along a subsequence of ϵ, $c \to c_0$. After we prove Lemma 2.4 we will find the value of c_0. Then $c \to c_0$ along all $\epsilon \to 0$.

We now return to the whole interval $(0,1)$ as we prove Lemma 2.4. We first construct a preliminary approximation $p = c\sum_j h_j$. The difference

$$\frac{du_\gamma}{d\gamma} - p$$

satisfies

$$-\epsilon^2\Delta\left(\frac{du_\gamma}{d\gamma} - p\right) + f'(u_\gamma)\left(\frac{du_\gamma}{d\gamma} - p\right) = \eta + O(\epsilon). \tag{A.8}$$

Together with the fact that

$$\frac{du_\gamma}{d\gamma}(x_j) = p(x_j),$$

we deduce that

$$\frac{du_\gamma}{d\gamma} - p = O(\epsilon) + O(|\eta|). \tag{A.9}$$

If we multiply (A.8) by a h_i and integrate, then

$$\eta = O\left(\epsilon^2\|\frac{du_\gamma}{d\gamma} - p\|_\infty\right) + O(\epsilon). \tag{A.10}$$

We conclude from (A.9) and (A.10) that

$$\frac{du_\gamma}{d\gamma} - p = O(\epsilon),\ \eta = O(\epsilon), \tag{A.11}$$

and consequently (recall $r = G_0[du_\gamma/d\gamma]$)

$$r = O(\epsilon). \tag{A.12}$$

We are now ready to consider the approximation q of $du_\gamma/d\gamma$. In the inner and matching regions,

$$\begin{aligned}
&-\epsilon^2\Delta q_j + f'(u_\gamma)q_j\\
&= -\epsilon^2\Delta(cH_j' + \epsilon^2R_j) + f'(u_\gamma)(cH_j' + \epsilon^2R_j)\\
&= c(f'(u_\gamma) - f'(H))H_j' - \epsilon^2(c_0PH_j'f''(H) + tv'(x_j) - \zeta) + \epsilon^2(f'(u_\gamma) - f'(H))R_j\\
&= \epsilon^2(c - c_0)f''(H)PH_j' - \epsilon^2(tv'(x_j) - \zeta) + \epsilon^2(f'(u_\gamma) - f'(H))R_j\\
&= -\epsilon^2(tv'(x_j) - \zeta) + o(\epsilon^2) + \epsilon^2O(\epsilon|t|)\\
&= \epsilon^2(\zeta - tv'(x_j)) + o(\epsilon^2).
\end{aligned}$$

Therefore, with the help of $\alpha \in (1/2, 1)$,

$$-\epsilon^2 \Delta q_j + f'(u_\gamma) q_j - \eta + \epsilon\gamma r + \epsilon v = -\eta + \epsilon\gamma r + \epsilon v(x_j) + \epsilon^2 \zeta + o(\epsilon^2). \tag{A.13}$$

In the outer region, the definition of q_0 in (2.16) implies that

$$-\epsilon^2 \Delta q_o + f'(u_\gamma) q_o - \eta + \epsilon\gamma r + \epsilon v = O(\epsilon^3), \tag{A.14}$$

because $\Delta q_o = O(\epsilon)$ on the outer region.

We now estimate the difference of q_j and q_o on a matching region. First, from (2.10) we find

$$\epsilon^2 \Delta q_j = O(\epsilon^3).$$

Then (A.13) implies that

$$f'(u_\gamma) q_j - \eta + \epsilon\gamma r + \epsilon\gamma v = \sigma + o(\epsilon^2),$$

where

$$\sigma = -\eta + \epsilon\gamma r(x_j) + \epsilon v(x_j) + \epsilon^2 \zeta. \tag{A.15}$$

Comparing this to (2.16) we deduce that

$$q_j - q_0 = O(|\sigma|) + o(\epsilon^2) \tag{A.16}$$

on the matching regions $(x_j - 2\epsilon^\alpha, x_j - \epsilon^\alpha)$ and $(x_j + \epsilon^\alpha, x_j + 2\epsilon^\alpha)$. Then we consider q in the matching region. Here by (A.16)

$$\begin{aligned}
&-\epsilon^2 \Delta q + f'(u_\gamma) q - \eta + \epsilon\gamma r + \epsilon v \\
&= -\epsilon^2 \Delta q_o + f'(u_\gamma) q_o - \eta + \epsilon\gamma r + \epsilon v + O(\|q_j - q_o\|_\infty) \\
&= -\epsilon^2 \Delta q + O(|\sigma|) + o(\epsilon^2) \\
&= -\epsilon^2 \Delta q - \epsilon^2 \Delta(\chi_j (q_j - q_o)) + O(|\sigma|) + o(\epsilon^2) \\
&= -\epsilon^2 (\chi_j''(q_j - q_o) + 2\chi_j'(q_j - q_o)' + \chi_j (q_j - q_o)'') + O(|\sigma|) + o(\epsilon^2) \\
&= O(|\sigma|) + o(\epsilon^2).
\end{aligned} \tag{A.17}$$

If we let $g = u - q$, then (A.13), (A.14), and (A.17) imply that

$$-\epsilon^2 \Delta g + f'(u_\gamma) g = \begin{cases} \sigma + o(\epsilon^2) & \text{in an inner region,} \\ O(|\sigma|) + o(\epsilon^2) & \text{in a matching region,} \\ O(\epsilon^3) & \text{in the outer region.} \end{cases} \tag{A.18}$$

We deduce from (A.18) and $g(x_j) = 0$ that

$$g = O(|\sigma|) + o(\epsilon^2). \tag{A.19}$$

On the other hand, if we multiply (A.18) by H_j' and integrate, then

$$\int_0^1 [(f'(u_\gamma) - f'(H)) g H_j'] \, dx = \epsilon \sigma_j + O(\epsilon^2 |\sigma|) + o(\epsilon^3).$$

But the integral on the left side is $O(\epsilon^2 \|g\|_\infty)$, from which we conclude that

$$\sigma = O(\epsilon \|g\|_\infty) + o(\epsilon^2). \tag{A.20}$$

Inserting (A.20) into (A.19) we find that

$$g = o(\epsilon^2), \tag{A.21}$$

proving Lemma 2.4; and substituting (A.21) into (A.20) we deduce that

$$\sigma = o(\epsilon^2). \tag{A.22}$$

Then we deduce

$$\eta = \epsilon v(x_j) + o(\epsilon). \tag{A.23}$$

Because $\overline{du_\gamma/d\gamma} = 0$,

$$0 = \overline{q} + o(\epsilon^2) = \epsilon \int_0^1 \frac{v^0(x_j^0) - v^0(x)}{f'(0)}\,dx + c\int_0^1 \sum_j h_j + o(\epsilon) = \epsilon\left(\frac{v^0(x_j^0)}{f'(0)} + c_0 K\right) + o(\epsilon).$$

Thus we obtain the value for c_0 in (2.8). ∎

Appendix B. Expansion of 2 ′(0).

In Appendices B and C we use the following simplified notations:

$$u := u_{\mathrm{B}},\ v := G_0[u_{\mathrm{B}} - a],\ \gamma := \gamma_{\mathrm{B}},\ \phi := \phi_{\mathrm{B}},\ \omega := \phi_{\mathrm{B}}^{\perp},\ f' := f'(H),\ f'' := f''(H), \text{ etc.} \tag{B.1}$$

The vector c_j in the expansion of ϕ satisfies $|c| = 1$.

Define a linear operator $\mathcal{L}$ by

$$\mathcal{L}U := U'' - f'U, \tag{B.2}$$

where U is defined on $\mathbf{R}$. Then

$$\mathcal{L}H_t = 0, \tag{B.3}$$

$$\mathcal{L}H_{tt} = f''H_t^2, \tag{B.4}$$

$$\mathcal{L}H_{ttt} = 3f''H_tH_{tt} + f'''H_t^3. \tag{B.5}$$

Let $u = s + \epsilon^2 p$, where s is given in (2.4). Then p satisfies

$$\epsilon^2 p'' - f'p = \frac{1}{\epsilon^2}[\gamma\epsilon G_0[u-a] - \text{const}] + \epsilon^2 f''\frac{p^2}{2} + O(\epsilon^4). \tag{B.6}$$

By Lemma 2.3, as $\epsilon \to 0$, $p(x_j + \epsilon\cdot) \to P$ in $C^2_{loc}(\mathbf{R})$, where P satisfies

$$\mathcal{L}P = \gamma(v^0)'(x_j^0)t. \tag{B.7}$$

Note that $P(t)$ is an odd function (and hence $P \perp H_t$). It is easy to compute that

$$\mathcal{L}P_t = f''H_tP + \gamma(v^0)'(x_j^0), \tag{B.8}$$

$$\mathcal{L}P_{tt} = (f''H_t)_tP + 2f''H_tP_t, \tag{B.9}$$

$$\mathcal{L}P_{ttt} = (f''H_t)_{tt}P + 3(f''H_t)_tP + 3f''H_tP_{tt}. \tag{B.10}$$

Recall Lemma 3.2. Set the decomposition

$$\phi(x) = \sum_{j=1}^{K} c_jh_j + \epsilon^2\omega,\ h_j \perp \omega, \tag{B.11}$$

for the zero principal eigenfunction ϕ. ω satisfies

$$\begin{aligned}&\epsilon^2\omega'' - \epsilon^2(m\pi)^2\omega - f'(u)\omega - \epsilon\gamma G_m[\omega]\\ &= -\frac{1}{\epsilon^2}\sum_{j=1}^{K} c_j[(f' - f'(u))h_j - \epsilon^2(m\pi)^2 h_j] + \frac{\gamma}{\epsilon}\sum_{k=1}^{K} c_k G_m[h_k].\end{aligned} \tag{B.12}$$

We further expand (B.12):

$$\begin{aligned}\epsilon^2\omega'' - f'\omega = &\sum_{j=1}^{K} c_j[f''H_t p - m^2\pi^2 H_t] \ + \frac{\gamma}{\epsilon}\sum_j c_j G_m[h_j]\\ &+\gamma\epsilon G_m[\omega] + \epsilon^2\left[f''p\omega + m^2\pi^2\omega + f'''\left(\sum_j c_j h_j\right)\frac{p^2}{2}\right] + O(\epsilon^3).\end{aligned} \tag{B.13}$$

As $\epsilon \to 0$, we have $\omega(x_j + \epsilon\cdot) \to c_j^0\Omega$ in $C^2_{loc}(\mathbf{R})$ just as in the proof of Lemma 3.2, where Ω satisfies

$$\mathcal{L}\Omega = f''H_t P - (m\pi)^2 H_t + \text{const},\ \Omega \text{ is even, and } \Omega \perp H_t. \tag{B.14}$$

Hence

$$\mathcal{L}\Omega_t = f''H_t\Omega + (f''H_t)_t P + f''H_t P_t - (m\pi)^2 H_{tt}. \tag{B.15}$$

Finally, we calculate $2g'(0)$. Since $\varphi_{\mathrm{B}}^2 = \phi^2(x)\cos^2(m\pi y)$, we decompose the solution of (5.5) into

$$2g'(0)(x,y) = \frac{g_1(x)}{2} + \frac{g_2(x)\cos(2m\pi y)}{2}, \tag{B.16}$$

where g_1 and g_2 are solutions of the following two equations:

$$\epsilon^2 g_1'' - f'(u)g_1 + \overline{f'(u)g_1} - \epsilon\gamma G_0[g_1] = f''(u)\phi^2 - \overline{f''(u)\phi^2},\ \ g_1'(0) = g_1'(1) = \overline{g_1} = 0; \tag{B.17}$$

$$\epsilon^2(g_2'' - 4m^2\pi^2 g_2) - f'(u)g_2 - \epsilon\gamma G_{2m}[g_2] = f''(u)\phi^2,\ g_2'(0) = g_2'(1) = 0. \tag{B.18}$$

Both equations are uniquely solvable, since the eigenvalues of the two operators in (B.17) and (B.18) are nonzero (the zero eigenvalue is associated with m); i.e., both operators are invertible.

We write

$$2g'(0) = \psi_1 + \epsilon^2\psi_2, \tag{B.19}$$

where

$$\psi_1(x,y) = \sum_j c_j^2 H_{tt}\left(\frac{x - x_j}{\epsilon}\right)\cos^2(m\pi y),\ \psi_2 = \frac{g_{11}}{2} + \frac{g_{21}}{2}\cos(2m\pi y). \tag{B.20}$$

Here

$$g_1 = \sum_j c_j^2 H_{tt}\left(\frac{x - x_j}{\epsilon}\right) + \epsilon^2 g_{11},\ g_2 = \sum_j c_j^2 H_{tt}\left(\frac{x - x_j}{\epsilon}\right) + g_{21}. \tag{B.21}$$

The equation for g_{11} is

$$
\begin{aligned}
&\epsilon^2 g_{11}'' - f'(u)g_{11} - \epsilon\gamma G_0[g_{11}] \\
&= \frac{\gamma c_j^2}{\epsilon} G_0[H_{tt}] + \frac{1}{\epsilon^2}[f''(u)\phi_m^2 - c_j^2 f'' H_t^2 - \overline{f''(u)\phi_m^2}] \\
&= \frac{\gamma c_j^2}{\epsilon} G_0[H_{tt}] + c_j^2 f''' H_t^2 p + 2f'' H_t c_j \omega + c_j^2 f'' H_{tt} p \\
&+\epsilon^2 \left[c_j^2 f^{(4)} H_t^2 \frac{p^2}{2} + c_j^2 f''' H_{tt} \frac{p^2}{2} + 2c_j f''' H_t p\omega + f''\omega^2 \right] + O(\epsilon^4) + C_1,
\end{aligned} \tag{B.22}
$$

where $C_1 = \epsilon^{-2}\overline{f''(u)\phi_m^2}$. By (B.15), it is easy to see that

$$
C_1 = \frac{1}{\epsilon^2}\int_0^1 f''(u)\phi_m^2 = \frac{1}{\epsilon^2}\sum_{j=1}^K f''(H+\epsilon^2 p)c_j^2 H_t^2 + o(1) = o(1). \tag{B.23}
$$

Similarly, the equation for g_{21} is

$$
\begin{aligned}
&\epsilon^2 g_{21}'' - 4m^2\pi^2 g_{21} - f'(u)g_{21} - \epsilon\gamma G_{2m}[g_{21}] \\
&= 4m^2\pi^2 c_j^2 H_{tt} + \frac{\gamma c_j^2}{\epsilon} G_{2m}[H_{tt}] + c_j^2 f''' H_t^2 p + 2f'' H_t c_j \omega + c_j^2 f'' H_{tt} p \\
&+\epsilon^2 \left[c_j^2 f^{(4)} H_t^2 \frac{p^2}{2} + c_j^2 f''' H_{tt} \frac{p^2}{2} + 2c_j f''' H_t p\omega + f''\omega^2 \right] + O(\epsilon^4).
\end{aligned} \tag{B.24}
$$

We state the following lemma.

LEMMA B.1. *As $\epsilon \to 0$, near x_j we have $g_{11}(x_j+\epsilon\cdot) \to (c_j^0)^2 G_{11}, g_{21}(x_j+\epsilon\cdot) \to (c_j^0)^2 G_{21}$, where G_{11} satisfies*

$$
\mathcal{L}G_{11} = f''' H_t^2 P + 2f'' H_t \Omega + f'' H_{tt} P,\ G_{11} \text{ is odd}
$$

and G_{21} satisfies

$$
\mathcal{L}G_{21} = f''' H_t^2 P + 2f'' H_t \Omega + f'' H_{tt} P + (2m\pi)^2 H_{tt},\ G_{21} \text{ is odd.}
$$

Proof. We only prove the convergence of g_{11}. The convergence of g_{21} is similar. To this end, let us decompose

$$
g_{11} = \sum_{j=1}^K \alpha_j (h_j - \overline{h_j}) + G_{11} + \hat{g}_{11},
$$

where $\hat{g}_{11} \perp h_j, j = 1, \ldots, K$, and $\int_0^1 \hat{g}_{11} = 0$. The key is to show that $\alpha_j = o(1)$.

Simple calculations show that $\hat{g}_{11}$ satisfies

$$
\epsilon^2 \hat{g}_{11}'' - f'(u)\hat{g}_{11} - \epsilon\gamma G_0[\hat{g}_{11}] = O\left(\epsilon^2 \sum_{j=1}^K |\alpha_j|\right) + o(1).
$$

Since $\hat{g}_{11} \perp h_j, j = 1, \ldots, K$, and $\int \hat{g}_{11} = 0$, standard arguments show that

$$
\hat{g}_{11} = O\left(\sum_{j=1}^K |\alpha_j|\epsilon^2\right) + o(1). \tag{B.25}
$$

We multiply (B.22) by h_j and integrate over $(0,1)$ to find

$$\begin{aligned}\epsilon C_1+\alpha_j\int_0^1(f'-f'(u))H_t^2 &= \int_0^1[c_j^2f'''H_t^2p+2c_jf''H_t\omega+c_j^2f''H_{tt}p]H_t+o(\epsilon^3)\\ &= \int_0^1[c_j^2f'''H_t^2p+c_j^2f''H_{tt}p]H_t+2c_j\int_0^1\mathcal{L}H_{tt}\omega+o(\epsilon^3)\\ &= \int_0^1[c_j^2f'''H_t^2p+3c_j^2f''H_{tt}p]H_t+o(\epsilon^3)\\ &= c_j^2\int_0^1(\mathcal{L}H_{ttt})p+o(\epsilon^3) \;=\; o(\epsilon^3).\end{aligned}$$

Thus we obtain the first identity

$$C_1+\epsilon^2\alpha_j\int_{\mathbf{R}}f''PH_t^2=o(\epsilon^2). \tag{B.26}$$

Next, we integrate (B.22) over $(0,1)$ and make use of (B.25) to deduce that

$$0=\int_0^1 f'(u)g_{11}=\int_0^1 f'(u)\left(\sum_j\alpha_j(h_j-\overline{h_j})+\hat{g}_{11}\right).$$

Thus we obtain the second identity

$$\sum_j\alpha_jf'(0)=o(1). \tag{B.27}$$

Substituting (B.27) into (B.26), we have that

$$C_1=o(\epsilon^2),\ \alpha_j=o(1), \tag{B.28}$$

and hence $\hat{g}_{11}=o(1)$. ∎

Appendix C. Proof of Lemma 5.2.

In this appendix we omit $\sum_j$ most of the time. When c_j appears in a quantity, $\sum_j$ is usually implied. We use the notation $A\approx B$ for $A-B=o(\epsilon^5)$.

Define a linear operator S by

$$S\psi:=\epsilon^2\Delta\psi-f'(u)\psi+\epsilon\gamma\Delta^{-1}\psi, \tag{C.1}$$

where ψ is a function on D. Recall ψ_1 and ψ_2 defined in (B.20). Note

$$\begin{aligned}S\psi_1 &= c_j^2\{(f'-f'(u))H_{tt}\cos^2(m\pi y)+f''H_t^2\cos^2(m\pi y)\\ &\quad -2\epsilon^2(m\pi)^2H_{tt}\cos(2m\pi y)+\epsilon\gamma\Delta^{-1}(H_{tt}\cos^2(m\pi y))\},\end{aligned} \tag{C.2}$$

$$\begin{aligned}S\psi_2 &= 2(m\pi)^2c_j^2H_{tt}\cos(2m\pi y)-\frac{\gamma c_j^2}{\epsilon}\Delta^{-1}(H_{tt}\cos^2(m\pi y))\\ &\quad +\frac{1}{\epsilon^2}(f''(u)\phi^2-c_j^2f''H_t^2)\cos^2(m\pi y)+\frac{c_j^2}{\epsilon^2}(f'(u)-f')H_{tt}\cos^2(m\pi y).\end{aligned} \tag{C.3}$$

Then

$$\begin{aligned}\int_D f''(u)\varphi^2(2g'(0)) \;&=\; \int_D(S(2g'(0)))(2g'(0))\\ &= \int_D(S\psi_1)\psi_1+2\epsilon^2\int_D(S\psi_2)\psi_1+\epsilon^4\int_D(S\psi_2)\psi_2 \;:=\; I_1+I_2+I_3,\end{aligned} \tag{C.4}$$

where the last equation defines I_1, I_2, and I_3. To prove Lemma 5.2 we compute

$$\int_D f''(u)\phi^2(2g'(0)) + \frac{1}{3}\int_D f'''(u)\phi^4 = I_1 + I_2 + I_3 + \frac{1}{3}\int_D f'''(u)\phi^4. \tag{C.5}$$

We start with I_2. From (C.3) we obtain

$$\begin{aligned}
I_2 &\approx 4\epsilon^3(m\pi)^2 c_j^4 \int_{\mathbf{R}} H_{tt}^2 \int_0^1 \cos^2(m\pi y)\cos(2m\pi y) \\
&\quad -2\epsilon\gamma c_j^4 \int_D (\Delta^{-1}(H_{tt}\cos^2(m\pi y)))H_{tt}\cos^2(m\pi y) \\
&\quad +2\epsilon c_j^2 \int_{\mathbf{R}} (f''(u)\phi^2 - c_j^2 f'' H_t^2)H_{tt} \int_0^1 \cos^4(m\pi y) \\
&\quad +2\epsilon c_j^4 \int_{\mathbf{R}} (f'(u) - f')H_{tt}^2 \int_0^1 \cos^4(m\pi y) \\
&\approx \epsilon^3(m\pi)^2 c_j^4 \int_{\mathbf{R}} H_{tt}^2 + \epsilon\gamma c_j^4 \int_0^1 (2G_0 + G_{2m})[H_{tt}]\frac{H_{tt}}{4} \\
&\quad +\frac{3\epsilon c_j^2}{4}\int_{\mathbf{R}} (f''(u)\phi^2 - c_j^2 f'' H_t^2)H_{tt} + \frac{3\epsilon c_j^4}{4}\int_{\mathbf{R}} (f'(u) - f')H_{tt}^2.
\end{aligned} \tag{C.6}$$

The last two terms in (C.6) are estimated as follows:

$$\begin{aligned}
&\frac{3\epsilon}{4}\int_{\mathbf{R}} (f''(u)\phi^2 - c_j^2 f'' H_t^2)H_{tt} \\
&\approx \frac{3\epsilon^3}{4}\int_{\mathbf{R}} \left(c_j^2 f''' p H_t^2 + 2c_j f'' H_t\omega + \epsilon^2 f''\omega^2 + 2\epsilon^2 c_j f''' p\omega + \epsilon^2 c_j^2 f^{(4)}\frac{p^2}{2}H_t^2\right)H_{tt}, \\
&\frac{3\epsilon}{4}\int_{\mathbf{R}} (f'(u) - f')H_{tt}^2 \\
&\approx \frac{3\epsilon^3}{4}\int_{\mathbf{R}} \left(f'' H_{tt}^2 p + \epsilon^2 f''' H_{tt}^2 \frac{p^2}{2}\right).
\end{aligned} \tag{C.7}$$

We substitute (C.7) into (C.6) to obtain

$$\begin{aligned}
I_2 &\approx \epsilon^3(m\pi)^2 c_j^4 \int_{\mathbf{R}} H_{tt}^2 + \epsilon\gamma c_j^4 \int_0^1 (2G_0 + G_{2m})[H_{tt}]\frac{H_{tt}}{4} \\
&\quad +\frac{3\epsilon^3 c_j^3}{4}\int_{\mathbf{R}} (c_j f''' H_t^2 H_{tt} p + c_j f'' H_{tt}^2 p + 2f'' H_t H_{tt}\omega) \\
&\quad +\frac{3\epsilon^5 c_j^2}{4}\int_{\mathbf{R}} \left(f'' H_{tt}\omega^2 + 2c_j f''' H_t H_{tt} p\omega + c_j^2 (f^{(4)} H_t^2 H_{tt} + f''' H_{tt}^2)\frac{p^2}{2}\right).
\end{aligned} \tag{C.8}$$

On the other hand,

$$\begin{aligned}
\frac{1}{3}\int_D f'''(u)\varphi^4 &= \frac{1}{8}\int_0^1 f'''(H + \epsilon^2 p)(c_j H_t + \epsilon^2\omega)^4 \\
&\approx \frac{\epsilon c_j^4}{8}\int_{\mathbf{R}} f''' H_t^4 + \frac{\epsilon^3 c_j^3}{8}\int_{\mathbf{R}} (c_j f^{(4)} H_t^4 p + 4f''' H_t^3\omega) \\
&\quad +\frac{\epsilon^5 c_j^2}{8}\int_{\mathbf{R}} \left(c_j^2 f^{(5)} H_t^4 \frac{p^2}{2} + 4c_j f^{(4)} H_t^3 p\omega + 6f''' H_t^2\omega^2\right).
\end{aligned}$$

We combine the last with (C.8) to deduce

$$
\begin{aligned}
& I_2 + \frac{1}{3}\int_D f'''(u)\varphi^4 \\
& \approx \frac{\epsilon c_j^4}{8}\int_{\mathbf{R}} f''' H_t^4 + \epsilon^3 (m\pi)^2 c_j^4 \int_{\mathbf{R}} H_{tt}^2 + \epsilon\gamma c_j^4 \int_0^1 (2G_0 + G_{2m})[H_{tt}]\frac{H_{tt}}{4} \\
& \quad + \frac{\epsilon^3 c_j^3}{8}\int_{\mathbf{R}} \{c_j (f^{(4)} H_t^4 + 6 f''' H_t^2 H_{tt} + 6 f'' H_{tt}^2) p + 4 (f''' H_t^3 + 3 f'' H_t H_{tt})\omega\} \\
& \quad + \frac{\epsilon^5 c_j^2}{8}\int_{\mathbf{R}} \Big\{ (6 f''' H_t^2 + 6 f'' H_{tt})\omega^2 + c_j (4 f^{(4)} H_t^3 + 12 f''' H_t H_{tt}) p\omega \\
& \quad + c_j^2 (f^{(5)} H_t^4 + 6 f^{(4)} H_t^2 H_{tt} + 6 f''' H_{tt}^2)\frac{p^2}{2}\Big\}.
\end{aligned} \tag{C.9}
$$

One term in the integral after $\epsilon^3 c_j^3/8$ is simplified using (B.5) and (B.13):

$$
\begin{aligned}
& \int_{\mathbf{R}} (f''' H_t^3 + 3 f'' H_t H_{tt})\omega \\
& = \int_{\mathbf{R}} (\mathcal{L} H_{ttt})\omega \;=\; \int_{\mathbf{R}} (\mathcal{L}\omega) H_{ttt} \;=\; \int_{\mathbf{R}} \{(f'(u) - f')H_{ttt}\omega + (\omega'' - f'(u)\omega) H_{ttt}\} \\
& = \epsilon^2 \int_{\mathbf{R}} f'' H_{ttt} p\omega + \epsilon^2 (m\pi)^2 \int_{\mathbf{R}} H_{ttt}\omega + \epsilon\gamma \int_0^1 G_m[\omega] H_{ttt} - (m\pi)^2 c_j \int_{\mathbf{R}} H_{tt}^2 \\
& + \frac{\gamma c_j}{\epsilon}\int_0^1 G_m[H_t] H_{ttt} + c_j \int_{\mathbf{R}} \left(f'' H_t H_{ttt} p + \epsilon^2 f''' H_t H_{ttt} \frac{p^2}{2}\right) + o(\epsilon^2).
\end{aligned} \tag{C.10}
$$

Here we have dropped $\int_0^1 G_m[\omega] H_{ttt} = \int_0^1 G_m[H_{ttt}]\omega = o(\epsilon)$. Substituting (C.10) into (C.9) we deduce

$$
\begin{aligned}
& I_2 + \frac{1}{3}\int_D f'''(u)\varphi^4 \\
& \approx \frac{\epsilon c_j^4}{8}\int_{\mathbf{R}} f''' H_t^4 + \frac{\epsilon^3 (m\pi)^2 c_j^4}{2}\int_{\mathbf{R}} H_{tt}^2 + \frac{\epsilon^5 (m\pi)^2 c_j^4}{2}\int_{\mathbf{R}} H_{ttt}\omega \\
& \quad + \epsilon\gamma c_j^4 \int_0^1 \left\{ (2G_0 + G_{2m})[H_{tt}]\frac{H_{tt}}{4} + G_m[H_t]\frac{H_{ttt}}{2}\right\} \\
& \quad + \frac{\epsilon^3 c_j^4}{8}\int_{\mathbf{R}} (f^{(4)} H_t^4 + 6 f''' H_t^2 H_{tt} + 6 f'' H_{tt}^2 + 4 f'' H_t H_{ttt}) p \\
& \quad + \frac{\epsilon^5 c_j^2}{8}\int_{\mathbf{R}} \Big\{ (6 f''' H_t^2 + 6 f'' H_{tt})\omega^2 + c_j (4 f^{(4)} H_t^3 + 12 f''' H_t H_{tt} + 4 f'' H_{ttt}) p\omega \\
& \quad + c_j^2 (f^{(5)} H_t^4 + 6 f^{(4)} H_t^2 H_{tt} + 4 f''' H_t H_{ttt} + 6 f''' H_{tt}^2)\frac{p^2}{2}\Big\}.
\end{aligned} \tag{C.11}
$$

Next we compute I_1. From (C.2) we deduce

$$
\begin{aligned}
I_1 \approx & \frac{3\epsilon c_j^4}{8}\int_{\mathbf{R}} f'' H_t^2 H_{tt} + \frac{3\epsilon c_j^4}{8}\int_{\mathbf{R}} (f' - f'(u)) H_{tt}^2 - \frac{\epsilon^3 (m\pi)^2 c_j^4}{2}\int_{\mathbf{R}} H_{tt}^2 \\
& + \epsilon\gamma c_j^4 \int_D (\Delta^{-1}(H_{tt}\cos^2(m\pi y)))(H_{tt}\cos^2(m\pi y))
\end{aligned}
$$

$$\approx \frac{3\epsilon c_j^4}{8}\int_{\mathbf{R}} f'' H_t^2 H_{tt} - \frac{3\epsilon^3 c_j^4}{8}\int_{\mathbf{R}} f'' H_{tt}^2 p - \frac{3\epsilon^5 c_j^4}{8}\int_{\mathbf{R}} f''' H_{tt}^2 \frac{p^2}{2} - \frac{\epsilon^3 (m\pi)^2 c_j^4}{2}\int_{\mathbf{R}} H_{tt}^2$$

$$-\epsilon\gamma c_j^4 \int_0^1 (2G_0 + G_{2m})[H_{tt}]\frac{H_{tt}}{8}. \tag{C.12}$$

(C.12) is added to (C.11). The ϵ order terms and the $\epsilon^3(m\pi)^2$ terms cancel out:

$$\begin{aligned}
&I_1 + I_2 + \frac{1}{3}\int_D f'''(u)\varphi^4 \\
&\approx \frac{\epsilon^5 (m\pi)^2 c_j^3}{2}\int_{\mathbf{R}} H_{ttt}\omega + \epsilon\gamma c_j^4 \int_0^1 \left\{(2G_0 + G_{2m})[H_{tt}]\frac{H_{tt}}{8} + G_m[H_t]\frac{H_{ttt}}{2}\right\} \\
&\quad + \frac{\epsilon^3 c_j^4}{8}\int_{\mathbf{R}} (f^{(4)} H_t^4 + 6f''' H_t^2 H_{tt} + 3f'' H_{tt}^2 + 4f'' H_t H_{ttt})p \\
&\quad + \frac{\epsilon^5 c_j^2}{8}\int_{\mathbf{R}} \Big\{(6f''' H_t^2 + 6f'' H_{tt})\omega^2 + c_j(4f^{(4)} H_t^3 + 12 f''' H_t H_{tt} + 4f'' H_{ttt})p\omega \\
&\quad + c_j^2 (f^{(5)} H_t^4 + 6f^{(4)} H_t^2 H_{tt} + 4f''' H_t H_{ttt} + 3f''' H_{tt}^2)\frac{p^2}{2}\Big\}.
\end{aligned} \tag{C.13}$$

The integral after $\epsilon^3 c_j^4/8$ is, by (B.6),

$$\begin{aligned}
&\int_{\mathbf{R}} (f^{(4)} H_t^4 + 6f''' H_t^2 H_{tt} + 3f'' H_{tt}^2 + 4f'' H_t H_{ttt})p \\
&= \int_{\mathbf{R}} (\mathcal{L} H_{tttt})p \;=\; \int_{\mathbf{R}} (\mathcal{L} p) H_{tttt} \\
&= \frac{\gamma}{\epsilon}\int_{\mathbf{R}} v H_{tttt} + \epsilon^2 \int_{\mathbf{R}} f'' H_{tttt}\frac{p^2}{2} + o(\epsilon^2) \;=\; \epsilon\gamma\int_{\mathbf{R}} v_{xx} H_{tt} + \epsilon^2\int_{\mathbf{R}} f'' H_{tttt}\frac{p^2}{2} + o(\epsilon^2) \\
&= -\epsilon\gamma \int_{\mathbf{R}} (H + \epsilon^2 p - a) H_{tt} + \epsilon^2 \int_{\mathbf{R}} f'' H_{tttt}\frac{p^2}{2} \;=\; \epsilon\gamma\int_{\mathbf{R}} H_t^2 + \epsilon^2 \int_{\mathbf{R}} f'' H_{tttt}\frac{p^2}{2} + o(\epsilon^2).
\end{aligned}$$

Hence (C.13) becomes

$$\begin{aligned}
&I_1 + I_2 + \frac{1}{3}\int_D f'''(u)\varphi^4 \\
&\approx \epsilon\gamma c_j^4 \int_0^1 \left\{(2G_0 + G_{2m})[H_{tt}]\frac{H_{tt}}{8} + G_m[H_t]\frac{H_{ttt}}{2}\right\} \\
&\quad + \frac{\epsilon^4 \gamma c_j^4}{8}\int_{\mathbf{R}} H_t^2 + \frac{\epsilon^5 (m\pi)^2 c_j^3}{2}\int_{\mathbf{R}} H_{ttt}\omega \\
&\quad + \frac{\epsilon^5 c_j^2}{8}\int_{\mathbf{R}} \Big\{(6f''' H_t^2 + 6f'' H_{tt})\omega^2 + c_j(4f^{(4)} H_t^3 + 12f''' H_t H_{tt} + 4f'' H_{ttt})p\omega \\
&\quad + c_j^2 (f^{(5)} H_t^4 + 6f^{(4)} H_t^2 H_{tt} + 4f''' H_t H_{ttt} + 3f''' H_{tt}^2 + f'' H_{tttt})\frac{p^2}{2}\Big\} \\
&= \epsilon\gamma c_j^4 \int_0^1 \left\{(2G_0 + G_{2m})[H_{tt}]\frac{H_{tt}}{8} + G_m[H_t]\frac{H_{ttt}}{2}\right\} \\
&\quad + \frac{\epsilon^4\gamma c_j^4}{8}\int_{\mathbf{R}} H_t^2 + \frac{\epsilon^5 (m\pi)^2 c_j^3}{2}\int_{\mathbf{R}} H_{ttt}\omega \\
&\quad + \frac{\epsilon^5 c_j^2}{8}\int_{\mathbf{R}} \left\{6(f'' H_t)_t \omega^2 + 4c_j (f'' Ht)_{tt} p\omega + c_j^2 (f'' H_t)_{ttt}\frac{p^2}{2}\right\}.
\end{aligned} \tag{C.14}$$

Finally, we compute I_3. By (C.3) we find

$$\begin{aligned}
I_3 &= \epsilon^4 \int_D (S\psi_2)\psi_2 \\
&\approx \epsilon^4 \Bigg\{ c_j^3 \int_D (c_j(f''H_t)_t p + 2f''H_t\omega)\cos^2(m\pi y)\left(\frac{g_{11}}{2} + \frac{g_{21}}{2}\cos(2m\pi y)\right) \\
&\quad +2(m\pi)^2 c_j^4 \int_D H_{tt}\cos(2m\pi)\left(\frac{g_{11}}{2} + \frac{g_{21}}{2}\cos(2m\pi y)\right) \\
&\quad -\frac{\gamma c_j^4}{\epsilon}\int_D (\Delta^{-1}(H_{tt}\cos^2(m\pi y)))\left(\frac{g_{11}}{2} + \frac{g_{21}}{2}\cos(2m\pi y)\right)\Bigg\} \\
&\approx \frac{\epsilon^5 (c_j^0)^4}{8}\int_{\mathbf{R}} ((f''(H_t)_t P + 2f''H_t\Omega)(2G_{11} + G_{21})) + \frac{\epsilon^5 (m\pi)^2 (c_j^0)^4}{2}\int_{\mathbf{R}} H_{tt}G_{21},
\end{aligned} \tag{C.15}$$

where we have used Lemma B.1. We have dropped the last integral of the second-last line because it is of order $o(\epsilon^2)$. Combining (C.14) and (C.15) we arrive at

$$\begin{aligned}
&I_1 + I_2 + I_3 + \frac{1}{3}\int_D f'''(u)\varphi^4 \\
&\approx \epsilon\gamma c_j^4 \int_0^1 \left\{(2G_0 + G_{2m})[H_{tt}]\frac{H_{tt}}{8} + G_m[H_t]\frac{H_{ttt}}{2}\right\} \\
&\quad + \frac{\epsilon^4\gamma c_j^4}{8}\int_{\mathbf{R}} H_t^2 + \frac{\epsilon^5 (m\pi)^2 (c_j^0)^4}{2}\int_{\mathbf{R}} H_{tt}(G_{21} - \Omega_t) \\
&\quad + \frac{\epsilon^5 (c_j^0)^4}{8}\int_{\mathbf{R}} \Bigg\{ 6(f''H_t)_t\Omega^2 + 4(f''Ht)_{tt}P\Omega + (f''H_t)_{ttt}\frac{P^2}{2} \\
&\quad +3((f''H_t)_t P + 2f''H_t\Omega)\Gamma \Bigg\}.
\end{aligned} \tag{C.16}$$

Here we have introduced

$$\Gamma := \frac{2G_{11} + G_{21}}{3}. \tag{C.17}$$

We simplify the last integral in (C.16). Let

$$\Omega = P_t + \Pi,\ \Gamma = P_{tt} + \Psi. \tag{C.18}$$

Note that by (3.12) and $\lambda(\gamma_{\mathrm{B}}) = 0$,

$$\mathcal{L}\Pi = (m\pi)^2 H_t + \text{const}, \quad \mathcal{L}\Psi = 2f''H_t\Pi + \frac{4(m\pi)^2}{3}H_{tt}. \tag{C.19}$$

The integral after $\epsilon^5 (c_j^0)^4/8$ in (C.16) is

$$\begin{aligned}
&\int_{\mathbf{R}} \left\{ (f''H_t)_{ttt}\frac{P^2}{2} + 6(f''H_t)_t P_t^2 + 4(f''H_t)_{tt}PP_t + 3((f''H_t)_t P + 2f''H_t P_t)P_{tt} \right\} \\
&+ \int_{\mathbf{R}} \{ 4(f''H_t)_{tt}P\Pi + 12(f''H_t)_t P_t\Pi + 6(f''H_t)_t\Pi^2 \\
&\quad +3((f''H)_t P + 2f''H_t P_t)\Psi + 6f''H_t\Pi\Psi + 6f''H_t\Pi P_{tt} \Bigg\}.
\end{aligned} \tag{C.20}$$

The first integral in (C.20) is 0 after integration by parts. To calculate the second integral note, by (B.9),

(C.21)

$$\int_{\mathbf{R}} 3((f''H_t)_t P + 2f''H_t P_t)\Psi$$
$$= 3\int_{\mathbf{R}} (\mathcal{L}P_{tt})\Psi = 3\int_{\mathbf{R}} (\mathcal{L}\Psi)P_{tt} = 6\int_{\mathbf{R}} f''H_t\Pi P_{tt} + 4(m\pi)^2\int_{\mathbf{R}} H_{tt}P_{tt},$$

(C.22)

$$\int_{\mathbf{R}} 6f''H_t\Pi\Psi$$
$$= 6\int_{\mathbf{R}} (\mathcal{L}\Pi_t - (m\pi)^2 H_{tt})\Psi \;=\; 6\int_{\mathbf{R}} \left(2f''H_t\Pi + \frac{4(m\pi)^2}{3}H_{tt}\right)\Pi_t - 6(m\pi)^2\int_{\mathbf{R}} H_{tt}\Psi$$
$$= -6\int_{\mathbf{R}} (f''H_t)_t\Pi^2 + 8(m\pi)^2\int_{\mathbf{R}} H_{tt}\Pi_t - 6(m\pi)^2\int_{\mathbf{R}} H_{tt}\Psi.$$

Substituting (C.21) and (C.22) into the second integral in (C.20) we find, with the help of (B.10) and (C.19),

(C.23)

$$\begin{aligned}
(\text{C.20}) = & \int_{\mathbf{R}} (4(f'H_t)_{tt}P + 12(f''H_t)_t P_t + 12f''H_t P_{tt})\Pi \\
& +4(m\pi)^2\int_{\mathbf{R}} H_{tt}P_{tt} + 8(m\pi)^2\int_{\mathbf{R}} H_{tt}\Pi_t - 6(m\pi)^2\int_{\mathbf{R}} H_{tt}\Psi \\
= & \, 4\int_{\mathbf{R}} (\mathcal{L}P_{ttt})\Pi + 4(m\pi)^2\int_{\mathbf{R}} H_{tt}P_{tt} + 8(m\pi)^2\int_{\mathbf{R}} H_{tt}\Pi_t - 6(m\pi)^2\int_{\mathbf{R}} H_{tt}\Psi \\
= & \, 4(m\pi)^2\int_{\mathbf{R}} H_t P_{ttt} + 4(m\pi)^2\int_{\mathbf{R}} H_{tt}P_{tt} + 8(m\pi)^2\int_{\mathbf{R}} H_{tt}\Pi_t - 6(m\pi)^2\int_{\mathbf{R}} H_{tt}\Psi \\
= & \, 8(m\pi)^2\int_{\mathbf{R}} H_{tt}\Pi_t - 6(m\pi)^2\int_{\mathbf{R}} H_{tt}\Psi.
\end{aligned}$$

(C.24)

Substituting (C.24) back into (C.16) we find

$$\begin{aligned}
& I_1 + I_2 + I_3 + \frac{1}{3}\int_D f'''(u)\varphi^4 \\
& \approx \epsilon\gamma c_j^4 \int_0^1 \left\{(2G_0 + G_{2m})[H_{tt}]\frac{H_{tt}}{8} + G_m[H_t]\frac{H_{ttt}}{2}\right\} + \frac{\epsilon^4\gamma c_j^4}{8}\int_{\mathbf{R}} H_t^2 \\
& + \frac{\epsilon^5(m\pi)^2(c_j^0)^4}{2}\int_{\mathbf{R}} H_{tt}\left(G_{21} - \Omega_t + 2\Pi_t - \frac{3}{2}\Psi\right).
\end{aligned} \tag{C.25}$$

Note that

$$\begin{aligned}
& \mathcal{L}\left(G_{21} - \Omega_t + 2\Pi_t - \frac{3}{2}\Psi\right) \\
& = 4(m\pi)^2 H_{tt} + f'''H_t^2 P + 2f''H_t\Omega + f''H_{tt}P \\
& \quad -2f''H_t\Omega - (f''H_t)_t P + f''H_t\Pi - (m\pi)^2 H_{tt}
\end{aligned} \tag{C.26}$$

$$+2f''H_t\Pi + 2(m\pi)^2 H_{tt} - \left(\frac{3}{2}\right) 2f''H_t\Pi - \left(\frac{3}{2}\right) \frac{4(m\pi)^2 H_{tt}}{3}$$
$$= 3(m\pi)^2 H_{tt}.$$

On the other hand, we may solve the last equation to find

$$G_{21} - \Omega_t + 2\Pi_t - \frac{3}{2}\Psi = \frac{3(m\pi)^2}{2} tH_t$$

since $\mathcal{L}(\frac{t}{2}H_t) = H_{tt}$. Hence the last integral in (C.25) is

$$\int_{\mathbf{R}} H_{tt}\left(G_{21} - \Omega_t + 2\Pi_t - \frac{3}{2}\Psi\right) = \frac{3(m\pi)^2}{2}\int_{\mathbf{R}} tH_tH_{tt} = -\frac{3(m\pi)^2\tau}{4}.$$

Putting this back into (C.25) we deduce

$$I_1 + I_2 + I_3 + \frac{1}{3}\int_D f'''(u)\varphi^4$$
$$\approx \epsilon\gamma c_j^4 \int_0^1 \left\{(2G_0 + G_{2m})[H_{tt}]\frac{H_{tt}}{8} + G_m[H_t]\frac{H_{ttt}}{2}\right\} + \frac{\epsilon^4\gamma c_j^4}{8}\int_{\mathbf{R}} H_t^2 - \frac{3\epsilon^5(m\pi)^2\tau(c_j^0)^4}{8}. \tag{C.27}$$

We now compute the first term in (C.27). Note that

$$\int_0^1 G_0[H_{tt}]H_{tt} = \epsilon^3\int_{\mathbf{R}} H_t^2 + o(\epsilon^4), \tag{C.28}$$

since $G_0[H_{tt}] = \epsilon^2(\overline{H(\frac{\cdot - x_j}{\epsilon})} - H(\frac{\cdot - x_j}{\epsilon}))$.

Recall that G_{2m} is identified with the Green function of

$$-G_{2m}'' + (2m\pi)^2 G_{2m} = \delta(\cdot - y), \quad G_{2m}'(0,y) = G_{2m}'(1,y) = 0.$$

G_{2m} splits to the fundamental solution part and the regular part:

$$G_{2m}(x,y) = \frac{1}{4m\pi}e^{-2m\pi|x-y|} - R_{2m}(x,y).$$

Note that R_{2m} is smooth in both variables x and y. We write down $G_{2m}(x,y)$ explicitly:

$$G_{2m}(x,y) = \frac{\cosh(2m\pi(1-|x-y|)) + \cosh(2m\pi(1-x-y))}{4m\pi\sinh(2m\pi)}.$$

Thus

$$R_{2m}(x,y) = \frac{1}{4m\pi}e^{-2m\pi|x-y|} - \frac{\cosh(2m\pi(1-|x-y|)) + \cosh(2m\pi(1-x-y))}{4m\pi\sinh(2m\pi)}.$$

We need to compute

$$R_{2m,xy}(y,y) := \left.\frac{\partial^2 R_{2m}}{\partial x\partial y}\right|_{x=y} = -m\pi + \frac{m\pi\cosh(2m\pi) - 2m\pi\cosh(2m\pi(1-2y))}{\sinh(2m\pi)}. \tag{C.29}$$

Then we have

$$G_{2m}[H_{tt}](x) = \int_0^1 G_{2m}(x,y) H_{tt}\left(\frac{y-x_j}{\epsilon}\right) dy.$$

By simple computations, we have that

$$\begin{aligned} G_{2m}[H_{tt}](x_j+\epsilon t) &= \epsilon \int_{-x_j/\epsilon}^{(1-x_j)/e} G_{2m}(x_j+\epsilon t, x_j+\epsilon z) H_{tt}(z)\, dz \\ &= \epsilon \int_{\mathbf{R}} \left[\frac{1}{4m\pi} e^{-2m\pi\epsilon|t-z|} - R_{2m}(x_j+\epsilon t, x_j+\epsilon z)\right] H_{zz} dz + o(\epsilon^4). \end{aligned} \tag{C.30}$$

We expand $e^{-2m\pi\epsilon|t-z|}$ to deduce

$$\begin{aligned} \int_{\mathbf{R}} e^{-2m\pi\epsilon|t-z|} H_{zz}\, dz &= \int_{\mathbf{R}} (1 - 2m\pi\epsilon|t-z| + 2(m\pi\epsilon)^2|t-z|^3 + O(\epsilon^3|t-z|^3)) H_{zz}\, dz \\ &= -4m\pi\epsilon H(t) + 4(m\pi\epsilon)^2 t + O(\epsilon^3). \end{aligned}$$

Hence (C.30) becomes

$$G_{2m}[H_{tt}](x_j+\epsilon t) = -\epsilon^2 H(t) + m\pi\epsilon^3 t - \epsilon \int_{\mathbf{R}} R_{2m}(x_j+\epsilon t, x_j+\epsilon z) H_{zz}\, dz. \tag{C.31}$$

Next we expand $R_{2m}(x_j+\epsilon t, x_j+\epsilon z)$ so that

$$\int_0^1 G_{2m}[H_{tt}]H_{tt} = \epsilon^3 \int_{\mathbf{R}} H_t^2 - m\pi\epsilon^4 - \epsilon^4 R_{2m,xy}(x_j^0, x_j^0) + o(\epsilon^4). \tag{C.32}$$

For the term involving G_m, by integrating by parts we obtain

$$\int_0^1 G_m[H_t]H_{ttt} = -\int_0^1 G_m^D[H_{tt}]H_{tt},$$

where $G_m^D[H_{tt}]$ is the Green function of

$$-(G_m^D)'' + (m\pi)^2 G_m^D = \delta(\cdot - y), \quad G_m^D(0,y)(0) = G_m^D(1,y) = 0. \tag{C.33}$$

The superscript D emphasizes the Dirichlet boundary condition. Similar to the Neumann boundary case we find

$$\begin{aligned} G_m^D(x,y) &= \frac{\cosh(m\pi(1-|x-y|)) - \cosh(m\pi(1-x-y))}{2m\pi\sinh(m\pi)}, \\ R_m^D(x,y) &:= \frac{1}{2m\pi} e^{-m\pi|x-y|} - \frac{\cosh(m\pi(1-|x-y|)) - \cosh(m\pi(1-x-y))}{2m\pi\sinh(m\pi)}, \\ R_{m,xy}^D(y,y) &:= \left.\frac{\partial^2 R_m^D}{\partial x \partial y}\right|_{x=y} = -\frac{m\pi}{2} + \frac{m\pi\cosh(m\pi) + m\pi\cosh(m\pi(1-2y))}{2\sinh(m\pi)}. \end{aligned} \tag{C.34}$$

By the same argument leading to (C.32), we arrive at

$$\int_0^1 G_m[H_t]H_{ttt} = -\epsilon^3 \int_{\mathbf{R}} H_t^2 + \frac{\epsilon^4 m\pi}{2} + \epsilon^4 R_{m,xy}^D(x_j^0, x_j^0) + o(\epsilon^4). \tag{C.35}$$

Substituting (C.28), (C.32), and (C.35) into (C.27), we obtain

$$\begin{aligned}
&I_1 + I_2 + I_3 + \frac{1}{3}\int_D f'''(u)\varphi^4 \\
&\approx \sum_{j=1}^K c_j^4 \left[\frac{\epsilon^5 \gamma m\pi}{8} - \frac{\epsilon^5\gamma}{8} R_{2m,xy}(x_j^0, x_j^0) + \frac{\epsilon^5\gamma}{2} R^D_{m,xy}(x_j^0, x_j^0) - \frac{3\epsilon^5 (m\pi)^4 \tau}{8} \right] \\
&\approx \epsilon^5 m\pi\gamma \sum_{j=1}^K (c_j^0)^4 \left[\frac{2+\cosh(2m\pi)}{8\sinh(2m\pi)} + \frac{\cosh(2m\pi(1-2x_j^0))}{8\sinh(2m\pi)} \right. \\
&\qquad\qquad \left. + \frac{\cosh(m\pi(1-2x_j^0))}{4\sinh(m\pi)} - \frac{3(m\pi)^3\tau}{8\gamma} \right]
\end{aligned}$$

using (C.29) and (C.34) (restoring the $\sum_j$ sign). This completes the proof. $\square$

Acknowledgments. We thank Professors R. Kohn and C. Muratov for conversations on the subject of block copolymers during the 50th annual SIAM conference in Philadelphia. Thanks also go to Professor D. Sattinger for discussion on the bifurcation theory. The support from The Institute of Mathematical Sciences at The Chinese University of Hong Kong is much appreciated.

REFERENCES

[1] D. Andelman, F. Broçhard, and J.-F. Joanny, *Phase transitions in Langmuir monolayers of polar molecules*, J. Chem. Phys., 86 (1987), pp. 3673–3681.

[2] R. Choksi, *Scaling laws in microphase separation of diblock copolymers*, J. Nonlinear Sci., 11 (2001), pp. 223–236.

[3] R. Choksi and X. Ren, *On the derivation of a density functional theory for microphase separation of diblock copolymers*, J. Statist. Phys., 113 (2003), pp. 151–176.

[4] M. G. Crandall and P. H. Rabinowitz, *Bifurcation, perturbation of simple eigenvalues, and linearized stability*, Arch. Rational Mech. Anal., 52 (1973), pp. 161–180.

[5] P. C. Fife and D. Hilhorst, *The Nishiura–Ohnishi free boundary problem in the* 1*D case*, SIAM J. Math. Anal., 33 (2001), pp. 589–606.

[6] M. Henry, *Singular limit of a fourth order problem arising in the micro-phase separation of diblock copolymers*, Adv. Differential Equations, 6 (2001), pp. 1049–1114.

[7] C. B. Muratov, *Theory of domain patterns in systems with long-range interactions of Coulomb type*, Phys. Rev. E, 66 (2002), 066108.

[8] Y. Nishiura, *Coexistence of infinitely many stable solutions to reaction-diffusion systems in the singular limit*, in Dynamics Reported: Expositions in Dynamical Systems, Vol. 3, C. R. K. T. Jones, U. Kirchgraber, and H. O. Walther, eds., Springer-Verlag, New York, 1994.

[9] Y. Nishiura and I. Ohnishi, *Some mathematical aspects of the microphase separation in diblock copolymers*, Phys. D, 84 (1995), pp. 31–39.

[10] I. Ohnishi, Y. Nishiura, M. Imai, and Y. Matsushita, *Analytical solutions describing the phase separation driven by a free energy functional containing a long-range interaction term*, Chaos, 9 (1999), pp. 329–341.

[11] T. Ohta and K. Kawasaki, *Equilibrium morphology of block copolymer melts*, Macromolecules, 19 (1986), pp. 2621–2632.

[12] X. Ren and J. Wei, *On the multiplicity of solutions of two nonlocal variational problems*, SIAM J. Math. Anal., 31 (2000), pp. 909–924.

[13] X. Ren and J. Wei, *Concentrically layered energy equilibria of the di-block copolymer problem*, European J. Appl. Math., 13 (2002), pp. 479–496.

[14] X. Ren and J. Wei, *On energy minimizers of the di-block copolymer problem*, Interfaces Free Bound., 5 (2003), pp. 193–238.

[15] X. Ren and J. Wei, *On the spectra of three-dimensional lamellar solutions of the diblock copolymer problem*, SIAM J. Math. Anal., 35 (2003), pp. 1–32.

[16] X. Ren and J. Wei, *Soliton-stripe patterns in charged Langmuir monolayers*, J. Nonlinear Sci., 13 (2003), pp. 603–624.

[17] X. Ren and J. Wei, *Triblock copolymer theory: Free energy, disordered phase and weak segregation*, Phys. D, 178 (2003), pp. 103–117.

[18] X. Ren and J. Wei, *Triblock copolymer theory: Ordered ABC lamellar phase*, J. Nonlinear Sci., 13 (2003), pp. 175–208.

[19] X. Ren and J. Wei, *The soliton-stripe pattern in the Seul–Andelman membrane*, Phys. D, 188 (2004), pp. 277–291.

[20] X. Ren and J. Wei, *Stability of spot and ring solutions of the diblock copolymer equation*, J. Math. Phys., 45 (2004), pp. 4106–4133.

[21] X. Ren and J. Wei, *Droplet solutions in the diblock copolymer problem with skewed monomer composition*, Calc. Var. Partial Differential Equations, to appear.

[22] X. Ren and J. Wei, *Sphere and Spherical Lamellar Patterns in the Ohta-Kawasaki Theory of Diblock Copolymer Melts*, preprint.

[23] D. H. Sattinger, *Stability of bifurcating solutions by Leray-Schauder degree*, Arch. Rational Mech. Anal., 43 (1971), pp. 154–166.

[24] J. V. Selinger, Z.-G. Wang, R. F. Bruinsma, and C. M. Knobler, *Chiral symmetry breaking in Langmuir monolayers and smectic films*, Phys. Rev. Lett., 70 (1993), pp. 1139–1142.

[25] M. Seul and D. Andelman, *Domain shapes and patterns: The phenomenology of modulated phases*, Science, 267 (1995), pp. 476–483.

[26] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.

[27] Y. Tsori, D. Andelman, and M. Schick, *Defects in lamellar diblock copolymers: Chevron- and Ω-shaped tilt boundaries*, Phys. Rev. E, 61 (2000), pp. 2848–2858.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 490–513

LARGE SOLUTIONS FOR A SYSTEM OF ELLIPTIC EQUATIONS ARISING FROM FLUID DYNAMICS*

J. I. DÍAZ†, M. LAZZO‡, AND P. G. SCHMIDT§

Abstract. This paper is concerned with the elliptic system

$$\Delta v = \phi, \qquad \Delta \phi = |\nabla v|^2 \tag{0.1}$$

posed in a bounded domain $\Omega \subset \mathbb{R}^N$, $N \in \mathbb{N}$. Specifically, we are interested in the existence and uniqueness or multiplicity of "large solutions," that is, classical solutions of (0.1) that approach infinity at the boundary of Ω. Assuming that Ω is a ball, we prove that the system (0.1) has a unique radially symmetric and nonnegative large solution with $v(0) = 0$ (obviously, v is determined only up to an additive constant). Moreover, if the space dimension N is sufficiently small, there exists exactly one additional radially symmetric large solution with $v(0) = 0$ (which, of course, fails to be nonnegative). We also study the asymptotic behavior of these solutions near the boundary of Ω and determine the exact blow-up rates; those are the same for all radial large solutions and independent of the space dimension. Our investigation is motivated by a problem in fluid dynamics. Under certain assumptions, the unidirectional flow of a viscous, heat-conducting fluid is governed by a pair of parabolic equations of the form

$$v_t - \Delta v = \theta, \qquad \theta_t - \Delta \theta = |\nabla v|^2, \tag{0.2}$$

where v and θ represent the fluid velocity and temperature, respectively. The system (0.1), with $\phi = -\theta$, is the stationary version of (0.2).

Key words. elliptic system, boundary blow-up, large solutions, radial solutions, existence and multiplicity, asymptotic behavior

AMS subject classifications. 35J60, 35J55, 35Q35

DOI. 10.1137/S0036141004443555

1. Introduction and main results. This paper is a contribution to the study of "explosive behavior" in certain systems of elliptic and parabolic PDEs. Our investigation is motivated by a question regarding the dynamics of a viscous, heat-conducting fluid.

In general, the flow of such a fluid is governed by a system of balance equations for momentum, mass, and energy. Under the assumptions of the so-called Boussinesq approximation, this system reduces to the Navier–Stokes equations for an incompressible fluid, along with a heat equation; the equations are nonlinearly coupled through the buoyancy force and viscous heating. If viscous heating (that is, the production of heat due to internal friction) is neglected, the resulting boundary and initial-boundary value problems are well posed in the same sense as for the classical Navier–Stokes equations without thermal coupling; but if viscous heating is taken into account, well-posedness is an open question. In fact, we conjecture that the solutions,

*Received by the editors May 6, 2004; accepted for publication (in revised form) January 19, 2005; published electronically October 14, 2005.
http://www.siam.org/journals/sima/37-2/44355.html

†Departamento de Matemática Aplicada, Universidad Complutense de Madrid, Avenida Complutense s/n, 28040 Madrid, Spain (ji_diaz@mat.ucm.es). The research of this author was partially supported by project REN2003-0223-C03 of the DGISGPI (Spain) and project RTN HPRN-CT-2002-00274 of the European Union.

‡Dipartimento di Matematica, Università degli Studi di Bari, via Orabona 4, 70125 Bari, Italy (lazzo@dm.uniba.it).

§Department of Mathematics and Statistics, Parker Hall, Auburn University, AL 36849–5310 (pgs@cam.auburn.edu).

in this case, may exhibit "explosive behavior." Such behavior would have implications for the viability of the Boussinesq approximation in situations where viscous heating cannot be neglected.

To address this issue, we are studying a simple prototype problem, which can be physically justified by considering a unidirectional flow, independent of distance in the flow direction:

$$v_t - \Delta v = \theta, \qquad \theta_t - \Delta\theta = |\nabla v|^2. \tag{1.1}$$

Here, v (the velocity) and θ (the temperature) are scalar functions of time t and position x; the spatial variable x varies over a bounded domain $\Omega \subset \mathbb{R}^N$ with $N \in \mathbb{N}$ ($N = 2$ in the physically relevant case, where Ω is the cross-section of the flow channel). The source terms θ and $|\nabla v|^2$ represent the buoyancy force and viscous heating, respectively. The system (1.1) must be supplemented by suitable initial conditions at time $t = 0$ and boundary conditions on the boundary $\partial\Omega$ of the domain Ω (for example, a homogeneous Dirichlet condition for v and a homogeneous Neumann condition for θ if the walls of the flow channel are impermeable and thermally insulated).

Note that we cannot hope to find weak solutions of the resulting initial-boundary value problem in the usual Hilbert-space setting: if v takes values in $H^1(\Omega)$, then the right-hand side of the second equation in (1.1) maps, a priori, only into $L^1(\Omega)$. However, local-in-time existence and uniqueness of a strong solution can be established by means of semigroup theory in a suitable L^p-space setting. We conjecture that this solution may blow up in finite time, in the sense that a suitable norm of (v, θ) approaches infinity as $t \to T^-$, for some $T > 0$. Preliminary analytical and numerical results for the parabolic problem will appear in a forthcoming publication.

In the present paper, we consider the stationary version of (1.1), that is, the elliptic system

$$-\Delta v = \theta, \qquad -\Delta\theta = |\nabla v|^2 \tag{1.2}$$

posed in a domain $\Omega \subset \mathbb{R}^N$ with $N \in \mathbb{N}$. Specifically, we are interested in the possibility of "boundary blow-up," that is, the existence of classical solutions (v, θ) of (1.2) with $|(v(x), \theta(x))| \to \infty$ as $\operatorname{dist}(x, \partial\Omega) \to 0$ (so-called large solutions). Note that the θ-component of any solution of (1.2) is superharmonic in Ω and thus cannot approach ∞ at the boundary (maximum principle); for a similar reason, v and θ cannot simultaneously approach $-\infty$ at the boundary. We therefore expect any large solution (v, θ) of (1.2) to satisfy $v(x) \to \infty$ and $\theta(x) \to -\infty$ as $\operatorname{dist}(x, \partial\Omega) \to 0$.

The preceding observation implies that large solutions of (1.2) cannot be expected to describe the asymptotics of explosive solutions of the parabolic system (1.1). Assuming, for example, that the temperature θ in (1.1) satisfies a homogeneous Neumann boundary condition on $\partial\Omega$, the temperature minimum is a nondecreasing function of time (parabolic maximum principle); thus, θ cannot approach $-\infty$ at the boundary. Nevertheless, boundary blow-up in the elliptic system (1.2) would have implications for the dynamics of the parabolic system (1.1) and its controllability. For example, large solutions of (1.2) may be used to construct "universal distributed bounds" (that is, interior bounds independent of the boundary data) for solutions of associated initial-boundary value problems and their steady states. We refer the reader to [3, 6, 7, 14, 23] and the references therein for similar arguments and applications in the context of other semilinear or quasilinear parabolic problems with superlinear nonlinearities.

Henceforth, we assume that Ω is a ball in $\mathbb{R}^N$, centered at the origin; that is, $\Omega = B_R^N(0)$ for some $R > 0$. For convenience, we introduce the function $\phi = -\theta$ and seek radially symmetric large solutions of the problem

$$\Delta v = \phi, \qquad \Delta\phi = |\nabla v|^2 \qquad \text{in } B_R^N(0), \tag{1.3}$$

that is, radial solutions (v, ϕ) with $|(v(x), \phi(x))| \to \infty$ as $|x| \to R^-$.

Remark 1.1. The problem (1.3) has a scaling property that we will exploit repeatedly. Suppose (v_1, ϕ_1) is a (large) solution of (1.3) with $R = R_1$. For $\lambda \in (0, \infty)$, let $R_\lambda := \lambda^{-1} R_1$. For $x \in B_{R_\lambda}^N(0)$, define

$$v_\lambda(x) := \lambda^2 v_1(\lambda x)\,, \quad \phi_\lambda(x) := \lambda^4 \phi_1(\lambda x)\,.$$

Then $(v_\lambda, \phi_\lambda)$ is a (large) solution of (1.3) with $R = R_\lambda$.

Remark 1.2. If (v, ϕ) is a (large) solution of (1.3), then so is $(v + c, \phi)$, for any constant $c \in \mathbb{R}$. Thus, we may restrict attention to solutions with $v(0) = 0$.

We will now state our main results, the first of which guarantees the existence of a unique (up to a shift in v) radially symmetric and nonnegative large solution for any space dimension.

Theorem 1.3. *For every $N \in \mathbb{N}$ and $R > 0$, the problem* (1.3) *has a unique radially symmetric large solution (v, ϕ) with $v(0) = 0$ and $\phi(0) > 0$. Both components of this solution are increasing functions of the radial variable r.*

If the space dimension is sufficiently small, there exists exactly one additional radially symmetric large solution with $v(0) = 0$, which, of course, fails to be nonnegative.

Theorem 1.4. *For every $N \in \mathbb{N}$ with $N \leq 10$ and every $R > 0$, the problem* (1.3) *has a unique radially symmetric large solution (v, ϕ) with $v(0) = 0$ and $\phi(0) < 0$. The ϕ-component of this solution is an increasing function of the radial variable r, while the v-component is decreasing to a negative minimum and increasing thereafter.*

Let us note that the bound on N in the above result is not sharp. In fact, based on numerical evidence (see Remarks 3.5 and 4.4), we conjecture that the solution of Theorem 1.4 exists if and only if $N \leq 14$.

With regard to asymptotic behavior, we find that, as expected, both components of a large solution approach infinity at the boundary, and we determine the exact blow-up rates; those are the same for all radially symmetric large solutions and independent of the space dimension. Here and in what follows, we write $f(x) \sim g(x)$ if the mappings $f, g : B_R^N(0) \to \mathbb{R}$ satisfy $f(x)/g(x) \to 1$ as $|x| \to R^-$.

Theorem 1.5. *Let (v, ϕ) be any radially symmetric large solution of* (1.3)*, for a given $N \in \mathbb{N}$ and $R > 0$. Then, as $|x| \to R^-$,*

$$v(x) \sim \frac{30}{(R - |x|)^2} \quad \textit{and} \quad \phi(x) \sim \frac{180}{(R - |x|)^4}\,.$$

The study of "explosive behavior," be it finite-time blow-up in evolutionary problems or boundary blow-up in stationary problems, has a long history going back to seminal work by Keller [15] and Osserman [20] in the 1950s; we refer the reader to the papers [2, 4, 8, 24] and the references therein. However, virtually all of the existing literature is concerned with scalar equations. Coupled systems of equations have been attacked only recently; see, for example, [5, 9, 10, 11, 16]. Due to the lack of variational structure and comparison principles, methods that have proven successful for scalar equations will, in general, fail to be useful for systems, even if the expected results are analogous. For example, our existence and multiplicity result for the problem (1.3) (existence of one large nonnegative solution for any space dimension, existence of a second large solution for sufficiently small space dimension) is analogous to a result by McKenna, Reichel, and Walter [17] for a class of scalar equations with variational structure. However, our method of proof is entirely different, and our result appears to be the first of its kind for an elliptic system. We expect that our work, while currently focussed on a very specific problem, will lead to general insights and new methods with potential applications to a much wider class of elliptic and parabolic systems.

The rest of the paper is organized as follows. In section 2 we reduce our problem to the study of a system of first-order ODEs, establish some basic properties of its solutions, and prove the existence and uniqueness of a nonnegative large radial solution for the problem (1.3); Theorem 1.3 is an immediate consequence of Proposition 2.5. Section 3 is devoted to the proof of Theorem 1.4 (existence of a second large radial solution for sufficiently small space dimension), which follows from Proposition 3.1. This section also includes a discussion of numerical experiments, suggesting a sharper version of Theorem 1.4, and observations about a related parameter-dependent fixed-point equation, leading to a Liouville-type theorem for the Dirichlet problem associated with the elliptic system (1.2). In section 4 we analyze the asymptotic behavior of large radial solutions of (1.3); Theorem 1.5 follows from Proposition 4.1, whose proof relies on dynamical-systems theory applied to an asymptotically autonomous and cooperative ODE system in $\mathbb{R}^3$. In an appendix at the end of the paper, we describe a Maple algorithm for the computer-aided construction of a priori bounds needed in the proof of Proposition 3.1.

2. Preliminaries and nonnegative large solutions. Given $N \in \mathbb{N}$ and $R > 0$, radially symmetric solutions of the problem (1.3) correspond to solutions of the ODE system

$$\begin{cases} v'' + \dfrac{N-1}{r}\, v' = \phi, \\ \phi'' + \dfrac{N-1}{r}\, \phi' = |v'|^2 \end{cases} \qquad \text{in } (0,R)$$

with $v'(0) = \phi'(0) = 0$; large solutions are those with $|(v(r),\phi(r))| \to \infty$ as $r \to R^-$. In view of Remark 1.2, we may impose the initial condition $v(0) = 0$. Finding radially symmetric large solutions of the problem (1.3) is therefore equivalent to finding initial conditions $\phi(0) = p$ such that the solution of the Cauchy problem

$$\begin{cases} v'' + \dfrac{N-1}{r}\, v' = \phi, & v(0) = 0,\ v'(0) = 0, \\ \phi'' + \dfrac{N-1}{r}\, \phi' = |v'|^2, & \phi(0) = p,\ \phi'(0) = 0 \end{cases} \tag{2.1}$$

exists on the interval $[0, R)$ and "blows up" at R.

Despite the singularity at $r = 0$ for $N > 1$, the Cauchy problem (2.1) is well posed. Indeed, for every $p \in \mathbb{R}$, there exists a unique maximal solution, which depends continuously on p (in the usual sense); see Lemma 2.3 for details.

Remark 2.1. The scaling property of the elliptic problem (1.3), as described in Remark 1.1, and the well-posedness of (2.1) imply that all solutions of the Cauchy problem with $p > 0$ ($p < 0$) are "rescalings" of the solution with $p = 1$ ($p = -1$). Indeed, if (v_1, ϕ_1) is the maximal solution with initial value $p = 1$ ($p = -1$), then the maximal solution with initial value $p > 0$ ($p < 0$) is given by $(v_\lambda, \phi_\lambda)$, as defined in Remark 1.1, with $\lambda = |p|^{1/4}$. Consequently, if the maximal solution with initial value $p = 1$ ($p = -1$) blows up at R_1, then the maximal solution with initial value $p > 0$ ($p < 0$) blows up at $R_p = |p|^{-1/4} R_1$.

REMARK 2.2. In light of the preceding remark, it is clear that the elliptic problem (1.3) has large radial solutions, for any given $R > 0$, if and only if the solutions of the Cauchy problem (2.1) with $p = \pm 1$ exhibit finite-time blow-up. More precisely, (1.3) has exactly one large radial solution with $v(0) = 0$ and $\phi(0) > 0$ if and only if the solution of (2.1) with $p = 1$ blows up in finite time; (1.3) has exactly one large radial solution with $v(0) = 0$ and $\phi(0) < 0$ if and only if the solution of (2.1) with $p = -1$ blows up in finite time. In particular, (1.3) cannot have more than two large radial solutions.

The Cauchy problem (2.1) is equivalent to the first-order system

$$\begin{cases} v' = w, & v(0) = 0, \\ w' + \dfrac{N-1}{r}\, w = \phi, & w(0) = 0, \\ \phi' = \psi, & \phi(0) = p, \\ \psi' + \dfrac{N-1}{r}\, \psi = w^2, & \psi(0) = 0. \end{cases}$$

Obviously, we can eliminate v and drop the first equation and initial condition; v is recovered from w via antidifferentiation. Furthermore, we may replace the nonnegative integer $N - 1$ with a continuous parameter $\mu \in \mathbb{R}_+$. Thus, we are led to the Cauchy problem

$$\begin{cases} w' + \dfrac{\mu}{r}\, w = \phi, & w(0) = 0, \\ \phi' = \psi, & \phi(0) = p, \\ \psi' + \dfrac{\mu}{r}\, \psi = w^2, & \psi(0) = 0. \end{cases} \tag{2.2}$$

LEMMA 2.3. *For every $\mu \in \mathbb{R}_+$ and $p \in \mathbb{R}$, the Cauchy problem* (2.2) *has a unique maximal, that is, noncontinuable, solution $(w, \phi, \psi) \in C^1([0, R), \mathbb{R}^3)$, for some $R \in (0, \infty]$. If $R < \infty$, then $|(w(r), \phi(r), \psi(r))| \to \infty$ as $r \to R^-$. Moreover, (w, ϕ, ψ) depends continuously on μ and p.*

Proof. What we claim is that, despite the singularity at $r = 0$ in the case $\mu > 0$, the Cauchy problem (2.2) has the usual, well-known properties of a regular initial-value problem in $\mathbb{R}^3$. Since we could not find a general result in the literature that would cover our problem, we provide a few remarks on the proof.

Note that the first equation in (2.2) can be written as $(r^\mu w)' = r^\mu \phi$. Together with the initial condition $w(0) = 0$, this is equivalent to the integral equation

$$w(r) = \int_0^r \left(\frac{s}{r}\right)^\mu \phi(s)\, ds. \tag{2.3}$$

Similarly, the remaining differential equations and initial conditions in (2.2) are equivalent to the integral equations

$$\phi(r) = p + \int_0^r \psi(s)\, ds \tag{2.4}$$

and

$$\psi(r) = \int_0^r \left(\frac{s}{r}\right)^\mu w^2(s)\, ds. \tag{2.5}$$

Since we have $0 < s/r < 1$ for $0 < s < r$, the "singular term" $(s/r)^\mu$ does not cause any difficulties in proving the existence and uniqueness of a solution (w, ϕ, ψ) in $C([0, \varepsilon], \mathbb{R}^3)$ of (2.3)–(2.5), for some $\varepsilon > 0$, by means of the contraction mapping principle. Clearly, w, ϕ, and ψ are continuously differentiable on $(0, \varepsilon]$ and satisfy the differential equations and initial conditions in (2.2). In fact, all three components are continuously differentiable on the *closed* interval $[0, \varepsilon]$. This is obvious for ϕ, but less so for w and ψ. Note, however, that

$$\begin{aligned}
\lim_{r\to 0^+} w'(r) &= \lim_{r\to 0^+} \Big(\phi(r) - \frac{\mu}{r}\, w(r)\Big) = p - \mu \lim_{r\to 0^+} \frac{1}{r}\int_0^r \Big(\frac{s}{r}\Big)^\mu \phi(s)\, ds \\
&= p - \mu \lim_{r\to 0^+} \frac{1}{r^{\mu+1}} \int_0^r s^\mu \phi(s)\, ds = p - \mu \lim_{r\to 0^+} \frac{r^\mu \phi(r)}{(\mu+1)\, r^\mu} \\
&= p - \mu \lim_{r\to 0^+} \frac{\phi(r)}{\mu+1} = p - \mu \frac{p}{\mu+1} = \frac{p}{\mu+1}\,,
\end{aligned}$$

where we used l'Hôspital's rule to get the fourth equality. Thus, $w \in C^1([0, \varepsilon], \mathbb{R})$ and $w'(0) = p/(\mu+1)$. Similarly, one shows that $\psi \in C^1([0, \varepsilon], \mathbb{R})$ with $\psi'(0) = 0$.

Once existence and uniqueness of a local C^1-solution are established, the remaining claims about maximal continuation and continuous dependence on parameters and initial data can be proved in the same way as for regular initial-value problems. □

LEMMA 2.4. *Let $(w, \phi, \psi) \in C^1([0, R), \mathbb{R}^3)$ be the maximal solution of the Cauchy problem* (2.2), *for some $\mu \in \mathbb{R}_+$ and $p \in \mathbb{R}$ with $p \neq 0$. Then the function ϕ is strictly increasing on $[0, R)$, and $L := \lim_{r\to R^-} \phi(r)$ is either zero or infinity. In fact,*

(a) *if $L < \infty$, then $R = \infty$ and $L = 0$;*

(b) *if $L = \infty$, then $R < \infty$ and $w(R^-) = \phi(R^-) = \psi(R^-) = \infty$.*

Proof. Taking into account the equations and initial conditions in (2.2), it is easy to see that the function $\Psi(r) := r^\mu \psi(r)$ is strictly increasing on $[0, R)$. As a consequence, Ψ (and thus ψ) is positive on $(0, R)$, and this implies that ϕ is strictly increasing on $[0, R)$, with $L := \lim_{r\to R^-} \phi(r) \in (p, \infty]$.

(a) Assume $L < \infty$, that is, ϕ is bounded. By (2.3), $w(r)$ grows at most linearly with r, and, by (2.5), $\psi(r)$ grows no faster than r^3. In particular, $|(w(r), \phi(r), \psi(r))|$ cannot go to infinity in finite time. Thus, $R = \infty$.

Now suppose that $L \neq 0$. If $L > 0$, choose a number $r_0 > 0$ such that $\phi(r) \geq L/2$ for every $r \geq r_0$. It follows that

$$w(r) \geq \int_0^{r_0} \Big(\frac{s}{r}\Big)^\mu \phi(s)\, ds + \frac{L}{2}\int_{r_0}^r \Big(\frac{s}{r}\Big)^\mu ds$$

for every $r \geq r_0$, and we conclude that $\lim_{r\to\infty} w(r) = \infty$ (note that the last integral is of order r). If $L < 0$, we infer in a similar way that $\lim_{r\to\infty} w(r) = -\infty$. In any case, we can choose a number $r_1 > 0$ such that $w^2(r) \geq 1$ for every $r \geq r_1$. As a consequence,

$$\psi(r) \geq \int_0^{r_1} \Big(\frac{s}{r}\Big)^\mu w^2(s)\, ds + \int_{r_1}^r \Big(\frac{s}{r}\Big)^\mu ds$$

for every $r \geq r_1$, and thus $\lim_{r\to\infty} \psi(r) = \infty$. But this implies

$$L = \lim_{r\to\infty} \phi(r) = p + \lim_{r\to\infty} \int_0^r \psi(s)\, ds = \infty\,,$$

a contradiction. It follows that $L = 0$.

(b) Assume $L = \infty$ and, by way of contradiction, suppose that $R = \infty$. Then there exist c_0, $r_0 > 0$ such that $\phi(r) \geq c_0$ for every $r \geq r_0$, and as in the proof of part (a) it follows that $w(r) \to \infty$ and $\psi(r) \to \infty$ as $r \to \infty$. In particular, we can choose $r_* > 0$ such that $w(r)$, $\phi(r)$, $\psi(r) > 0$ for every $r \geq r_*$. Define $\eta := w\,\phi\,\psi$. Then we have

$$\begin{aligned}\eta' &= \phi^2\psi + w\psi^2 + w^3\phi - \frac{2\mu}{r}\,w\,\phi\,\psi \\ &= Q\big(w,\phi,\psi\big)\,\eta^{13/12} - \frac{2\mu}{r}\,\eta \qquad \text{in } [r_*,\infty),\end{aligned} \tag{2.6}$$

where Q is defined by

$$Q(x,y,z) := \frac{y^2z + xz^2 + x^3y}{(xyz)^{13/12}}\,,$$

for x, y, $z > 0$. Note that $Q = Q_1 + Q_2 + Q_3$, with

$$Q_1 := \frac{y^2z}{(xyz)^{13/12}}\,, \quad Q_2 := \frac{xz^2}{(xyz)^{13/12}}\,, \quad Q_3 := \frac{x^3y}{(xyz)^{13/12}}\,.$$

It is easy to see that $Q_1^5\,Q_2^4\,Q_3^3 \equiv 1$, which implies that $\max(Q_1, Q_2, Q_3) \geq 1$. Hence, we have $Q(x,y,z) \geq 1$ for all x, y, $z > 0$, and (2.6) yields

$$\eta' \geq \eta\left(\eta^{1/12} - \frac{2\mu}{r_*}\right) \qquad \text{in } [r_*,\infty). \tag{2.7}$$

Recall that $w(r)$, $\phi(r)$, $\psi(r) \to \infty$ as $r \to \infty$ and choose $r^* \geq r_*$ such that $\eta(r^*) > (2\mu/r_*)^{12}$. Then the maximal solution ζ of the initial-value problem

$$\zeta' = \zeta\Big(\zeta^{1/12} - \frac{2\mu}{r_*}\Big)\,, \qquad \zeta(r^*) = \eta(r^*)$$

approaches infinity in finite time. But due to (2.7), ζ is bounded from above by η on $[r^*,\infty)$. This is a contradiction, and it follows that R is finite.

In order to prove our last claim, we first note that both w and ψ have (proper or improper) limits as $r \to R^-$. Indeed, since ϕ is eventually positive, the function $W(r) := r^\mu w(r)$ is eventually increasing, and thus has a limit as $r \to R^-$. As we observed earlier, the same holds for the function $\Psi(r) := r^\mu\psi(r)$. Since R is finite, it follows that $w(r)$ and $\psi(r)$, too, have limits as $r \to R^-$. Moreover, since R is finite, all three of the functions w, $\phi\,,\psi$ would be bounded if one of them were. But ϕ is unbounded (by assumption) and thus w and ψ are unbounded as well. Clearly, this implies that $w(R^-) = \phi(R^-) = \psi(R^-) = \infty$. ∎

PROPOSITION 2.5. *For every $\mu \in \mathbb{R}_+$, the maximal solution of the Cauchy problem* (2.2) *with $p = 1$ blows up in finite time.*

Proof. Fix $\mu \in \mathbb{R}_+$ and let $(w,\phi,\psi) \in C^1([0,R),\mathbb{R}^3)$ be the maximal solution of (2.2) with $p = 1$. According to Lemma 2.4, ϕ is increasing and $L := \lim_{r\to R^-}\phi(r)$ is either zero or infinity. Since $\phi(0) > 0$, we have $L = \infty$, and then part (b) of the same lemma implies that R is finite. ∎

Proof of Theorem 1.3. Thanks to Remark 2.2, the preceding proposition guarantees that the problem (1.3), for arbitrary $N \in \mathbb{N}$ and $R > 0$, has exactly one large radial solution (v, ϕ) with $v(0) = 0$ and $\phi(0) > 0$. By Lemma 2.4, ϕ is a strictly increasing function of the radial variable r, and the same then holds for v. (Note that, by Lemma 2.4, $\phi(r)$ approaches infinity as $r \to R^-$, and so do $v'(r)$ and $\phi'(r)$. That the same holds for $v(r)$ is not obvious at this point, but will follow from the blow-up estimates in section 4.) ∎

Figure 1 shows computed profiles of the nonnegative large radial solutions (v, ϕ) of the problem (1.3), with $R = 1$, for two values of the space dimension N (see Remark 4.4 for comments regarding the numerical method).

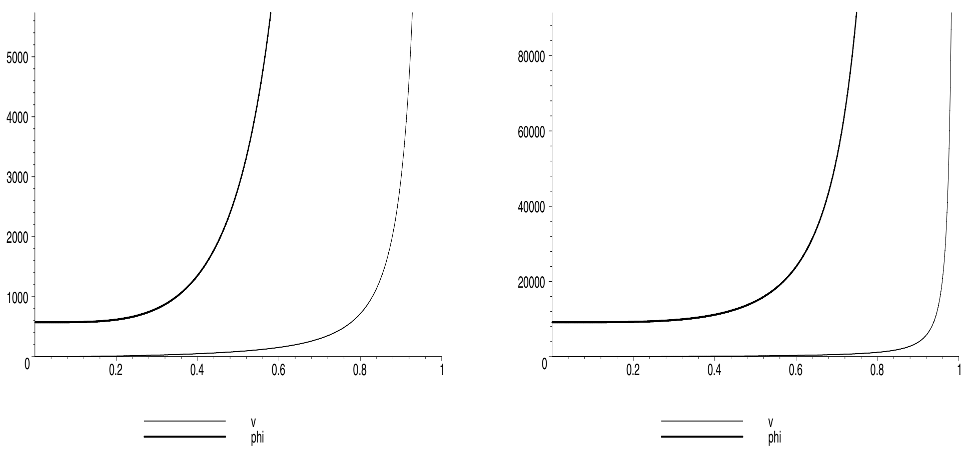


FIG. 1. *Large radial solutions* (v, ϕ) *with* $v(0) = 0$ *and* $\phi(0) > 0$ *of problem* (1.3) *with* $R = 1$ *for* $N = 1$ (*left*) *and* $N = 6$ (*right*).

3. Existence of a second large solution. To prove Theorem 1.4, we need to investigate for which values of $\mu \in \mathbb{R}_+$ (if any) the maximal solution of the Cauchy problem (2.2) with $p = -1$ blows up in finite time.

It is easy to see that blow-up occurs at least if $\mu \in [0, 1]$. Indeed, suppose that $\mu \in \mathbb{R}_+$ and that the corresponding maximal solution (w, ϕ, ψ) of (2.2) with $p = -1$ exists globally. Lemma 2.4 then implies that $\phi(r) \to 0$ as $r \to \infty$; thus, $\int_0^\infty \psi(s)\,ds = 1$, due to (2.4). On the other hand, since $\Psi(r) := r^\mu \psi(r)$ is strictly increasing for $r \geq 0$, we have $c := \Psi(1) > 0$ and $\psi(r) = \Psi(r)\,r^{-\mu} \geq c\,r^{-\mu}$ for all $r \geq 1$, which implies $\int_0^\infty \psi(s)\,ds \geq c \int_1^\infty s^{-\mu}\,ds$. Unless μ is greater than 1, this shows that $\int_0^\infty \psi(s)\,ds = \infty$, and we arrive at a contradiction.

Recalling Remark 2.2, we conclude that the elliptic problem (1.3) has a second large radial solution at least if the space dimension N is 1 or 2. The following proposition allows us to draw the same conclusion for any space dimension up to and including 10.

PROPOSITION 3.1. *For every* $\mu \in [0, 9]$, *the maximal solution of the Cauchy problem* (2.2) *with* $p = -1$ *blows up in finite time.*

Proof. Fix $\mu \in \mathbb{R}_+$ and let $(w, \phi, \psi) \in C^1([0, R), \mathbb{R}^3)$ be the corresponding maximal solution of (2.2) with $p = -1$. By Lemma 2.4, R is finite if and only if ϕ is eventually positive. We will prove the proposition by constructing explicit lower bounds for ϕ that are eventually positive if μ is small enough.

Clearly, since ϕ is an increasing function, we have

$$\phi(r) \geq \phi(0) = -1 =: \underline{\phi}_0(r)$$

for all $r \in [0, R)$. From this, we derive a lower bound for w; indeed,

$$w(r) = \int_0^r \left(\frac{s}{r}\right)^\mu \phi(s)\, ds \geq \int_0^r \left(\frac{s}{r}\right)^\mu \underline{\phi}_0(s)\, ds = -\frac{r}{\mu+1} =: \underline{w}_1(r)$$

for all $r \in [0, R)$. As long as $w \leq 0$, a lower bound for w yields an upper bound for ψ; in particular,

$$\psi(r) = \int_0^r \left(\frac{s}{r}\right)^\mu w^2(s)\, ds \leq \int_0^r \left(\frac{s}{r}\right)^\mu \underline{w}_1^2(s)\, ds = \frac{r^3}{(\mu+1)^2(\mu+3)} =: \overline{\psi}_1(r)$$

for all $r \in [0, R)$ with $w(r) \leq 0$ (note that $w(r) \leq 0$ implies $\underline{w}_1 \leq w \leq 0$ on $[0, r]$). Next, we find an upper bound for ϕ, namely,

$$\phi(r) = -1 + \int_0^r \psi(s)\, ds \leq -1 + \int_0^r \overline{\psi}_1(s)\, ds = -1 + \frac{r^4}{4(\mu+1)^2(\mu+3)} =: \overline{\phi}_1(r),$$

still valid for all $r \in [0, R)$ with $w(r) \leq 0$. Continuation of this process yields an upper bound for w,

$$w(r) \leq \int_0^r \left(\frac{s}{r}\right)^\mu \overline{\phi}_1(s)\, ds = -\frac{r}{\mu+1} + \frac{r^5}{4(\mu+1)^2(\mu+3)(\mu+5)} =: \overline{w}_1(r),$$

valid for all $r \in [0, R)$ with $w(r) \leq 0$, and then a lower bound for ψ,

$$\psi(r) \geq \int_0^r \left(\frac{s}{r}\right)^\mu \overline{w}_1^2(s)\, ds =: \underline{\psi}_1(r),$$

valid for all $r \in [0, r_0]$, where $r_0 := (4(\mu+1)(\mu+3)(\mu+5))^{1/4}$ is the unique positive root of $\overline{w}_1$ (note that $r_0 < R$ and $w \leq \overline{w}_1 \leq 0$ on $[0, r_0]$). Finally, we obtain an improved lower bound for ϕ,

$$\phi(r) \geq -1 + \int_0^r \underline{\psi}_1(s)\, ds =: \underline{\phi}_1(r),$$

valid for all $r \in [0, r_0]$.

If $\underline{\phi}_1(r_0)$ were nonnegative, the same would hold for $\phi(r_0)$, and this would imply blow-up. Now, $\underline{\phi}_1(r_0)$ is easily seen to be a rational function of μ, with a unique positive root μ_1, near 3.512, and positive on the interval $[0, \mu_1)$. Hence, blow-up does occur if $\mu \leq \mu_1$.

To improve this result, we iterate the preceding estimates and construct the improved bounds $\underline{w}_2$, $\overline{\psi}_2$, $\overline{\phi}_2$, $\overline{w}_2$, $\underline{\psi}_2$, and $\underline{\phi}_2$; since $w \leq \overline{w}_2 \leq \overline{w}_1 \leq 0$ on $[0, r_0]$, these are still valid on the entire interval $[0, r_0]$. The actual construction is best done with the aid of a computer-algebra system. All the bounds being polynomials, the computations amount to symbolic operations on the coefficients and can be implemented very efficiently. (A more naïve approach, using symbolic antidifferentiation, is likely to fail.) In an appendix at the end of the paper, we describe a Maple implementation of the algorithm.

Once the improved lower bound $\underline{\phi}_2$ for ϕ is constructed, we may determine the sign of $\underline{\phi}_2(r_0)$; this is once again a rational function of μ, with a unique positive root μ_2, near 4.307, and positive on the interval $[0,\mu_2)$. We conclude that blow-up does occur if $\mu \le \mu_2$.

Any attempt to push this method further by performing another round of estimates turns out to be futile—the computational cost is prohibitive, the gain marginal (the positive root of $\underline{\phi}_3(r_0)$ is located near 4.311). Instead, we will extend the lower bound $\underline{\phi}_2$ of ϕ beyond the interval $[0,r_0]$. To this end, let $r_1 \in (r_0,R)$ be such that $\phi(r_1) \le 0$. Then we have $\phi \le \overline{\phi}_2$ on $[0,r_0]$ and $\phi \le 0$ on $[r_0,r_1]$. It follows that for all $r \in [r_0,r_1]$,

$$w(r) \le \int_0^{r_0} \left(\frac{s}{r}\right)^{\mu} \overline{\phi}_2(s)\,ds = \left(\frac{r_0}{r}\right)^{\mu} \int_0^{r_0} \left(\frac{s}{r_0}\right)^{\mu} \overline{\phi}_2(s)\,ds = \alpha \left(\frac{r_0}{r}\right)^{\mu},$$

where $\alpha := \overline{w}_2(r_0) \le 0$, and then

$$\psi(r) \ge \int_0^{r_0} \left(\frac{s}{r}\right)^{\mu} \overline{w}_2^2(s)\,ds + \int_{r_0}^{r} \left(\frac{s}{r}\right)^{\mu} \alpha^2 \left(\frac{r_0}{s}\right)^{2\mu} ds = \beta \left(\frac{r_0}{r}\right)^{\mu} - \gamma \left(\frac{r_0}{r}\right)^{2\mu-1},$$

where $\beta := \gamma + \underline{\psi}_2(r_0)$, $\gamma := r_0\alpha^2/(\mu-1)$, and we have implicitly assumed that $\mu \ne 1$ (the case $\mu = 1$ will not be needed). Finally, we see that for all $r \in [r_0,r_1]$,

$$\begin{aligned}\phi(r) &\ge -1 + \int_0^{r_0} \underline{\psi}_2(s)\,ds + \int_{r_0}^{r} \left(\beta \left(\frac{r_0}{s}\right)^{\mu} - \gamma \left(\frac{r_0}{s}\right)^{2\mu-1}\right) ds \\ &= \widetilde{\alpha} \left(\frac{r_0}{r}\right)^{2(\mu-1)} - \widetilde{\beta} \left(\frac{r_0}{r}\right)^{\mu-1} + \widetilde{\gamma}\,,\end{aligned}$$

with $\widetilde{\alpha} := (1/2)r_0\gamma/(\mu-1)$, $\widetilde{\beta} := r_0\beta/(\mu-1)$, $\widetilde{\gamma} := (1/2)r_0(2\beta-\gamma)/(\mu-1) + \underline{\phi}_2(r_0)$, and hence,

$$0 \ge (\mu-1)^2\phi(r) \ge a \left(\frac{r_0}{r}\right)^{2(\mu-1)} - b \left(\frac{r_0}{r}\right)^{\mu-1} + c\,, \tag{3.1}$$

where

$$a := (\mu-1)^2\widetilde{\alpha} = \frac{1}{2} r_0^2\, \overline{w}_2^2(r_0)\,,$$

$$b := (\mu-1)^2\widetilde{\beta} = r_0^2\, \overline{w}_2^2(r_0) + (\mu-1)r_0\underline{\psi}_2(r_0)\,,$$

$$c := (\mu-1)^2\widetilde{\gamma} = \frac{1}{2} r_0^2\, \overline{w}_2^2(r_0) + (\mu-1)r_0\underline{\psi}_2(r_0) + (\mu-1)^2\underline{\phi}_2(r_0)\,.$$

The estimate (3.1) holds for every $r \in [r_0,R)$, provided that $\phi(r) \le 0$, and for any value of μ (trivially if $\mu = 1$). The coefficients a, b, and c are rational functions of μ (note that r_0^4 is a polynomial in μ, while $r^2\overline{w}_2^2(r)$, $r\underline{\psi}_2(r)$, and $\underline{\phi}_2(r)$ are polynomials in r^4, whose coefficients are rational functions of μ); this facilitates their symbolic computation and analysis. In particular, it is easily verified (see the appendix) that c has a unique root, $\bar{\mu} \approx 9.073$, in the interval $(1,\infty)$ and is positive on $(1,\bar{\mu})$. It follows that if $\mu \in (1,\bar{\mu})$, the right-hand side of the inequality (3.1) becomes positive as $r \to \infty$. But then, the inequality cannot hold for all $r \ge r_0$, which implies that R is finite. Since this is already known to be true if $\mu \in [0,1]$, the proposition is proved. $\square$

Remark 3.2. We emphasize that the proof of Proposition 3.1, while relying heavily on the use of a computer-algebra system, does not involve any numerical techniques or floating-point arithmetic. In fact, all the computations amount to symbolic algebra on the coefficients of certain polynomials. The crucial fact that the coefficient c in the estimate (3.1), a rational function of μ, has a unique root, $\bar{\mu} \approx 9.073$, in the interval $(1, \infty)$ can be verified by applying Descartes's rule of signs (and the intermediate-value theorem) to the numerator polynomial (the roots of the denominator polynomial are negative integers). We refer the reader to the appendix for implementation details.

Remark 3.3. The estimates in the proof of Proposition 3.1 involve some deliberate choices, but are in some sense optimal. Of course, it is computationally much less expensive to use the bounds $\overline{w}_1$, $\underline{\psi}_1$, and $\underline{\phi}_1$ (instead of $\overline{w}_2$, $\underline{\psi}_2$, and $\underline{\phi}_2$) for the "tail estimate" (3.1); also, the estimate itself is simpler in this case, since the coefficient a vanishes. Again, the coefficient c has a unique root in the interval $(1, \infty)$, but it is located near 5.606, leading to a much weaker result. A slight improvement is achieved by using the polynomial bounds $\overline{w}_1$, $\underline{\psi}_1$, and $\underline{\phi}_1$ only on the smaller interval $[0, \tilde{r}_0]$ (instead of $[0, r_0]$), where $\tilde{r}_0 := (4(\mu+1)^2(\mu+3))^{1/4}$ is the unique positive root of $\overline{\phi}_1$. The coefficient c in (3.1) then has a unique positive root near 5.955. Using the improved bounds $\overline{w}_2$, $\underline{\psi}_2$, and $\underline{\phi}_2$ (as in the proof of Proposition 3.1), but on the smaller interval $[0, \tilde{r}_0]$ (instead of $[0, r_0]$), yields a coefficient c with a unique positive root near 7.709. It is natural to ask whether it would make sense to construct the bounds $\overline{w}_3$, $\underline{\psi}_3$, and $\underline{\phi}_3$ before proceeding with the "tail estimate." The answer is negative; in fact, the symbolic computations would require an enormous amount of virtual memory, without leading to a tangibly improved result (the relevant root of the coefficient c in (3.1) is located near 9.170).

Remark 3.4. The estimates in the proof of Proposition 3.1 yield explicit a priori bounds for the zero z_0 of ϕ (the ϕ-component of the maximal solution of the Cauchy problem (2.2) with $\mu \in \mathbb{R}_+$ and $p = -1$), assuming that it exists. Clearly, a lower bound is given by $\tilde{r}_0 := (4(\mu+1)^2(\mu+3))^{1/4}$, the unique positive root of $\overline{\phi}_1$. To establish an upper bound, note that either $z_0 \le r_0$, where $r_0 := (4(\mu+1)(\mu+3)(\mu+5))^{1/4}$ is the unique positive root of $\overline{w}_1$, or $z_0 > r_0$. In the latter case, which can arise only if $\mu > \mu_2 \approx 4.307$, (3.1) implies that

$$a \left(\frac{r_0}{z_0}\right)^{2(\mu-1)} - b \left(\frac{r_0}{z_0}\right)^{\mu-1} + c \;\le\; 0$$

and thus

$$s_1 \;\le\; \left(\frac{r_0}{z_0}\right)^{\mu-1} \;\le\; s_2\,,$$

where $s_{1,2} := (b \pm \sqrt{d})/(2a)$ with $d := b^2 - 4ac$ (the discriminant d is positive for $\mu \ne 1$, zero for $\mu = 1$). As long as $\mu > 1$ and s_1 is positive (which is the case for $1 < \mu < \bar{\mu} \approx 9.073$), it follows that $z_0 \le r_0 s_1^{1/(1-\mu)}$. Hence, an upper bound for z_0 is given by r_0 if $0 \le \mu \le \mu_2$ and by $r_0 s_1^{1/(1-\mu)}$ if $\mu_2 < \mu < \bar{\mu}$. A little computation shows that this upper bound may be written, more concisely yet equivalently, as $r_0 \max(1, s_0^{1/(1-\mu)})$, where $s_0 := (b - \operatorname{sign}(\mu-1)\sqrt{d})/(2a)$. Summarizing, we have

$$\tilde{r}_0 \le z_0 \le r_0 \max(1, s_0^{1/(1-\mu)})$$

for every $\mu \in [0, \bar{\mu})$; the upper bound is a continuous function of μ, positive on $[0, \bar{\mu})$, with a vertical asymptote at $\bar{\mu}$.

Proof of Theorem 1.4. Due to Remark 2.2, Proposition 3.1 implies that the elliptic problem (1.3), for arbitrary $R > 0$, has exactly one large radial solution (v, ϕ) with $v(0) = 0$ and $\phi(0) < 0$ if $\mu = N - 1 \leq 9$, that is, if $N \leq 10$. Lemma 2.4 shows that ϕ is a strictly increasing function of the radial variable r and crosses zero at a point $z_0 \in (0, R)$. Hence, the function $W(r) := r^\mu w(r)$, with $w = v'$, is strictly decreasing for $r < z_0$ and strictly increasing for $r > z_0$; it crosses zero at a point $z_1 \in (z_0, R)$. Thus, v' is negative on $(0, z_1)$, positive on (z_1, R), and consequently, v is strictly decreasing to a negative minimum at z_1, strictly increasing thereafter. ∎

Remark 3.5. Numerical evidence suggests that Proposition 3.1 (and with it, Theorem 1.4) may be significantly improved. In fact, there appears to be a number $\tilde{\mu}$, approximately equal to 13.755, such that the maximal solution of the Cauchy problem (2.2) with $p = -1$ blows up in finite time if and only if $\mu < \tilde{\mu}$. Consequently, we conjecture that the large solution of Theorem 1.4 exists if and only if $N \leq 14$.

Figure 2 depicts computed profiles of the large radial solutions (v, ϕ) with $v(0) = 0$ and $\phi(0) < 0$ of the problem (1.3), with $R = 1$, for several values of the space dimension N (see Remark 4.4 for comments on the numerical method). In particular, the solution is shown for $N = 10$, the largest space dimension for which we proved its existence, and for $N = 14$, the largest space dimension for which we found it numerically. Of course, we can compute the large solution (v, ϕ) with $v(0) = 0$ and $\phi(0) < 0$ of the radial version of (1.3) for every value of $\mu = N - 1$, not necessarily integer, up to $\tilde{\mu} \approx 13.755$. As $\mu \to \tilde{\mu}$, the ϕ-component of the solution appears to approach $c_R\delta_R - c_0\delta_0$, for some positive constants c_0 and c_R, where δ_0 and δ_R denote the Dirac distributions centered at 0 and R, respectively.

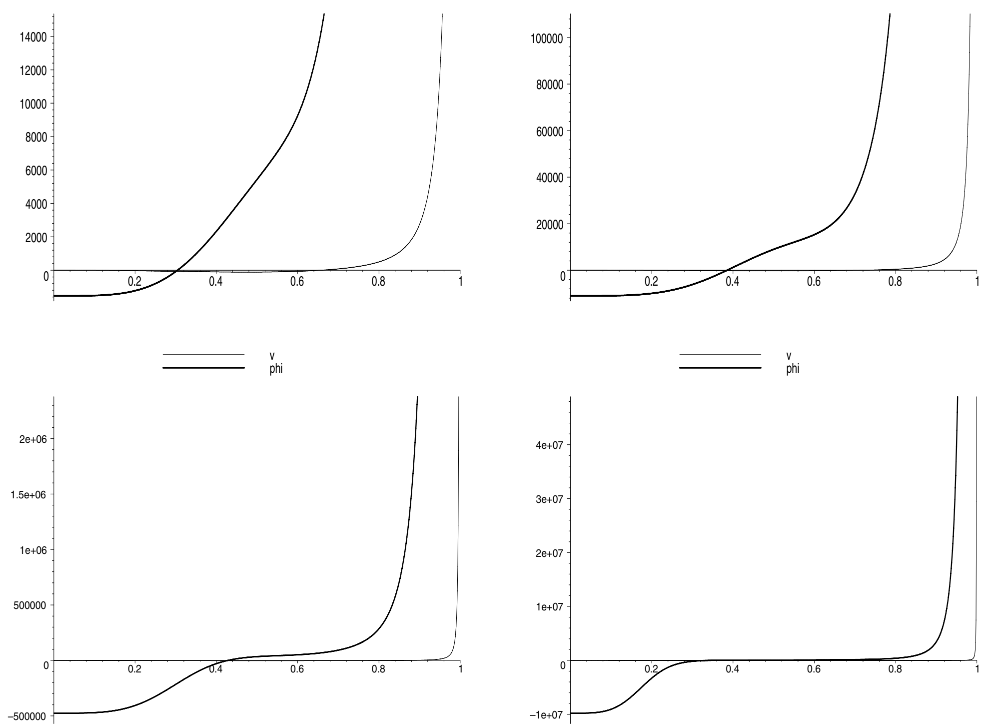


Fig. 2. *Large radial solutions (v, ϕ) with $v(0) = 0$ and $\phi(0) < 0$ of problem (1.3) with $R = 1$ for $N = 1$ (top left), $N = 3$ (top right), $N = 10$ (bottom left), and $N = 14$ (bottom right).*

We will now describe how the question of finite-time blow-up in the Cauchy problem (2.2) can be recast as a question regarding the existence of nontrivial solutions of a related boundary-value problem or, equivalently, a parameter-dependent fixed-point equation. This approach will allow us to exploit standard tools of nonlinear analysis (such as the degree of mapping and bifurcation theory) and to gain some additional information not otherwise available. As a corollary, we will obtain a Liouville-type result (existence of a positive solution) for the Dirichlet problem associated with the elliptic system (1.2), which is of independent interest.

Given $\mu \in \mathbb{R}_+$, the maximal solution (w, ϕ, ψ) of the Cauchy problem (2.2) with $p = -1$ blows up in finite time if and only if ϕ crosses zero at some point $r > 0$. Due to the scaling property of the system (see Remark 2.1), this happens if and only if there exists a (necessarily negative and unique) initial value p such that the ϕ-component of the corresponding maximal solution of (2.2) crosses zero at $r = 1$. In other words, the maximal solution of (2.2) with $p = -1$ blows up in finite time if and only if the boundary-value problem

$$\begin{cases} w' + \dfrac{\mu}{r}\, w = \phi, & w(0) = 0, \\ \phi' = \psi, & \phi(1) = 0, \\ \psi' + \dfrac{\mu}{r}\, \psi = w^2, & \psi(0) = 0 \end{cases} \tag{3.2}$$

has a (necessarily unique) nontrivial solution.

The problem (3.2) can be written as a parameter-dependent fixed-point equation of the form

$$u = T(\mu, u) \tag{3.3}$$

in $X := C([0,1], \mathbb{R}^3)$, where $T : \mathbb{R}_+ \times X \to X$ is a completely continuous operator, defined by

$$T(\mu, u)(r) := \left(\int_0^r \left(\frac{s}{r}\right)^{\mu} \phi(s)\, ds\, ,\ -\int_r^1 \psi(s)\, ds\, ,\ \int_0^r \left(\frac{s}{r}\right)^{\mu} w^2(s)\, ds \right),$$

for $\mu \in \mathbb{R}_+$, $u = (w, \phi, \psi) \in X$, and $r \in [0, 1]$. We are interested in the structure of the solution set

$$\Sigma := \{(\mu, u) \in \mathbb{R}_+ \times X : u = T(\mu, u)\}.$$

Clearly, Σ contains the branch of trivial solutions of (3.3), $\mathbb{R}_+ \times \{0\}$. Also, as mentioned above, (3.3) cannot have more than one nontrivial solution for any $\mu \in \mathbb{R}_+$; hence, the set $\Sigma \setminus (\mathbb{R}_+ \times \{0\})$ is the graph of a function $\mu \mapsto u_\mu$. Let M denote the domain of this function, that is,

$$M := \{\mu \in \mathbb{R}_+ : u = T(\mu, u) \text{ for some } u \in X \setminus \{0\}\}, \tag{3.4}$$

and define

$$\mu^* := \sup(\{\mu \in \mathbb{R}_+ \,:\, [0, \mu] \subset M\}), \tag{3.5}$$

with the understanding that $\sup(\emptyset) = 0$ and $\sup(A) = \infty$ if $A \subset \mathbb{R}_+$ is unbounded.

Remark 3.6. From the discussion leading to (3.3), it is evident that M, as defined in (3.4), coincides with the set of all $\mu \in \mathbb{R}_+$ for which the maximal solution of the Cauchy problem (2.2) with $p = -1$ blows up in finite time. Thus, M contains the interval $[0, 9]$, by Proposition 3.1. The subsequent arguments will prove this once again (we will only need the a priori estimates from Remark 3.4).

Remark 3.7. We can characterize M as the set of all $\mu \in \mathbb{R}_+$ for which the ϕ-component of the maximal solution of (2.2) with $p = -1$ is eventually positive. Since the solution depends continuously on μ, this property is stable under small perturbations of μ. It follows that the set M is open in $\mathbb{R}_+$ and, hence, μ^* is not an element of M. On the other hand, it is clear from the definition of μ^* that M contains the interval $[0, \mu^*)$.

Remark 3.8. Suppose that $\mu \in M$, $u_\mu = (w_\mu, \phi_\mu, \psi_\mu)$ is the corresponding nontrivial solution of (3.3), and $u = (w, \phi, \psi)$ is the corresponding maximal solution of the Cauchy problem (2.2) with $p = -1$. Then, due to the scaling property of the problem, u_μ is a "rescaling" of u. In fact, denoting the unique zero of ϕ by z_μ, we have $w_\mu(r) = z_\mu^3\, w(z_\mu r)$, $\phi_\mu(r) = z_\mu^4\, \phi(z_\mu r)$, and $\psi_\mu(r) = z_\mu^5\, \psi(z_\mu r)$, for all $r \in [0, 1]$. Since $u = (w, \phi, \psi)$ depends continuously on μ, so does z_μ. It follows that the function $\mu \mapsto u_\mu$ is continuous as a mapping from $M \subset \mathbb{R}_+$ into X.

Lemma 3.9. *Given $\mu \in M$, let u_μ denote the unique nontrivial solution of* (3.3). *Then $\|u_\mu\|_\infty \geq 4(\mu + 1)^2(\mu + 3) \geq 12$. Moreover, there exists a continuous function $\bar{f} : [0, \bar{\mu}) \to \mathbb{R}_+$, with $\bar{\mu} \approx 9.073$, such that $\|u_\mu\|_\infty \leq \bar{f}(\mu)$, provided that $\mu \in [0, \bar{\mu})$.*

Proof. Fix a number $\mu \in M$, let $u_\mu = (w_\mu, \phi_\mu, \psi_\mu)$ denote the corresponding nontrivial solution of (3.3), and let $u = (w, \phi, \psi)$ denote the corresponding maximal solution of the Cauchy problem (2.2) with $p = -1$. According to Remark 3.8, we have $\phi_\mu(r) = z_\mu^4\, \phi(z_\mu r)$ for all $r \in [0, 1]$, where z_μ is the zero of ϕ; in particular, $|\phi_\mu(0)| = z_\mu^4$. Recalling the a priori bounds in Remark 3.4, we obtain the estimate

$$|\phi_\mu(0)| \geq 4(\mu + 1)^2(\mu + 3) \geq 12 \tag{3.6}$$

and, furthermore, the existence of a continuous function $f : [0, \bar{\mu}) \to [1, \infty)$, with $\bar{\mu} \approx 9.073$, such that

$$|\phi_\mu(0)| \leq 4(\mu + 1)(\mu + 3)(\mu + 5) f(\mu)\,, \tag{3.7}$$

provided that $\mu \in [0, \bar{\mu})$. Next, observe that

$$\|\phi_\mu\|_\infty = |\phi_\mu(0)|\,, \qquad \|w_\mu\|_\infty \leq \frac{|\phi_\mu(0)|}{\mu + 1} \leq |\phi_\mu(0)|\,,$$

and

$$\|\psi_\mu\|_\infty \leq \frac{|\phi_\mu(0)|^2}{(\mu + 1)^2(\mu + 3)} \leq \frac{1}{3}\, |\phi_\mu(0)|^2\,;$$

since $|\phi_\mu(0)| \geq 12$ by (3.6), this implies that $\|(w_\mu, \phi_\mu, \psi_\mu)\|_\infty \leq |\phi_\mu(0)|^2$. Consequently, we have

$$|\phi_\mu(0)| \leq \|u_\mu\|_\infty \leq |\phi_\mu(0)|^2,$$

and now the assertions of the lemma follow from (3.6) and (3.7). ∎

PROPOSITION 3.10. *For every $\mu \in [0,\mu^*)$, with μ^* defined by (3.5), the unique nontrivial solution u_μ of (3.3) has a fixed-point index of -1. The graph $\mathcal{C} := \{(\mu,u_\mu) : \mu \in [0,\mu^*)\}$ is an unbounded, continuous curve in $\mathbb{R}_+ \times X$, and μ^* is greater than 9.*

Proof. We begin by computing the (Leray–Schauder) fixed-point index of the map $T(0,\cdot)$ in u_0, the nontrivial solution of (3.3) for $\mu = 0$. (While the existence of u_0 was established previously, the following argument will prove it once again.) Inspired by similar reasoning in [1], we define a completely continuous operator $S : \mathbb{R}_+ \times X \to X$, with $X := C([0,1],\mathbb{R}^3)$, by

$$S(\lambda,u)(r) := \Big(\int_0^r \phi(s)\,ds\,,\ -\int_r^1 \psi(s)\,ds\,,\ \int_0^r \big(w^2(s)+\lambda\big)\,ds\Big),$$

for $\lambda \in \mathbb{R}_+$, $u = (w,\phi,\psi) \in X$, and $r \in [0,1]$, and consider the parameter-dependent fixed-point problem in X,

$$u = S(\lambda,u). \tag{3.8}$$

Note that $S(0,\cdot) = T(0,\cdot)$; that is, if $\lambda = 0$, then (3.8) coincides with (3.3) with $\mu = 0$.

Now let $\lambda \in \mathbb{R}_+$ and suppose that $u_{0\lambda} = (w_{0\lambda},\phi_{0\lambda},\psi_{0\lambda})$ is a solution of (3.8). Since $\phi_{0\lambda}'' = w_{0\lambda}^2 + \lambda \geq 0$, the function $\phi_{0\lambda}$ is convex; thus, $\phi_{0\lambda}(r) \leq \phi_{0\lambda}(0)(1-r)$ for all $r \in [0,1]$. Arguing as in the proof of Proposition 3.1, we derive an upper bound for $w_{0\lambda}$, then a lower bound for $\psi_{0\lambda}$. Since $|\phi_{0\lambda}(0)| = \int_0^1 \psi_{0\lambda}(s)\,ds$, the lower bound for $\psi_{0\lambda}$ yields a quadratic inequality for $|\phi_{0\lambda}(0)|$, namely, $|\phi_{0\lambda}(0)|^2 - 24\,|\phi_{0\lambda}(0)| + 12\lambda \leq 0$. It follows that $\lambda \leq 12$ and $|\phi_{0\lambda}(0)| \leq 24$. This shows that (3.8) does not have any solutions if $\lambda > 12$; moreover, the uniform bound on $|\phi_{0\lambda}(0)|$ implies a uniform bound on $\|u_{0\lambda}\|_\infty$.

Choosing a sufficiently large $\rho > 0$, we infer that $\deg\big(Id_X - S(\lambda,\cdot), B_\rho^X(0), 0\big)$ is well defined for $\lambda \in \mathbb{R}_+$, independent of λ (due to homotopy invariance), and in fact equal to zero (since (3.8) has no solutions for $\lambda > 12$). It follows that

$$\deg\big(Id_X - T(0,\cdot), B_\rho^X(0), 0\big) = \deg\big(Id_X - S(0,\cdot), B_\rho^X(0), 0\big) = 0\,.$$

However, a routine homotopy argument shows that the index of $T(0,\cdot)$ in the trivial fixed point 0 equals 1. This proves, once again, the existence of the nontrivial fixed point u_0 (and thereby the fact that $\mu^* > 0$) and shows, more importantly, that the index of $T(0,\cdot)$ in u_0 is -1. But then, by homotopy along the continuous curve $\mathcal{C} := \{(\mu,u_\mu) : \mu \in [0,\mu^*)\}$, the fixed-point index of $T(\mu,\cdot)$ in u_μ is -1 for every $\mu \in [0,\mu^*)$.

We are now in a position to complete the proof of the proposition by applying a Rabinowitz-type argument (see [21]). Suppose that the curve $\mathcal{C}$ is bounded. Then $\mu^* < \infty$ and there exists a constant $\rho > 0$ such that $\|u_\mu\|_\infty < \rho$ for all $\mu \in [0,\mu^*)$. Also, due to Lemma 3.9, $\|u_\mu\|_\infty \geq 12$ for all $\mu \in [0,\mu^*)$ and, according to Remark 3.7, the equation (3.3) has no nontrivial solution for $\mu = \mu^*$. It follows that $\deg\big(Id_X - T(\mu,\cdot), B_\rho^X(0) \setminus \overline{B}_1^X(0), 0\big)$ is well defined for $\mu \in [0,\mu^*]$, independent of μ, and in fact equal to zero. Clearly, this contradicts the fact that $T(0,\cdot)$ has index -1 in u_0. It follows that $\mathcal{C}$ is unbounded in $\mathbb{R}_+ \times X$. In conjunction with Lemma 3.9, according to which $\mathcal{C} \cap ([0,\hat\mu] \times X)$ is bounded for every $\hat\mu \in \mathbb{R}_+$ with $\hat\mu < \bar\mu \approx 9.073$, this proves, once again, that $\mu^* \geq \bar\mu$ and thus $\mu^* > 9$. ∎

REMARK 3.11. In light of the numerical evidence described in Remark 3.5, we conjecture that $\mu^* \approx 13.755$, that the set M coincides with the interval $[0, \mu^*)$, and that the solution branch $\mathcal{C}$ bifurcates from infinity at μ^*.

We conclude this section with a comment on the Dirichlet problem for the elliptic system (1.2) on a ball in $\mathbb{R}^N$,

$$\begin{cases} -\Delta v = \theta & \text{in } B_R^N(0), \\ -\Delta\theta = |\nabla v|^2 & \text{in } B_R^N(0), \\ v = \theta = 0 & \text{on } \partial B_R^N(0). \end{cases} \tag{3.9}$$

Assuming that $R = 1$ (due to the scaling property, this entails no loss of generality), there is a one-to-one correspondence between the radial solutions of (3.9) and the solutions of the boundary-value problem (3.2) or the equivalent fixed-point equation (3.3), with $\mu = N - 1$, $\theta = -\phi$, and $v(r) = -\int_r^1 w(s)\,ds$ for $r \in [0, 1]$. Hence, Proposition 3.10 implies that (3.9) has a unique nontrivial radial solution (v, θ) as long as the space dimension N does not exceed 10; it is easily checked that both components of this solution are positive and strictly decreasing functions of the radial variable r.

COROLLARY 3.12. *For every $N \in \mathbb{N}$ with $N \leq 10$ and every $R > 0$, the Dirichlet problem* (3.9) *has a unique nontrivial radially symmetric solution (v, θ). Both components of this solution are positive and decreasing functions of the radial variable r.*

Our numerical evidence (see Remark 3.5) suggests that the solution of Corollary 3.12 exists, in fact, if and only if $N \leq 14$. Figure 3 shows computed profiles of this solution for $R = 1$ and the two extreme values of the space dimension, $N = 1$ and $N = 14$. Of course, we can compute the nontrivial solution (v, θ) of the radial version of (3.9) for any $\mu \in [0, \tilde{\mu})$, with $\tilde{\mu} \approx 13.755$, in place of the integer $N - 1$. As $\mu \to \tilde{\mu}$, the θ-component of the solution appears to approach a multiple of δ_0 (the Dirac distribution centered at 0).

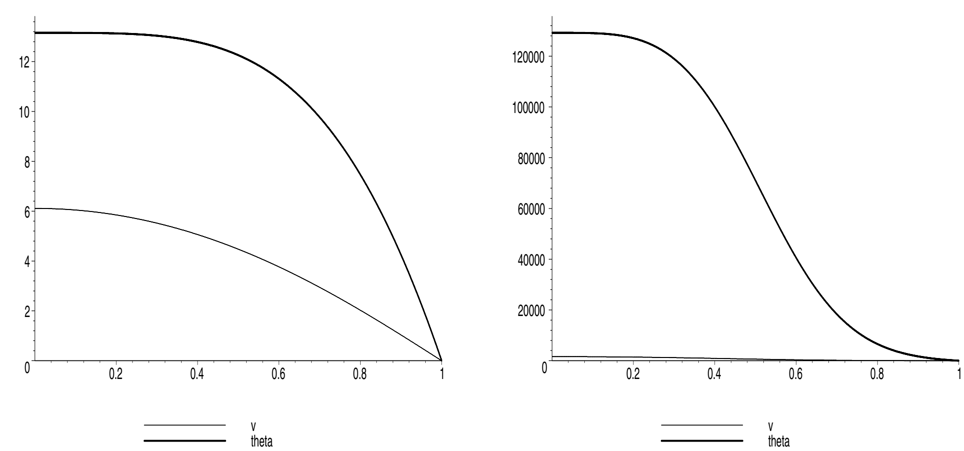


FIG. 3. *Positive radial solutions (v, θ) of the Dirichlet problem* (3.9) *with $R = 1$ for $N = 1$* (*left*) *and $N = 14$* (*right*).

4. Asymptotic behavior. Let $(w, \phi, \psi) \in C^1([0,R), \mathbb{R}^3)$ be the maximal solution of the Cauchy problem (2.2) for a given $\mu \in \mathbb{R}_+$ and $p \in \mathbb{R}$. By Lemma 2.4, we know that R is finite if and only if $\phi(r)$ is eventually positive and that in this case $w(r)$, $\phi(r)$, $\psi(r) \to \infty$ as $r \to R^-$. In view of the existing literature on boundary blow-up in elliptic equations (see, for example, [4, 24]), it is natural to expect asymptotic behavior of the form $Q/(R-r)^q$, with positive constants Q and q. In fact, we will prove the following result.

PROPOSITION 4.1. *Let $(w, \phi, \psi) \in C^1([0,R), \mathbb{R}^3)$ be the maximal solution of the Cauchy problem* (2.2)*, for a given $\mu \in \mathbb{R}_+$ and $p \in \mathbb{R}$, and suppose that R is finite. Then, as $r \to R^-$,*

$$w(r) \sim \frac{60}{(R-r)^3}\,, \quad \phi(r) \sim \frac{180}{(R-r)^4}\,, \quad \psi(r) \sim \frac{720}{(R-r)^5}\,.$$

Let us note that if all three of the functions w, ϕ, ψ exhibit asymptotic behavior of the form $Q/(R-r)^q$, it is easy to see that the constants Q and q are necessarily as above.

We will prove Proposition 4.1 under the assumption that $R = 1$; thanks to the scaling property, this entails no loss of generality. The proof will be achieved by analyzing a system of equations derived from (2.2) by a suitable change of variables.

Given any solution (w, ϕ, ψ) of (2.2), define functions α, β, γ by

$$\alpha(r) := \frac{(1-r)^3}{60}\, w(r)\,, \quad \beta(r) := \frac{(1-r)^4}{180}\, \phi(r)\,, \quad \gamma(r) := \frac{(1-r)^5}{720}\, \psi(r)\,;$$

then (α, β, γ) is a solution of

$$\begin{cases} (1-r)\left(\alpha' + \dfrac{\mu}{r}\,\alpha\right) = 3(\beta - \alpha)\,, & \alpha(0) = 0\,, \\ (1-r)\,\beta' = 4(\gamma - \beta)\,, & \beta(0) = \dfrac{p}{180}\,, \\ (1-r)\left(\gamma' + \dfrac{\mu}{r}\,\gamma\right) = 5(\alpha^2 - \gamma)\,, & \gamma(0) = 0\,. \end{cases} \tag{4.1}$$

Just like (2.2), the system (4.1) is singular at $r = 0$, but this does not affect the well-posedness of the initial value problem; in addition, (4.1) is singular at $r = 1$.

REMARK 4.2. Suppose that (w, ϕ, ψ) is the maximal solution of (2.2), for a given $\mu \in \mathbb{R}_+$ and $p \in \mathbb{R}$, with interval of existence $[0, R_p)$; let (α, β, γ) be the corresponding solution of (4.1), as defined above. Clearly, if $R_p < 1$, then (α, β, γ) ceases to exist before reaching the singularity at $r = 1$; in fact, $\alpha(r)$, $\beta(r)$, $\gamma(r) \to \infty$ as $r \to R_p^-$. Also, if $R_p > 1$, then (α, β, γ) can be continued beyond the singularity at $r = 1$, and $\alpha(r)$, $\beta(r)$, $\gamma(r) \to 0$ as $r \to 1^-$. Finally, if $R_p = 1$, then (α, β, γ) exists up to the singularity at $r = 1$, but the behavior near the singularity is not obvious. The assertion of Proposition 4.1 (with $R = 1$) is that, in this case, $\alpha(r)$, $\beta(r)$, $\gamma(r) \to 1$ as $r \to 1^-$.

REMARK 4.3. Recall that for every $\mu \in \mathbb{R}_+$, there is exactly one initial value $p_\mu^+ > 0$ and at most one initial value $p_\mu^- < 0$ such that the maximal solution of (2.2) with $p = p_\mu^\pm$ blows up at $r = 1$. Let p_μ^* denote one such value. Due to the scaling property of the problem, all solutions with $\operatorname{sign}(p) = \operatorname{sign}(p_\mu^*)$ blow up in finite time. Moreover, there is a one-to-one correspondence between the initial value p and the exit time R_p of the solution; in fact, $R_p^4 = p_\mu^*/p$ (see Remark 2.1). It follows that if $p > p_\mu^* > 0$ or $p < p_\mu^* < 0$, then $R_p < 1$, and consequently $\alpha(r)$, $\beta(r)$, $\gamma(r) \to \infty$ as $r \to R_p^-$. Also, if $0 < p < p_\mu^*$ or $p_\mu^* < p < 0$, then $R_p > 1$, and consequently $\alpha(r)$, $\beta(r)$, $\gamma(r) \to 0$ as $r \to 1^-$.

REMARK 4.4. The observations in the preceding remark allow us to use a shooting method to numerically approximate p_μ^+ and, if it exists, p_μ^-, that is, the critical initial values p for which the maximal solution of (2.2), for a given $\mu \in \mathbb{R}_+$, blows up at $r = 1$. Solving the Cauchy problem (4.1) with $p = p_\mu^\pm$, we can then construct the solutions of (2.2) that blow up at $r = 1$, and thereby the large radial solutions of the problem (1.3) with $R = 1$. All the graphs in the preceding sections were generated in this way (with a suitable rescaling in the case of Figure 3).

Our experiments suggest that p_μ^- exists if and only if $\mu < \tilde{\mu}$, for some number $\tilde{\mu} \approx 13.755$, which, due to the scaling property, must coincide with the number μ^* defined in (3.5). In fact, we find that p_μ^- is a strictly decreasing function of μ that approaches $-\infty$ as $\mu \to \tilde{\mu}$.

To prove Proposition 4.1 for $R = 1$, we must show that all three components of the maximal solution (α, β, γ) of the Cauchy problem (4.1) with $\mu \in \mathbb{R}_+$ and $p = p_\mu^\pm$ converge to 1 as $r \to 1^-$ (recall Remarks 4.2 and 4.3). While our numerical experiments leave no doubt about this (see Figure 4 for examples of computed solutions), the proof requires a small detour in dynamical systems; we refer the reader to [22] for terminology and basic properties.

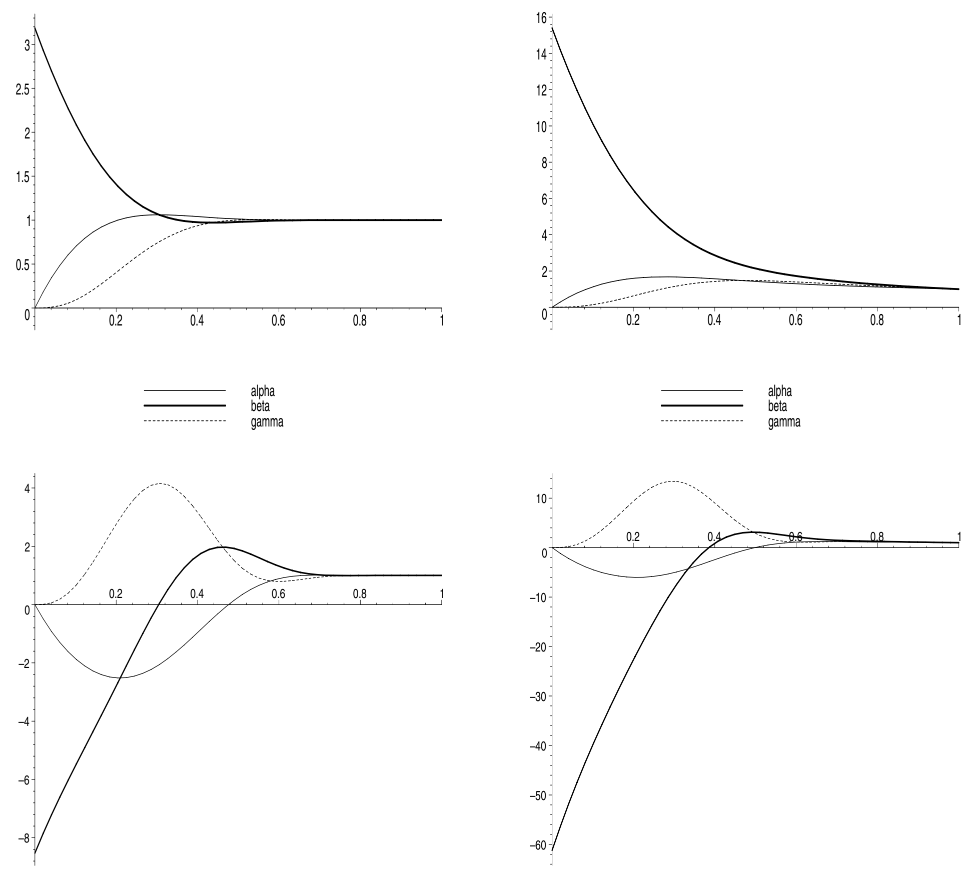


FIG. 4. *Critical solutions* (α, β, γ) *of problem* (4.1) *for* $\mu = 0$, $p = p_0^+ \approx 180 \times 3.185739467$ (*top left*), $\mu = 2$, $p = p_2^+ \approx 180 \times 15.39874038$ (*top right*), $\mu = 0$, $p = p_0^- \approx -180 \times 8.525812707$ (*bottom left*), *and* $\mu = 2$, $p = p_2^- \approx -180 \times 61.20559852$ (*bottom right*).

It is convenient to perform another change of variables in the system (4.1), letting $r = 1 - e^{-t}$ and $a(t) = \alpha(r)$, $b(t) = \beta(r)$, $c(t) = \gamma(r)$. With this rescaling of the independent variable, (4.1) is equivalent to

$$\text{(4.2)} \qquad \begin{cases} a' + \dfrac{\mu}{e^t - 1}\, a = 3(b - a), & a(0) = 0, \\ b' = 4(c - b), & b(0) = \dfrac{p}{180}, \\ c' + \dfrac{\mu}{e^t - 1}\, c = 5(a^2 - c), & c(0) = 0. \end{cases}$$

Note that the singularity of (4.1) at $r = 1$ has been moved to $t = \infty$; moreover, the system (4.2) is *autonomous* for $\mu = 0$ and *asymptotically autonomous* for $\mu > 0$. For notational convenience, we write the system of differential equations in (4.2) as

$$\text{(4.3)} \qquad x' + \frac{\mu}{e^t - 1}\, E(x) = F(x),$$

where $x = (a, b, c)$ takes values in $\mathbb{R}^3$, $E : \mathbb{R}^3 \to \mathbb{R}^3$ is the linear mapping defined by $E(a, b, c) := (a, 0, c)$, and $F : \mathbb{R}^3 \to \mathbb{R}^3$ is the vector field defined by

$$F(a, b, c) := \big(3(b - a), 4(c - b), 5(a^2 - c)\big).$$

Remark 4.5. For $t > 0$ and $a \geq 0$, the system (4.3), with arbitrary $\mu \in \mathbb{R}_+$, satisfies the well-known Kamke condition and, thus, a comparison principle (see, for example, [25]). To be precise, let t_0, $t_1 \in [0, \infty]$ with $t_0 < t_1$ and suppose that x_1, $x_2 \in C([t_0, t_1), \mathbb{R}^3) \cap C^1((t_0, t_1), \mathbb{R}^3)$. If x_1 is a subsolution of (4.3) with $a_1 \geq 0$, if x_2 is a supersolution of (4.3), and if $x_1(t_0) \leq x_2(t_0)$ $(x_1(t_0) < x_2(t_0))$, then we have $x_1(t) \leq x_2(t)$ $(x_1(t) < x_2(t))$ for all $t \in [t_0, t_1)$.

By a subsolution (supersolution) of (4.3) we mean a function $x = (a, b, c)$ satisfying the differential inequality obtained from (4.3) by replacing "=" with "$\leq$" ("$\geq$"). Also, given vectors x_1, $x_2 \in \mathbb{R}^3$, we write $x_1 \leq x_2$ or $x_2 \geq x_1$ ($x_1 < x_2$ or $x_2 > x_1$) if the respective inequality holds componentwise, and we call a vector $x \in \mathbb{R}^3$ nonnegative (positive) if $x \geq \bar{0}$ ($x > \bar{0}$), where $\bar{0} := (0, 0, 0)$.

Remark 4.6. For every $\lambda \in [0, \infty]$, let $\bar{\lambda}$ denote the vector $(\lambda, \lambda, \lambda)$. For arbitrary $\mu \in \mathbb{R}_+$, $\bar{0}$ is a solution of (4.3), and $\bar{1}$ is a supersolution (a solution if $\mu = 0$). More generally, for every $\lambda \in [0, 1]$, $\bar{\lambda}$ is a supersolution. Furthermore, for every $\lambda \in (1, \infty)$, there exists a number $\tau \in [0, \infty)$, depending only on λ and μ, such that $((\lambda + 1)/2, \lambda, \lambda)$ is a subsolution on the interval (τ, ∞) (where $\tau = 0$ if $\mu = 0$).

Proposition 4.7. *For a given $\mu \in \mathbb{R}_+$, let x be a nonnegative maximal forward solution of* (4.3)*.*

(a) *If x is unbounded, then x blows up in finite time and approaches $\overline{\infty}$.*

(b) *If x is bounded, then x converges to either $\bar{0}$ or $\bar{1}$.*

Proof. Fix $\mu \in \mathbb{R}_+$ and let $x = (a, b, c)$ be a nonnegative maximal forward solution of the system (4.3).

(a) Suppose that $x = (a, b, c)$ is unbounded. First we will show that b is unbounded. By way of contradiction, suppose that $b \leq b_0$ for some positive constant b_0. Then we have

$$a' \leq a' + \frac{\mu}{e^t - 1}\, a = 3(b - a) \leq 3(b_0 - a)\,,$$

which implies that a is bounded. A similar argument then shows that c is bounded as well, and this contradicts the unboundedness of $x = (a, b, c)$. Thus, b is unbounded.

Now fix a number $p_0 > 0$ such that the solution (w_0, ϕ_0, ψ_0) of the initial-value problem (2.2) with $p = p_0$ blows up at a point $R_0 < 1$. The corresponding solution $x_0 = (a_0, b_0, c_0)$ of the initial-value problem (4.2) then blows up at $-\log(1 - R_0)$. Recall that the b-component of the trajectory $x = (a, b, c)$ is unbounded and choose a point τ in the interval of existence of x such that $b(\tau) \geq p_0/180$; define $\tilde{x} := x(\tau + \cdot\,)$. Then we have

$$\tilde{x}' + \frac{\mu}{e^t - 1}\, E(\tilde{x}) \;\geq\; \tilde{x}' + \frac{\mu}{e^{\tau+t} - 1}\, E(\tilde{x}) \;=\; F(\tilde{x})$$

and

$$\tilde{x}(0) = x(\tau) = \big(a(\tau), b(\tau), c(\tau)\big) \geq \Big(0, \frac{p_0}{180}, 0\Big) = x_0(0)\,.$$

Thanks to the comparison principle in Remark 4.5, it follows that $\tilde{x} \geq x_0$. In particular, $\tilde{x}$ blows up in finite time, and then so does x. By reasoning as in the proof of Lemma 2.4(b), it is now easy to verify that all three components of x approach infinity.

(b) Suppose that x is bounded. First, consider the autonomous case, $\mu = 0$. The vector field F is cooperative in the half-space $a \geq 0$ and, in particular, in the nonnegative cone $\mathbb{R}^3_+$; moreover, $\operatorname{div}(F) \equiv -12$ and F has exactly two zeros, at $\bar{0}$ and $\bar{1}$. Thus, F generates a monotone, volume-contracting semiflow Φ in $\mathbb{R}^3_+$, with exactly two equilibria, at $\bar{0}$ and $\bar{1}$. The equilibrium at $\bar{0}$ is a stable node; the equilibrium at $\bar{1}$ is a saddle point with a two-dimensional stable manifold (the eigenvalues are 1 and $-13/2 \pm i\sqrt{71}/2$).

Moreover, the system can be embedded into a cooperative system in all of $\mathbb{R}^3$ by replacing the nonlinear term a^2 in the third component of the vector field F with $a|a|$; the extended system still has negative divergence, and it has equilibria at $\bar{0}$ and $\pm\bar{1}$. Morris Hirsch proved (see [13, Theorem 1]) that every compact limit set of a cooperative or competitive system in $\mathbb{R}^3$ is either a cycle or contains an equilibrium. Another result of Hirsch's (see [12, Theorem 7]) guarantees that a cooperative system in $\mathbb{R}^3$ with negative divergence cannot have any cycles. Moreover, for our particular system, it is easy to see that any compact limit set containing one of the equilibria is in fact a singleton. Combining these results we infer that every bounded (forward or backward) trajectory of the system converges. In particular, the trajectory x converges to either $\bar{0}$ or $\bar{1}$.

Now consider the nonautonomous case, $\mu > 0$. As we observed before, the system (4.3) is asymptotically autonomous. An old result of Markus [18] implies that the ω-limit set K of the trajectory x is a nonempty compact and connected subset of $\mathbb{R}^3_+$; moreover, $\operatorname{dist}(x(t), K) \to 0$ as $t \to \infty$, and K is invariant under the semiflow Φ of the autonomous limit system, that is, (4.3) with $\mu = 0$. A more recent result by Mischaikow, Smith, and Thieme (see [19, Theorem 1.8]) implies that K is also *chain-recurrent* under Φ.

We claim that $K \subset \{\bar{0}, \bar{1}\}$. By way of contradiction, suppose there is a point $z \in K \setminus \{\bar{0}, \bar{1}\}$. In light of what we proved for the autonomous case, since K is compact and Φ-invariant, the Φ-trajectory through z must converge, both forward and backward in time. Backward in time, it can only converge to $\bar{1}$ (since $\bar{0}$ is stable). Thus, z belongs to the unstable manifold of $\bar{1}$, and it follows that, forward in time, the Φ-trajectory through z can only converge to $\bar{0}$. Hence, K consists of the two equilibria, $\bar{0}$ and $\bar{1}$, and a heteroclinic orbit connecting the two; such a set is obviously not chain-recurrent. The contradiction proves that $K \subset \{\bar{0}, \bar{1}\}$. In fact, since K is nonempty and connected, we have either $K = \{\bar{0}\}$ or $K = \{\bar{1}\}$; that is, x converges to either $\bar{0}$ or $\bar{1}$. $\square$

COROLLARY 4.8. *Let $\mu \in \mathbb{R}_+$ and $p \in \mathbb{R}$ be such that the maximal solution (w, ϕ, ψ) of the Cauchy problem* (2.2) *blows up at $r = 1$. Then the maximal solution (a, b, c) of* (4.2) *converges to $\bar{1}$.*

Proof. Under the assumptions of the corollary, (w, ϕ, ψ) approaches $\overline{\infty}$ at $r = 1$; thus, the corresponding solution (α, β, γ) of (4.1) is eventually positive and exists on $[0, 1)$ (see Remark 4.2). This means that $x = (a, b, c)$ is eventually positive and exists on $[0, \infty)$. By part (a) of Proposition 4.7, it follows that x is bounded, and then part (b) of the same proposition implies that x converges to either $\bar{0}$ or $\bar{1}$. Now suppose that $x(t) \to \bar{0}$ as $t \to \infty$ and choose $t_0 \in (0, \infty)$ such that $\bar{0} < x(t_0) < \bar{1}$. Since the solution of (4.2) depends continuously on p, we can find a value $\tilde{p}$ close to p, with $|\tilde{p}| > |p|$ such that the corresponding maximal solution $\tilde{x}$ of (4.2) exists at $t = t_0$ and satisfies $\bar{0} < \tilde{x}(t_0) < \bar{1}$. Since $\bar{0}$ is a solution and $\bar{1}$ is a supersolution of (4.3), the comparison principle in Remark 4.5 implies that $\bar{0} < \tilde{x}(t) < \bar{1}$ for all $t \geq t_0$, as long as $\tilde{x}$ exists. From Remark 4.3, however, we know that $\tilde{x}$ goes to $\overline{\infty}$ (in finite time). This contradiction proves that x does not converge to $\bar{0}$, and therefore it must converge to $\bar{1}$. ∎

Proof of Proposition 4.1. Due to the scaling property of the system (2.2) (see Remark 2.1), it suffices to prove the assertion of the proposition for $R = 1$; in this case, it is an immediate consequence of Corollary 4.8 (see Remark 4.2). ∎

Proof of Theorem 1.5. Since every large radial solution (v, ϕ) of the problem (1.3), for a given $R > 0$, corresponds to a solution (w, ϕ, ψ) of (2.2) that blows up at R, the asymptotic behavior of ϕ is clear from Proposition 4.1. Moreover, the asymptotic behavior of v, given by $v(r) = \int_0^r w(s)\, ds$ for $0 \leq r < R$, follows readily from that of w. ∎

In closing, we note that Hirsch's results on cooperative systems in $\mathbb{R}^3$ (see [12, 13]) allow us to completely describe the dynamics of the monotone, volume-contracting semiflow Φ in $\mathbb{R}^3_+$, induced by the vector field F. First, it is easily verified that Φ is, in fact, *strongly* monotone (even though F is irreducible only for $a > 0$). As shown in the first part of the proof of Proposition 4.7(b), Hirsch's results imply that every forward trajectory of Φ either converges to $\bar{0}$ (a stable node) or to $\bar{1}$ (a saddle point), or it approaches $\overline{\infty}$, necessarily in finite time. Clearly, both $\bar{0}$ and $\overline{\infty}$ are stable attractors. In fact, using the sub- and super-solutions constructed in Remark 4.6, we see that the open order intervals $(\bar{0}, \bar{1})$ and $(\bar{1}, \overline{\infty})$ are positively invariant and contained in the basins of attraction of $\bar{0}$ and $\overline{\infty}$, respectively. The two basins of attraction are separated by the (two-dimensional) stable manifold $W_s(\bar{1})$ of the saddle point. The (one-dimensional) unstable manifold $W_u(\bar{1})$ has a positive tangent vector at $\bar{1}$, which implies that $W_u(\bar{1}) \setminus \{\bar{1}\}$ is contained in the union of the order intervals $(\bar{0}, \bar{1})$ and $(\bar{1}, \overline{\infty})$. Thus, every forward trajectory on $W_u(\bar{1}) \setminus \{\bar{1}\}$ either converges to $\bar{0}$ or approaches $\overline{\infty}$. It follows that $W_u(\bar{1}) \setminus \{\bar{1}\}$ consists of two heteroclinic orbits connecting $\bar{1}$ to $\bar{0}$ and $\overline{\infty}$, respectively.

Appendix. The following algorithm allows the construction of increasing sequences of polynomial lower bounds and decreasing sequences of polynomial upper bounds for the maximal solution (w, ϕ, ψ) of the Cauchy problem (2.2) with $p = -1$ and arbitrary $\mu \in \mathbb{R}_+$. The bounds being polynomials, the computations amount to symbolic algebra on the coefficients. We used Maple (Version 9.5) to perform these computations; the relevant commands are provided below.

Given a polynomial P in r (whose coefficients are rational functions of another variable), the command

```
L:=normal(CoefficientList(P,r)):
```

generates the list of coefficients of P (in increasing order, starting with the zero-order term) and writes each coefficient in "normal form" (that is, as a quotient of polynomials). The command

```
n:=nops(L):
```

gives the number of entries in the list L; that is, $n-1$ is the degree of P.

Algorithm: Construction of Polynomial Bounds.

0. Specify a lower (or an upper) bound ϕ_0 for ϕ (ϕ_0 a suitable polynomial in r).

```
with(PolynomialTools):
phi0:=-1;  # Other choices are possible.
```

1. Compute a lower (upper) bound w_0 for w: $w_0 := \int_0^r (s/r)^\mu \phi_0(s)\,ds$.

```
L:=normal(CoefficientList(phi0,r)):  n:=nops(L):
      w0:=sum(L[i]*r^i/(mu+i),i=1..n);
```

2. Compute an upper (lower) bound ψ_0 for ψ, valid as long as $\max(w, w_0) \le 0$: $\psi_0 := \int_0^r (s/r)^\mu w_0^2(s)\,ds$.

```
L:=normal(CoefficientList(w0^2,r)):  n:=nops(L):
      psi0:=sum(L[i]*r^i/(mu+i),i=1..n);
```

3. Compute an upper (lower) bound ϕ_0 for ϕ: $\phi_0 := -1 + \int_0^r \psi_0(s)\,ds$.

```
L:=normal(CoefficientList(psi0,r)):  n:=nops(L):
      phi0:=-1+sum(L[i]*r^i/i,i=1..n);
```

4. Return to Step 1 (or proceed to "Computation of coefficient c" below).

Starting with the trivial lower bound $\underline{\phi}_0 := -1$ for ϕ, the algorithm produces the bounds $\underline{w}_1$, $\overline{\psi}_1$, $\overline{\phi}_1$, $\overline{w}_1$, $\underline{\psi}_1$, $\underline{\phi}_1$, ..., referred to in the proof of Proposition 3.1; these bounds are valid on the interval $[0, r_0]$, where $r_0 := (4(\mu+1)(\mu+3)(\mu+5))^{1/4}$ is the unique positive root of $\overline{w}_1$ (note that for all $k \in \mathbb{N}$, $\underline{w}_1 \le \underline{w}_k \le w \le \overline{w}_k \le \overline{w}_1 \le 0$ on $[0, r_0]$).

Now suppose that w_0, ψ_0, and ϕ_0 have been constructed according to Steps 1–3 of the algorithm, starting with an initially negative upper bound for ϕ. Then we have $w \le w_0 \le 0$, $\psi \ge \psi_0 \ge 0$, and $\phi \ge \phi_0$ on some interval $[0, r_0]$ with $r_0 > 0$. The arguments in the proof of Proposition 3.1 show that the following "tail estimate" holds for every $r > r_0$, provided that $\phi(r) \le 0$:

$$0 \ge (\mu-1)^2\phi(r) \ge a\left(\frac{r_0}{r}\right)^{2(\mu-1)} - b\left(\frac{r_0}{r}\right)^{\mu-1} + c\,,$$

where

$$a := \frac{1}{2}r_0^2\, w_0^2(r_0)\,, \qquad b := r_0^2\, w_0^2(r_0) + (\mu-1)r_0\psi_0(r_0)\,,$$

$$c := \frac{1}{2}r_0^2\, w_0^2(r_0) + (\mu-1)r_0\psi_0(r_0) + (\mu-1)^2\phi_0(r_0)\,.$$

If we use the bounds $w_0 = \overline{w}_k$, $\psi_0 = \underline{\psi}_k$, and $\phi_0 = \underline{\phi}_k$, for some $k \in \mathbb{N}$, and choose r_0 to be the positive root of $\overline{w}_1$ (or the positive root of $\overline{\phi}_1$), the coefficients a, b, c, and various other relevant quantities, such as $\phi_0(r_0)$, are rational functions of μ; in fact, r_0^4 is a polynomial in μ, while $r^2 w_0^2(r)$, $r\psi_0(r)$, and $\phi_0(r)$ are polynomials in r^4, whose coefficients are rational functions of μ. This facilitates the symbolic computation of these quantities and allows us to locate their positive roots by inspecting the coefficients of the respective numerator polynomials (the roots of the denominator polynomials are negative integers). Recall that finding the positive roots of c, in particular, is a crucial step in the proof of Proposition 3.1.

Algorithm: Computation of Coefficient c.

0. Specify $\rho := r_0^4$ (ρ a suitable polynomial in μ).

```
rho:=4*(mu+1)*(mu+3)*(mu+5):
# Alternatively:  rho:=4*(mu+1)^2*(mu+3):
```

1. Compute coefficients of $(1/2)r^2 w_0^2(r) + (\mu-1)r\psi_0(r) + (\mu-1)^2\phi_0(r)$.

```
P:=r^2*w0^2/2+(mu-1)*r*psi0+(mu-1)^2*phi0:
L:=normal(CoefficientList(P,r)):  n:=(nops(L)-1)/4+1:
```

2. Compute coefficient $c = c_1/c_2$, with numerator polynomial c_1 and denominator polynomial c_2.

```
c:=normal(sum(L[4*i-3]*rho^(i-1),i=1..n)):
c1:=sort(numer(c));  c2:=denom(c);
```

If we choose $w_0 = \overline{w}_2$, $\psi_0 = \underline{\psi}_2$, $\phi_0 = \underline{\phi}_2$, and $\rho = 4(\mu+1)(\mu+3)(\mu+5)$, as in the proof of Proposition 3.1, then the numerator polynomial c_1 of c has exactly two positive roots (by Descartes's rule of signs and the intermediate-value theorem), one near 0.747 and the other near 9.073.

Acknowledgments. This project was started while M. Lazzo and P. G. Schmidt were visiting the Universidad Politécnica de Madrid and the Universidad Complutense de Madrid; both wish to express their gratitude to J. F. Padial and J. I. Díaz for their generous support. Most of the work was done while M. Lazzo was visiting Auburn University; she would like to thank the faculty and staff of the Department of Mathematics and Statistics for their kind hospitality. Both M. Lazzo and P. G. Schmidt are grateful for stimulating discussions with W. Shen and G. Hetzer at Auburn. Finally, the authors wish to thank the anonymous referees for their constructive criticism and valuable suggestions.

REFERENCES

[1] C. Azizieh, P. Clément, and E. Mitidieri, *Existence and a priori estimates for positive solutions of p-Laplace equations*, J. Differential Equations, 184 (2002), pp. 422–442.

[2] C. Bandle and H. Brunner, *Blowup in diffusion equations: A survey*, J. Comput. Appl. Math., 97 (1998), pp. 3–22.

[3] C. Bandle, G. Díaz, and J. I. Díaz, *Solutions d'équations de réaction-diffusion non linéaires explosant au bord parabolique*, C. R. Acad. Sci. Paris Sér. I Math., 318 (1994), pp. 455–460.

[4] C. Bandle and M. Marcus, *"Large" solutions of semilinear elliptic equations: Existence, uniqueness and asymptotic behaviour*, J. Anal. Math., 58 (1992), pp. 9–24.

[5] F.-C. Şt. Cîrstea and V. D. Rădulescu, *Entire solutions blowing up at infinity for semilinear elliptic systems*, J. Math. Pures Appl., 81 (2002), pp. 827–846.

[6] M. G. Crandall, P. L. Lions, and P. E. Souganidis, *Maximal solutions and universal bounds for some partial differential equations of evolution*, Arch. Rational Mech. Anal., 105 (1989), pp. 163–190.

[7] J. I. Díaz, *Obstruction and some approximate controllability results for the Burgers equation and related problems*, in Control of Partial Differential Equations and Applications (Proc. Int. Conf., Laredo, 1994), Lecture Notes in Pure and Appl. Math. 174, E. Casas, ed., Marcel Dekker, New York, 1996, pp. 63–76.

[8] G. Díaz and R. Letelier, *Explosive solutions of quasilinear elliptic equations: Existence and uniqueness*, Nonlinear Anal., 20 (1993), pp. 97–125.

[9] J. García-Melián and J. D. Rossi, *Boundary blow-up solutions to elliptic systems of competitive type*, J. Differential Equations, 206 (2004), pp. 156–181.

[10] J. García-Melián and A. Suárez, *Existence and uniqueness of positive large solutions to some cooperative elliptic systems*, Adv. Nonlinear Stud., 3 (2003), pp. 193–206.

[11] M. Ghergu and V. D. Rădulescu, *Explosive solutions of semilinear elliptic systems with gradient term*, RACSAM Rev. R. Acad. Cien. Exactas Fis. Nat. Ser. A. Mat., 97 (2003), pp. 467–475.

[12] M. W. Hirsch, *Systems of differential equation that are competitive or cooperative.* V. *Convergence in* 3*-dimensional systems*, J. Differential Equations, 80 (1989), pp. 94–106.

[13] M. W. Hirsch, *Systems of differential equations that are competitive or cooperative.* IV: *Structural stability in three-dimensional systems*, SIAM J. Math. Anal., 21 (1990), pp. 1225–1234.

[14] S. Kamin and L. Véron, *Existence and uniqueness of the very singular solution of the porous media equation with absorption*, J. Anal. Math., 51 (1988), pp. 245–258.

[15] J. B. Keller, *On solutions of* $\Delta u = f(u)$, Comm. Pure Appl. Math., 10 (1957), pp. 503–510.

[16] A. V. Lair and A. W. Wood, *Existence of entire large positive solutions of semilinear elliptic systems*, J. Differential Equations, 164 (2000), pp. 380–394.

[17] P. J. McKenna, W. Reichel, and W. Walter, *Symmetry and multiplicity for nonlinear elliptic differential equations with boundary blow-up*, Nonlinear Anal., 28 (1997), pp. 1213–1225.

[18] L. Markus, *Asymptotically autonomous differential systems*, in Contributions to the Theory of Nonlinear Oscillations 3, Ann. of Math. Stud. 36, S. Lefschetz, ed., Princeton University Press, Princeton, NJ, 1956, pp. 17–29.

[19] K. Mischaikow, H. Smith, and H. R. Thieme, *Asymptotically autonomous semiflows: Chain recurrence and Lyapunov functions*, Trans. Amer. Math. Soc., 347 (1995), pp. 1669–1685.

[20] R. Osserman, *On the inequality* $\Delta u \geq f(u)$, Pacific J. Math., 7 (1957), pp. 1641–1647.

[21] P. H. Rabinowitz, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.

[22] H. L. Smith, *Monotone Dynamical Systems: An Introduction to the Theory of Competitive and Cooperative Systems*, Math. Surveys Monogr. 41, AMS, Providence, RI, 1995.

[23] J. L. Vázquez and M. Walías, *Existence and uniqueness of solutions of diffusion-absorption equations with general data*, Differential Integral Equations, 7 (1994), pp. 15–36.

[24] L. Véron, *Semilinear elliptic equations with uniform blow-up on the boundary*, J. Anal. Math., 59 (1992), pp. 231–250.

[25] W. Walter, *Differential and Integral Inequalities*, Springer-Verlag, Berlin, 1970.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 514–530

A MODEL FOR THE OPTIMAL PLANNING OF AN URBAN AREA*

GIUSEPPE BUTTAZZO† AND FILIPPO SANTAMBROGIO‡

Abstract. We propose a model to describe the optimal distributions of residents and services in a prescribed urban area. The cost functional takes into account the transportation costs (according to a Monge–Kantorovich-type criterion) and two additional terms which penalize concentration of residents and dispersion of services. The tools we use are the Monge–Kantorovich mass transportation theory and the theory of nonconvex functionals defined on measures.

Key words. urban planning, mass transportation, nonconvex functionals over measures

AMS subject classifications. 49J45, 49K99, 90B06

DOI. 10.1137/S0036141003438313

1. Introduction.

The efficient planning of a city is a tremendously complicated problem, both for the high number of parameters which are involved as well as for the several relations which intervene among them (price of the land, kind of industries working in the area, quality of the life, prices of transportations, geographical obstacles, etc.). Perhaps a careful description of the real situations could be only obtained through evolution models which take into account the dynamical behavior of the different parameters involved.

An interesting mathematical model for the description of the equilibrium structure of a city is presented by Carlier and Ekeland in [3], where Monge–Kantorovich optimal transport theory plays an important role.

In the present paper we consider a geographical area as given, and we represent it through a subset Ω of $\mathbb{R}^n$ ($n = 2$ in the applications to concrete urban planning problems). We want to study the optimal location in Ω of a mass of inhabitants, which we denote by μ, as well as of a mass of services (working places, stores, offices, etc.), which we denote by ν. We assume that μ and ν are probability measures on Ω. This means that the total amounts of population and production are fixed as problem data, and this is a difference from the model in [3]. The measures μ and ν represent the unknowns of our problem that have to be found in such a way that a suitable total cost functional $\mathfrak{F}(\mu, \nu)$ is minimized. The definition of this total cost functional takes into account some criteria we want the two densities μ and ν to satisfy:

(i) there is a transportation cost for moving from the residential areas to the services areas;
(ii) people do not want to live in areas where the density of population is too high;
(iii) services need to be concentrated as much as possible in order to increase efficiency and decrease management costs.

*Received by the editors December 4, 2003; accepted for publication (in revised form) November 29, 2004; published electronically October 27, 2005. This work is part of the European Research Training Network "Homogenization and Multiple Scales"(HMS2000) under contract HPRN-2000-00109.

http://www.siam.org/journals/sima/37-2/43831.html

†Dipartimento di Matematica, Università di Pisa, via Buonarroti 2, 56127 Pisa, Italy (buttazzo@dm.unipi.it). The first author also acknowledges the support of the project "Problemi di Ottimizzazione in Teoria del Trasporto ed Applicazioni a Problemi di Pianificazione Urbana" of the Italian GNAMPA and of the project "Calcolo delle Variazioni" of the Italian Ministry of Education.

‡Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy (f.santambrogio@sns.it).

Fact (i) will be described through a Monge–Kantorovich mass transportation model; the transportation cost will indeed be given by using a p-Wasserstein distance ($p \geq 1$). We set

$$T_p(\mu,\nu) = W_p^p(\mu,\nu) = \inf_\gamma \left(\int_{\Omega\times\Omega} |x-y|^p \gamma(dx,dy) \right), \tag{1.1}$$

where the infimum is taken over all possible transport plans γ between μ and ν (i.e., probabilities on the product space having μ, ν as marginal measures). We refer to [7] for the whole theory on mass transportation. When $p = 1$ we are in the classical Monge case, and for this particular case we refer to [1] and [5].

Fact (ii) will be described by a penalization functional, a kind of total unhappiness of citizens due to high density of population, obtained by integrating with respect to the citizens' density their personal unhappiness.

Fact (iii) is modeled by a third term representing costs for managing services once they are located according to the distribution ν, taking into account that efficiency depends strongly on how much ν is concentrated.

The cost functional we will consider is then

$$\mathfrak{F}^p(\mu,\nu) = T_p(\mu,\nu) + F(\mu) + G(\nu), \tag{1.2}$$

and thus the optimal location of μ and ν will be determined by the minimization problem

$$\min \left\{ \mathfrak{F}^p(\mu,\nu) \ : \ \mu,\nu \text{ probabilities on } \Omega \right\}. \tag{1.3}$$

In this way, our model takes into consideration only the optimization of a total welfare parameter of the city, disregarding the effects on each single citizen. In particular, no equilibrium condition is considered. This may appear as a fault in the model, since the personal welfare of the citizens (depending on the population density of their zone and on the cost of moving from home to services) could be nonconstant. As a consequence, nonstable optimal solutions may occur, where some citizens would prefer to move elsewhere in the city in order to get better conditions. However, this is not the case, since our model also disregards prices of land and houses in the city, since they do not affect the total wealth of the area. It turns out that by a proper, market-determined choice of prices, welfare differences could be compensated and equilibrium recovered. This fact turns out to be a major difference between our model and the model in [3], both for the importance given in [3] to the variable represented by the price of land and for the fact that Carlier and Ekeland specifically look for an equilibrium solution instead of an optimal one.

The present paper, after this introduction, contains three sections. Section 2 is devoted to presenting precise choices for the functionals F and G and justifying them as reasonable choices. In the same section we also give a simple existence result for an optimal solution (μ,ν) as a starting point for the rest of the paper. In section 3 we consider the functional on μ obtained by keeping the measure ν as fixed: in this case the functional G does not play any role, and we obtain a convex minimization problem, which is interesting in itself. We also obtain some necessary optimality conditions in the very general case where no assumption is taken on the fixed measure ν. In section 4 we apply these results to the case where G is of the particular form presented in section 2, which forces ν to be atomic (i.e., services are concentrated in countably many points of the city area Ω). In the case where Ω is bounded, we give a quite precise description of the solution (μ, ν), and then we give an existence result also for the case $\Omega = \mathbb{R}^n$.

Both in the case $\Omega = \mathbb{R}^n$ and Ω bounded, optimal choices for μ and ν are given by the formation of a certain number of subcities, which are circular areas with a pole of services in the center (an atom for the measure ν) around which the population is distributed with a decreasing radial density.

Since we have considered only a very simplified model, our goal is neither to suggest a realistic way to design the ideal city nor to describe in a variational way the formation of existing cities. Nevertheless, from the analysis of our optimality results (and in particular from the subcities phenomena we referred to), we can infer some conclusions.

- Our model is not a proper choice to describe the shape of a single existing city, since the delocalization of services we find in an optimal solution does not reflect what reality suggests (in fact, we find finitely many disjoint, independent subcities with services only in the center).
- Our model is likely to be more realistic on a larger scale, when Ω represents a large urban area composed of several cities: in this case every atom of the optimal ν stands for the center of one of them and includes a complex system of services, located downtown, whose complexity cannot be seen in this scale.
- In our model the concentrated measure ν gives a good representation of the areas where services are offered to citizens and not of areas where commodities are produced (factories), due to the assumption that no land is actually occupied by the service poles (since ν is atomic).
- We do not believe that our model may actually be used to plan a future city or to improve the efficiency of an existing one, as a consequence of its oversimplified nature. However, we do not exclude the possibility of using it in the planning of less complex agglomerations, such as tourist villages, university campuses, etc.
- We conclude by stressing that the same model may be applied as a first simplified approach to other kinds of problems, where we have to choose in some efficient way the distributions of two different parameters, the first spread and the second concentrated, keeping them as close as possible to each other in some mass transportation sense.

2. The model. We now define the three terms appearing in our functional $\mathfrak{F}^p$. We must go through the definition of F and G, since the first term will be a Monge–Kantorovich transport cost, as explained in the previous section. For the functional F we take

$$F(\mu) = \begin{cases} \int_\Omega f(u(x))\,dx & \text{if } \mu = u \cdot \mathcal{L}^n,\ u \in L^1(\Omega), \\ +\infty & \text{otherwise,} \end{cases} \tag{2.1}$$

where the integrand $f : [0, +\infty] \to [0, +\infty]$ is assumed to be lower semicontinuous and convex, with $f(0) = 0$ and superlinear at infinity, that is,

$$\lim_{t \to +\infty} \frac{f(t)}{t} = +\infty. \tag{2.2}$$

In this form we have a local semicontinuous functional on measures. Without loss of generality, by subtracting constants to the functional F, we can suppose $f'(0) = 0$. Due to the assumption $f(0) = 0$, the ratio $f(t)/t$ is an incremental ratio of the convex function f and thus it is increasing in t. Then, if we write the functional F as

$$\int_\Omega \frac{f(u(x))}{u(x)} u(x)\, dx,$$

we can see the quantity $f(u)/u$, which is increasing in u, as the unhappiness of a single citizen when he lives in a place where the population density is u. Integrating it with respect to $\mu = u \cdot \mathcal{L}^n$ gives a quantity to be seen as the total unhappiness of the population.

As far as the concentration term $G(\nu)$ is concerned, we set

$$G(\nu) = \begin{cases} \sum_{i=0}^{\infty} g(a_i) & \text{if } \nu = \sum_{i=0}^{\infty} a_i \delta_{x_i}, \\ +\infty & \text{if } \nu \text{ is not atomic.} \end{cases} \tag{2.3}$$

We require the function g to be subadditive, lower semicontinuous, and such that $g(0) = 0$ and

$$\lim_{t \to 0} \frac{g(t)}{t} = +\infty. \tag{2.4}$$

Every single term $g(a_i)$ in the sum in (2.3) represents the cost of building and managing a service pole of size a_i, located at the point $x_i \in \Omega$.

In our model, as already pointed out, we fix as a datum the total production of services; moreover, in each service pole the production is required as a quantity proportionally depending on its size (or on the number of inhabitants making use of such a pole). We may define the productivity P of a pole of mass (size) a as the ratio between the production and the cost to get such a production. Then we have $P(a) = a/g(a)$ and

$$\sum_{i=0}^{\infty} g(a_i) = \sum_{i=0}^{\infty} \frac{a_i}{P(a_i)}.$$

As a consequence of assumption (2.4) we have that the productivity in very small service poles is near 0.

Notice that in the functional G we do not take into account distances between service poles. It would be interesting to consider nonlocal functionals involving such distances, taking into account possible cooperation and the consequent gain in efficiency. Those functionals could be matters of investigation in a subsequent paper, in which the results shown in the next section (since they do not depend on the choice of G) could be useful as well as in the present setting.

For the problem introduced in (1.3), existence results are straightforward, especially when we use as an environment a compact set Ω. In fact, functionals of the form of both F and G have been studied in a general setting by Bouchitté and Buttazzo in [2], and lower semicontinuity results were proven.

THEOREM 2.1. *Suppose Ω is compact, $p \geq 1$, and f and g satisfy the conditions listed above. Then the minimization problem* (1.3) *has at least one solution.*

Proof. By the direct method of calculus of variations, this result is an easy consequence of the weak-* compactness of the space $\mathcal{P}(\Omega)$, the space of probability measures on Ω when Ω itself is compact, and of the weak-* semicontinuity of the functional $\mathfrak{F}^p$. The second and third terms in (1.2) are, in fact, local semicontinuous functionals (due to results in [2]), while the first term is nothing but a Wasserstein distance raised to a certain power. Since it is known that in compact spaces this distance metrizes the weak-* topology, T_p is actually continuous. ◻

In [6], where we first presented the model, other existence results were shown. For instance, the case of a noncompact bounded convex set $\Omega \subset \mathbb{R}^n$ was considered. We will not go through this proof here and will discuss just one existence result in a noncompact setting, obtained as a consequence of a proper use of the optimality conditions presented in the next section.

3. A necessary condition of optimality. In this section we find optimality conditions for probability measures on Ω minimizing the functional

$$\mathfrak{F}^p_\nu(\mu) = T_p(\mu,\nu) + F(\mu).$$

It is clear that if (μ,ν) is an optimal pair for the whole functional $\mathfrak{F}^p$, μ is a minimizer for $\mathfrak{F}^p_\nu$. The goal of this section is to derive optimality conditions for $\mathfrak{F}^p_\nu$, for any ν, without any link to the minimization of $\mathfrak{F}^p$. The main part of the section will be devoted to presenting an approach obtained by starting with the easier case $p > 1$ and ν "regular" in some sense and then recovering the general case by an approximation argument. The reason for doing so relies on some conditions ensuring uniqueness properties of the Kantorovich potential. Similar approximation arguments were also used in [6]: purely atomic probability measures (i.e., finite sums of Dirac masses) were considered first, and then, by approximation, the result was extended to any measure ν. At the end of the section we also provide a sketch of a different proof, suggested to us by an anonymous referee, which is based on some convex analysis tools and strongly uses the convex structure of the problem.

For simplicity, let us call domains those sets which are the closure of a nonempty connected open subset of $\mathbb{R}^n$ with negligible boundary. From now on Ω will be a bounded domain and its diameter will be denoted by D. The function f in (2.1) will be assumed to be strictly convex and C^1, and we will denote by k the continuous, strictly increasing function $(f')^{-1}$. Strict convexity of f will ensure uniqueness for the minimizer of $\mathfrak{F}^p_\nu$.

LEMMA 3.1. *If μ is optimal for $\mathfrak{F}^p_\nu$, then for any other probability measure μ_1 with density u_1 such that $\mathfrak{F}^p_\nu(\mu_1) < +\infty$, the following inequality holds:*

$$T_p(\mu_1,\nu) - T_p(\mu,\nu) + \int_\Omega f'(u(x))[u_1(x) - u(x)]dx \geq 0.$$

Proof. For any $\varepsilon > 0$, due to the convexity of the transport term, it holds that

$$\begin{aligned} T_p(\mu,\nu) + F(\mu) &\leq T_p(\mu + \varepsilon(\mu_1 - \mu)) + F(\mu + \varepsilon(\mu_1 - \mu),\nu) \\ &\leq T_p(\mu,\nu) + \varepsilon(T_p(\mu_1,\nu) - T_p(\mu,\nu)) + F(\mu + \varepsilon(\mu_1 - \mu)). \end{aligned}$$

Therefore the quantity $T_p(\mu_1,\nu) - T_p(\mu,\nu) + \varepsilon^{-1}\left[F(\mu + \varepsilon(\mu_1 - \mu)) - F(\mu)\right]$ is nonnegative. If we let $\varepsilon \to 0$, we obtain the thesis if we prove

$$\lim_{\varepsilon\to 0} \int \frac{f(u + \varepsilon(u_1 - u)) - f(u)}{\varepsilon}\, d\mathcal{L}^n = \int f'(u)(u_1 - u)\, d\mathcal{L}^n.$$

By using the monotonicity of the incremental ratios of convex functions we can see that, for $\varepsilon < 1$,

$$\frac{|f(u + \varepsilon(u_1 - u)) - f(u)|}{\varepsilon} \leq |f(u) - f(u_1)|.$$

This is sufficient in order to apply Lebesgue dominated convergence theorem, since $F(\mu)$ and $F(\mu_1)$ are finite. ∎

LEMMA 3.2. *Let us suppose $\nu = \nu^s + v \cdot \mathcal{L}^n$, with $v \in L^\infty(\Omega)$, $\nu^s \perp \mathcal{L}^n$, $v > 0$ a.e. in Ω. If μ is optimal for $\mathfrak{F}^p_\nu$, then $u > 0$ a.e. in Ω.*

Proof. The lemma will be proven by contradiction. We will find, if the set $A = \{u = 0\}$ is not negligible, a measure μ_1 for which Lemma 3.1 is not verified. Let N be a Lebesgue-negligible set where ν^s is concentrated and t is an optimal transport map between μ and ν. Such an optimal transport exists, since $\mu \ll \mathcal{L}^n$. A proof of this fact can be found in [7] as long as we deal with the case $p > 1$, while for $p = 1$ we refer to [1].

Let $B = t^{-1}(A)$. Up to modifying t on the μ-negligible set A, we may suppose $B \cap A = \emptyset$. Set $\mu_1 = 1_{B^c} \cdot \mu + 1_{A\setminus N} \cdot \nu$; it is a probability measure with density u_1 given by $1_{B^c}u + 1_A v = 1_{B^c\setminus A}u + 1_A v$ (this equality comes from $u = 0$ on A). We have

$$F(\mu_1) = \int_{B^c\setminus A} f(u)\, d\mathcal{L}^n + \int_A f(v)\, d\mathcal{L}^n \le F(\mu) + ||f(v)||_\infty |\Omega| < +\infty.$$

Setting

$$t^*(x) = \begin{cases} t(x) & \text{if } x \in (A \cup B)^c, \\ x & \text{if } x \in (A \cup B), \end{cases}$$

we can see that t^* is a transport map between μ_1 and ν. In fact, for any Borel set $E \subset \Omega$, we may express $(t^*)^{-1}(E)$ as the disjoint union of $E \cap A$, $E \cap B$, and $t^{-1}(E) \cap B^c \cap A^c$. Thus,

$$\begin{aligned} \mu_1((t^*)^{-1}(E)) &= \nu(E \cap A) + \nu(E \cap B \cap A) + \mu(t^{-1}(E) \cap B^c \cap A^c) \\ &= \nu(E \cap A) + \mu(t^{-1}(E \cap A^c)) = \nu(E), \end{aligned}$$

where we used the fact that $A \cap B = \emptyset$ and that A^c is a set of full measure for μ. Consequently,

$$T_p(\mu_1, \nu) \le \int_{(A\cup B)^c} |x - t(x)|^p u(x) dx < \int_\Omega |x - t(x)|^p u(x) dx = T_p(\mu, \nu). \tag{3.1}$$

From this it follows that for μ_1 Lemma 3.1 is not satisfied, since the integral term $\int_\Omega f'(u)(u_1 - u) d\mathcal{L}^n$ is nonpositive, because $u_1 > u$ only on A, where $f'(u)$ vanishes. The strict inequality in (3.1) follows from the fact that if $\int_{A\cup B} |x - t(x)|^p u(x) dx = 0$, then for a.e. $x \in B$ it holds $u(x) = 0$ or $x = t(x)$, which, by definition of B, implies $x \in A$; in both cases we are led to $u(x) = 0$. This would give $\nu(A) = \mu(B) = 0$, contradicting the assumptions $|A| > 0$ and $v > 0$ a.e. in Ω. ∎

We need some results from duality theory in mass transportation that can be found in [7]. In particular, we point out the notation of c-transform (a kind of generalization of the well-known Legendre transform): given a function χ on Ω we define its c-transform (or c-conjugate function) by

$$\chi^c(y) = \inf_{x\in\Omega} c(x, y) - \chi(x).$$

We will generally use $c(x, y) = |x - y|^p$.

THEOREM 3.3. *Under the same hypotheses of Lemma 3.2, assuming also that $p > 1$, if μ is optimal for $\mathfrak{F}^p_\nu$ and we denote by ψ the unique, up to additive constants,* *Kantorovich potential for the transport between μ and ν, there exists a constant l such that the following relation holds:*

$$u = k(l - \psi) \text{ a.e. in } \Omega. \tag{3.2}$$

Proof. Let us choose an arbitrary measure μ_1 with bounded density u_1 (so that $F(\mu_1) < +\infty$) and define $\mu_\varepsilon = \mu + \varepsilon(\mu_1 - \mu)$. Let us denote by ψ_ε a Kantorovich potential between μ_ε and ν, chosen so that all the functions ψ_ε vanish at a same point. We can use the optimality of μ to write

$$T_p(\mu_\varepsilon, \nu) + F(\mu_\varepsilon) - T_p(\mu, \nu) - F(\mu) \geq 0.$$

By means of the duality formula, as $T_p(\mu_\varepsilon, \nu) = \int \psi_\varepsilon d\mu_\varepsilon + \int \psi_\varepsilon^c d\nu$ and $T_p(\mu, \nu) \geq \int \psi_\varepsilon d\mu + \int \psi_\varepsilon^c d\nu$, we can write

$$\int \psi_\varepsilon d(\mu_\varepsilon - \mu) + F(\mu_\varepsilon) - F(\mu) \geq 0.$$

Recalling that $\mu_\varepsilon - \mu = \varepsilon(\mu_1 - \mu)$ and that

$$F(\mu_\varepsilon) - F(\mu) = \int \left(f(u + \varepsilon(u_1 - u)) - f(u)\right) d\mathcal{L}^n,$$

we can divide by ε and pass to the limit. We know from Lemma 3.4 that ψ_ε converge towards the unique Kantorovich potential ψ for the transport between μ and ν. For the limit of the F part we use Lebesgue dominated convergence, as in Lemma 3.1. We then obtain at the limit

$$\int_\Omega (\psi(x) + f'(u(x)))(u_1(x) - u(x))dx \geq 0.$$

This means that for every probability μ_1 with bounded density u_1 we have

$$\int (\psi(x) + f'(u(x)))u_1(x)\, dx \geq \int (\psi(x) + f'(u(x)))u(x)\, dx.$$

Define first $l = \operatorname{ess\,inf}_{x \in \Omega} \psi(x) + f'(u(x))$. The left-hand side, by properly choosing u_1, can be made as close to l as we want. Then we get that the function $\psi + f'(u)$, which is $\mathcal{L}^n$-a.e., and so also μ-a.e., greater than l, integrated with respect to the probability μ gives a result less than or equal to l. It follows that

$$\psi(x) + f'(u(x)) = l, \quad \mu\text{-a.e. } x \in \Omega.$$

Together with the fact that, by Lemma 3.2, $u > 0$ a.e., we get an equality valid $\mathcal{L}^n$-a.e., and so it holds that

$$f'(u) = l - \psi.$$

We can then compose with k and get the thesis. $\square$

To establish Lemma 3.4, which we used in the proof of Theorem 3.3, we have first to point out the following fact. In the transport between two probabilities, if we look at the cost $c(x, y) = |x - y|^p$ with $p > 1$, there exists just one Kantorovich potential, up to additive constants, provided the absolutely continuous part of one of the measures has strictly positive density a.e. in the domain Ω.

LEMMA 3.4. *Let ψ_ε be Kantorovich potentials for the transport between $\mu_\varepsilon = \mu + \varepsilon(\mu_1 - \mu)$ and ν, all vanishing at a same point $x_0 \in \Omega$. Suppose that $\mu = u \cdot \mathcal{L}^n$ and $u > 0$ a.e. in Ω, and let ψ be the unique Kantorovich potential between μ and ν vanishing at the same point; then ψ_ε converge uniformly to ψ.*

Proof. First, notice that the family $(\psi_\varepsilon)_\varepsilon$ is equicontinuous since any c-concave function with respect to the cost $c(x,y) = |x-y|^p$ is pD^{p-1}-Lipschitz continuous (and Kantorovich potentials are optimal c-concave functions in the duality formula). Moreover, thanks to $\psi_\varepsilon(x_0) = 0$, we also get equiboundedness and thus, by the Ascoli–Arzelà theorem, the existence of uniform limits up to subsequences. Let $\overline{\psi}$ be one of these limits, arising from a certain subsequence. From the optimality of ψ_ε in the duality formula for μ_ε and ν we have, for any c-concave function φ,

$$\int \psi_\varepsilon \, d\mu_\varepsilon + \int \psi_\varepsilon^c \, d\nu \geq \int \varphi \, d\mu_\varepsilon + \int \varphi^c \, d\nu.$$

We want to pass to the limit as $\varepsilon \to 0$: we have uniform convergence of ψ_ε but we need uniform convergence of ψ_ε^c as well. To get it, note that

$$\psi_\varepsilon^c(x) = \inf_y |x-y|^p - \psi_\varepsilon(y), \quad \overline{\psi}^c(x) = \inf_y |x-y|^p - \overline{\psi}(y),$$
$$|\psi_\varepsilon^c(x) - \overline{\psi}^c(x)| \leq ||\psi_\varepsilon - \overline{\psi}||_\infty.$$

Passing to the limit as $\varepsilon \to 0$ along the considered subsequence we get, for any φ,

$$\int \overline{\psi} \, d\mu + \int \overline{\psi}^c \, d\nu \geq \int \varphi \, d\mu + \int \varphi^c \, d\nu.$$

This means that $\overline{\psi}$ is a Kantorovich potential for the transport between μ and ν. Then, taking into account that $\overline{\psi}(x_0) = 0$, we get the equality $\overline{\psi} = \psi$. Then we derive that the whole sequence converges to ψ. ◻

We now highlight that the relation we have proved in Theorem 3.3 enables us to choose a density u which is continuous. Moreover, it is also continuous in a quantified way, since it coincides with k composed with a Lipschitz function with a fixed Lipschitz constant. As a next step we will try to extend such results to the case of general ν and then to the case $p = 1$. The uniform continuity property we proved will be essential for an approximation process.

In order to go through our approximation approach, we need the following lemma, requiring the well-known theory of Γ-convergence. For all details about this theory, we refer to [4].

LEMMA 3.5. *Given a sequence $(\nu_h)_h$ of probability measures on Ω, supposing $\nu_h \rightharpoonup \nu$ and $p > 1$, it follows that the sequence of functionals $(\mathfrak{F}^p_{\nu_h})_h$ Γ-converges to the functional $\mathfrak{F}^p_\nu$ with respect to weak-$*$ topology on $\mathcal{P}(\Omega)$. Moreover, if ν is fixed and we let p vary, we have Γ-convergence, according to the same topology, of the functionals $\mathfrak{F}^p_\nu$ to the functional $(\mathfrak{F}^1_\nu)$ as $p \to 1$.*

Proof. For the first part of the statement, just notice that the Wasserstein distance is a metrization of weak-$*$ topology: consequently, since $T_p(\mu,\nu) = W_p^p(\mu,\nu)$, as $\nu_h \rightharpoonup \nu$ we have uniform convergence of the continuous functionals $T_p(\cdot, \nu_h)$. This implies Γ-convergence and pointwise convergence. In view of Proposition 6.25 in [4], concerning Γ-convergence of sums, we achieve the proof. The second assertion follows the same scheme once we notice that, for each $p > 1$ and every pair (μ,ν) of probability measures, it holds that

$$W_1(\mu,\nu) \leq W_p(\mu,\nu) \leq D^{1-\frac{1}{p}} W_1^{\frac{1}{p}}(\mu,\nu).$$

This gives uniform convergence of the transport term, as

$$\begin{aligned} T_p(\mu,\nu) - T_1(\mu,\nu) &\le (D^{p-1}-1)T_1(\mu,\nu) \\ &\le D(D^{p-1}-1) \to 0. \\ T_p(\mu,\nu) - T_1(\mu,\nu) &\ge T_1^p(\mu,\nu) - T_1(\mu,\nu) \\ &\ge (p-1)c(T_1(\mu,\nu)) \ge \bar{c}\,(p-1) \to 0, \end{aligned}$$

where $c(t) = t\log t$, $\bar{c} = \inf c$, and we used the fact $T_1(\mu,\nu) \le D$. ◻

We now state in the form of lemmas two extensions of Theorem 3.3

LEMMA 3.6. *Suppose $p > 1$ and fix an arbitrary $\nu \in \mathcal{P}(\Omega)$. If μ is optimal for $\mathfrak{F}^p_\nu$, then there exists a Kantorovich potential ψ for the transport between μ and ν such that* (3.2) *holds.*

Proof. We choose a sequence $(\nu_h)_h$ approximating ν in such a way that each ν_h satisfies the assumptions of Theorem 3.3. By Lemma 3.5 and the properties of Γ-convergence, the space $\mathcal{P}(\Omega)$ being compact and the functional $\mathfrak{F}^p_\nu$ having an unique minimizer (see, for instance, Chapter 7 in [4]), we get that $\mu_h \rightharpoonup \mu$, where each μ_h is the unique minimizer of $\mathfrak{F}^p_{\nu_h}$. Each measure μ_h is absolutely continuous with density u_h. We use (3.2) to express u_h in terms of Kantorovich potentials ψ_h and get uniform continuity estimates on u_h. We would like to extract converging subsequences by the Ascoli–Arzelà theorem, but we also need equiboundedness. We may obtain this by using together the integral bound $\int u_h d\mathcal{L}^n = \int k(-\psi_h) d\mathcal{L}^n = 1$ and the equicontinuity. So, up to subsequences, we have the following situation:

$$\begin{gathered} \mu_h = u_h \cdot \mathcal{L}^n, \qquad u_h = k(-\psi_h), \\ u_h \to u, \qquad \psi_h \to \psi \text{ uniformly}, \\ \mu_h \rightharpoonup \mu, \quad \mu = u \cdot \mathcal{L}^n, \qquad \nu_h \rightharpoonup \nu, \end{gathered}$$

where we have absorbed the constants l into the Kantorovich potentials. Clearly it is sufficient to prove that ψ is a Kantorovich potential between μ and ν to reach our goal.

To see this, we consider that for any c-concave function φ, it holds that

$$\int \psi_h \, d\mu_h + \int \psi_h^c \, d\nu_h \ge \int \varphi \, d\mu_h + \int \varphi^c \, d\nu_h.$$

The thesis follows passing to the limit with respect to h, as in Lemma 3.4. ◻

The next step is proving the same relation when ν is generic and $p = 1$. We are in the same situation as before, and we simply need approximation results on Kantorovich potentials in the more difficult situation when the cost functions $c_p(x,y) = |x-y|^p$ vary with p.

LEMMA 3.7. *Suppose $p = 1$ and fix an arbitrary $\nu \in \mathcal{P}(\Omega)$. If μ is optimal for $\mathfrak{F}^1_\nu$, then there exists a Kantorovich potential ψ for the transport between μ and ν with cost $c(x,y) = |x-y|$ such that* (3.2) *holds.*

Proof. For any $p > 1$ we consider the functional $\mathfrak{F}^p_\nu$ and its unique minimizer μ_p. Thanks to Lemma 3.6 we get the existence of densities u_p and Kantorovich potential ψ_p between μ_p and ν with respect to the cost c_p, such that

$$\mu_p = u_p \cdot \mathcal{L}^n, \quad u_p = k(-\psi_p).$$

By the Ascoli–Arzelà compactness result, as usual, we may suppose, up to subsequences,

$$u_p \to u, \quad \psi_p \to \psi \text{ uniformly},$$

and due to the Γ-convergence result in Lemma 3.5, since $\mathfrak{F}^1_\nu$ has an unique minimizer denoted by μ we also get

$$\mu_p \rightharpoonup \mu, \qquad \mu = u \cdot \mathcal{L}^n.$$

As in Lemma 3.6, we simply need to prove that ψ is a Kantorovich potential between μ and ν for the cost c_1. The limit function ψ is Lipschitz continuous with Lipschitz constant less than or equal to $\liminf_{p\to 1} pD^{p-1} = 1$, since it is approximated by ψ_p. Consequently ψ is c-concave for $c = c_1$. We need to show that it is optimal in the duality formula.

Let us recall that, for any real function φ and any cost function c, it holds that $\varphi^{cc} \geq \varphi$ and φ^{cc} is a c-concave function whose c-transform is $\varphi^{ccc} = \varphi^c$. Consequently, by the optimality of ψ_p, we get

$$\int \psi_p d\mu_p + \int \psi_p^{c_p} d\nu \geq \int \varphi^{c_p c_p} d\mu_p + \int \varphi^{c_p} d\nu \geq \int \varphi d\mu_p + \int \varphi^{c_p} d\nu. \tag{3.3}$$

We want to pass to the limit in the inequality between the first and the last term. We start by proving that, for an arbitrary sequence $(\varphi_p)_p$, if $\varphi_p \to \varphi_1$, we have the uniform convergence $\varphi_p^{c_p} \to \varphi_1^{c_1}$. Let us take into account that we have uniform convergence on bounded sets of $c_p(x,y) = |x-y|^p$ to $c_1(x,y) = |x-y|$. Then we have

$$\varphi_p^{c_p}(x) = \inf_y |x-y|^p - \varphi_p(y), \quad \varphi_1^{c_1}(x) = \inf_y |x-y| - \varphi_1(y),$$
$$|\varphi_p^{c,p}(x) - \varphi_1^{c,1}(x)| \leq ||c_p - c_1||_\infty + ||\varphi_p - \varphi_1||_\infty,$$

which gives us the convergence we needed. We then obtain, passing to the limit as $p \to 1$ in (3.3),

$$\int \psi \, d\mu + \int \psi^{c_1} \, d\nu \geq \int \varphi \, d\mu + \int \varphi^{c_1} \, d\nu.$$

By restricting this inequality to all φ which are c_1-concave, we get that ψ is a Kantorovich potential for the transport between μ and ν and the cost c_1. □

We can now state the main theorem of this section, whose proof consists only of putting together all the results we have obtained above.

THEOREM 3.8. *Let Ω be a bounded domain in $\mathbb{R}^n$, f be a C^1 strictly convex function, $p \geq 1$, and ν be a probability measure on Ω. Then there exists a unique measure $\mu \in \mathcal{P}(\Omega)$ minimizing $\mathfrak{F}^p_\nu$ and it is absolutely continuous with density u. Moreover, there exists a Kantorovich potential ψ for the transport between μ and ν and the cost $c(x,y) = |x-y|^p$ such that $u = k(-\psi)$ holds, where $k = (f')^{-1}$.*

Consequences on the regularity of u come from this expression, which gives Lipschitz-type continuity, and from the relationship between Kantorovich potentials and optimal transport, which can be expressed through some PDEs. It is not difficult, for instance, in the case $p = 2$, to obtain a Monge–Ampère equation for the density u.

As we have already mentioned, we provide a sketch of an alternative proof to Theorem 3.8. The idea of such a proof consists of looking at the subdifferential of the functional $\mathfrak{F}^p_\nu$ in order to get optimality conditions on the unique minimizer measure μ and its density u (here we will identify any absolutely continuous probability measure with its density).

Sketch of Proof. Step 1. Consider the minimizing probability μ with density $u \in L^1(\Omega)$ and define the vector space $X = span\,(L^\infty(\Omega), \{u\})$, with dual

$$X' = \left\{ \xi \in L^1(\Omega) : \int_\Omega |\xi| u \, d\mathcal{L}^n < +\infty \right\}.$$

Then we consider the minimization problem for the functional H defined on X by

$$H(v) = \begin{cases} \mathfrak{F}^p_\nu(v) & \text{if } v \in \mathcal{P}(\Omega), \\ +\infty & \text{otherwise.} \end{cases}$$

It is clear that u minimizes H. We will prove

$$\partial H(u) = \left\{ f'(u) + \psi : \psi \text{ maximizes } \int_\Omega \phi d\mu + \int_\Omega \phi^c d\nu \text{ for } \phi \in X' \right\} \tag{3.4}$$

and then consider as an optimality condition $0 \in \partial H(u)$. The subdifferential ∂H of the convex functional H is to be considered in the sense of the duality between X and X'. Notice that, in this setting, the c-transform ϕ^c of a function $\phi \in X'$ has to be defined replacing the inf with an ess inf. Finally, in order to achieve the proof, it is sufficient to recognize that for a function ψ attaining the maximum in the duality formula, it necessarily holds that $\psi = \psi^{cc}$ a.e. on $\{u > 0\}$ and that this, together with $0 = f'(u) + \psi$, implies $\psi = \psi^{cc} \wedge 0$. This means that ψ is an optimal c-concave function (since it is expressed as an infimum of two c-concave functions) in the duality formula between μ and ν, and so it is a Kantorovich potential. In this way the thesis of Theorem 3.8 is achieved, provided (3.4) is proved.

Step 2. By using the same computations as in Lemma 3.1, for any $u_1 \in X \cap \mathcal{P}(\Omega)$, if we set $u_\varepsilon = u + \varepsilon(u_1 - u)$, we may prove that

$$\lim_{\varepsilon \to 0} \frac{F(\mu_\varepsilon) - F(\mu)}{\varepsilon} = \int_\Omega f'(u)(u_1 - u) \, d\mathcal{L}^n.$$

Notice that, since $\int_\Omega f'(u)|u_1 - u| \, d\mathcal{L}^n < +\infty$, by choosing $u_1 = 1/|\Omega|$ it follows that $f'(u)$ and $f'(u)u$ are L^1 functions; i.e., $f'(u) \in X'$. Then it is possible to prove that this implies $\partial H(u) = f'(u) + \partial T(u)$, where T is the convex functional $T_p(\cdot, \nu)$.

Step 3. It remains to prove that

$$\partial T(u) = \left\{ \psi : \psi \text{ maximizes } \int_\Omega \phi d\mu + \int_\Omega \phi^c d\nu \text{ for } \phi \in X' \right\}. \tag{3.5}$$

In fact, if we define $K(\phi) = \int_\Omega \phi^c d\nu$, the key point is to prove that K is concave and upper semicontinuous in ϕ. Then, by standard convex analysis tools, (3.5) is a consequence of the equality $T(v) = \sup_\phi v \cdot \phi + K(\phi)$, where $v \cdot \phi$ stands for the duality product between X and X' and equals $\int_\Omega v\phi \, d\mathcal{L}^n$.

4. Applications to urban planning problems (with atomic services). In this section we go through the consequences that Theorem 3.8 has in the problem of minimizing $\mathfrak{F}^p$, when this functional is built by using a term G as in (2.3), which forces the measure ν, representing services, to be purely atomic. We have two goals: trying to have an explicit expression for u in the case of a bounded domain Ω and proving an existence result in the case $\Omega = \mathbb{R}^n$.

THEOREM 4.1. *Suppose* (μ,ν) *is optimal for problem* (1.3)*. Suppose also that the function* g *is locally Lipschitz in* $]0,1]$*: then* ν *has finitely many atoms and is of the form* $\nu=\sum_{i=1}^m a_i\delta_{x_i}$*.*

Proof. It is clear that ν is purely atomic, i.e., a countable sum of Dirac masses. We want to show their finiteness. Consider $a=\max a_i$ (such a maximum exists since $\lim_i a_i=0$ and $a_i>0$) and let L be the Lipschitz constant of g on $[a,1]$. Now consider an atom with mass a_i and modify ν by moving its mass onto the atom x_j whose mass a_j equals a, obtaining a new measure ν'. The G-part of the functional decreases, while it may happen that the transport part increases. Since we do not change μ, the F-part remains the same. By optimality of ν we get $T_p(\mu,\nu)+G(\nu)\le T_p(\mu,\nu')+G(\nu')$ and thus

$$g(a_i)-La_i\le g(a_i)+g(a)-g(a+a_i)\le T_p(\mu,\nu')-T_p(\mu,\nu)\le a_iD.$$

This implies

$$\frac{g(a_i)}{a_i}\le D+L,$$

and by the assumption on the behavior of g at 0, this gives a lower bound δ on a_i. Since we have proved that every atom of ν has a mass greater than δ, we may conclude that ν has finitely many atoms. ∎

Now we can use the results from last section.

THEOREM 4.2. *For any* $\nu\in\mathcal{P}(\Omega)$ *such that* ν *is purely atomic and composed by finitely many atoms at the points* $x_1,\dots,x_m$*, if* μ *minimizes* $\mathfrak{F}^p_\nu$*, there exist constants* c_i *such that*

$$u(x)=k\left((c_1-|x-x_1|^p)\vee\cdots\vee(c_m-|x-x_m|^p)\vee 0\right). \tag{4.1}$$

In particular the support of u *is the intersection with* Ω *of a finite union of balls centered around the atoms of* ν*.*

Proof. On the Kantorovich potential ψ appearing in Theorem 3.8, we know that

$$\begin{aligned}\psi(x)+\psi^c(y)&=|x-y|^p\quad\forall(x,y)\in spt(\gamma),\\ \psi(x)+\psi^c(y)&\le|x-y|^p\quad\forall(x,y)\in\Omega\times\Omega,\end{aligned}$$

where γ is an optimal transport plan between μ and ν. Taking into account that ν is purely atomic we obtain, defining $c_i=\psi^c(x_i)$,

$$\begin{aligned}-\psi(x)&=c_i-|x-x_i|^p\quad\mu\text{-a.e. }x\in\Omega_i,\\ -\psi(x)&\ge c_i-|x-x_i|^p\quad\forall x\in\Omega,\,\forall i,\end{aligned}$$

where $\Omega_i=t^{-1}(x_i)$, where t is an optimal transport map between μ and ν. Since $\mu-$a.e. point in Ω is transported to a point x_i, we know that $u=0$ a.e. in the complement of $\bigcup_i\Omega_i$. Since, by $f'(u)=-\psi$, it holds that $-\psi(x)\ge 0$, one gets that everywhere in Ω the function $-\psi$ is greater than each of the terms $c_i-|x-x_i|^p$ and 0, while a.e. it holds equality with at least one of them. By changing u on a negligible set, one obtains (4.1). The support of μ, consequently, turns out to be composed of the union of the intersection with Ω of the balls $B_i=B(x_i,c_i^{1/p})$. ∎

Theorem 4.2 allows us to have an almost explicit formula for the density of μ. Formula (4.1) becomes more explicit when the balls B_i are disjoint. We now give a sufficient condition on ν under which this fact occurs.

LEMMA 4.3. *There exists a positive number $\overline{R}$, depending on the function k, such that any of the balls B_i has a radius not exceeding $\overline{R}$. In particular, for any atomic probability ν such that the distance between any two of its atoms is larger than $2\overline{R}$, the balls B_i are disjoint.*

Proof. Set $R_i = c_i^{1/p}$ and notice that

$$1 = \int_\Omega u \geq \int_{B_i} k(c_i - |x - x_i|^p)\, dx = \int_0^{R_i} k(R_i^p - r^p) n\omega_n r^{n-1}\, dr,$$

where the number ω_n stands for the volume of the unit ball in $\mathbb{R}^n$. This inequality gives the required upper bound on R_i, since

$$\int_0^{R_i} k(R_i^p - r^p) n\omega_n r^{n-1}\, dr \geq C \int_0^{R_i - 1} n r^{n-1}\, dr = C(R_i - 1)^n. \qquad \square$$

When the balls B_i are disjoint, we have $B_i = \Omega_i$ for every i and we get a simple relation between radii and masses corresponding to each atom. The constants c_i can then be found by using $R_i = c_i^{1/p}$. In fact, by imposing the equality of the mass of μ in the ball and of ν in the atom, the radius $R(m)$ corresponding to a mass m satisfies

$$m = \int_0^{R(m)} k(R(m)^p - r^p) n\omega_n r^{n-1} dr. \tag{4.2}$$

For instance, if $f(s) = s^2/2$, we have

$$R(m) = \left(\frac{m(n+p)}{\omega_n p} \right)^{\frac{1}{n+p}}.$$

The second aim of this section is to obtain an existence result for the problem (1.3) when $\Omega = \mathbb{R}^n$. A difference from the bounded case is the fact that we must look for minimization among all pairs of measures in $\mathcal{W}_p(\mathbb{R}^n)$, the p-th Wasserstein metric space (i.e., the space of measures $\lambda \in \mathcal{P}(\mathbb{R}^n)$ such that $\int |x|^p \lambda(dx) < +\infty$, endowed with the distance W_p), rather than in $\mathcal{P}(\mathbb{R}^n)$.

We start with some simple results about the minimization problem for $\mathfrak{F}_\nu^p$.

LEMMA 4.4. *For every fixed $\nu \in \mathcal{P}(\mathbb{R}^n)$ there exists a (unique if f is strictly convex) minimizer μ for $\mathfrak{F}_\nu^p$: it belongs to $\mathcal{W}_p(\mathbb{R}^n)$ if and only if $\nu \in \mathcal{W}_p(\mathbb{R}^n)$, and if ν does not belong to this space, the functional $\mathfrak{F}_\nu^p$ is infinite on the whole $\mathcal{W}_p(\mathbb{R}^n)$. Moreover, if ν is compactly supported, the same happens for μ.*

Proof. The existence of μ comes from the direct method of the calculus of variations and the fact that if $(T_p(\mu_h, \nu))_h$ is bounded, then $(\mu_h)_h$ is tight. The behavior of the functional with respect to the space $\mathcal{W}_p(\mathbb{R}^n)$ is trivial. Finally, the last assertion can be proved by contradiction, supposing $\mu(B(0,R)^c) > 0$ for every $R < +\infty$ and replacing μ with

$$\mu_R = 1_{B_R} \cdot \mu + \frac{\mu(B_R^c)}{|B_r|} 1_{B_r} \cdot \mathcal{L}^n,$$

where $B(0,r)$ is a ball containing the support of ν. By optimality, we should have

$$T_p(\mu_R, \nu) + F(\mu_R) \geq T_p(\mu, \nu) + F(\mu), \tag{4.3}$$

but we have

$$T_p(\mu_R,\nu) - T_p(\mu,\nu) \le -((R-r)^p - (2r)^p)\mu(B_R^c), \tag{4.4}$$

$$F(\mu_R) - F(\mu) \le \int_{B_r} \left[f\left(u + \frac{\mu(B_R^c)}{|B_r|} \right) - f(u) \right] d\mathcal{L}^n. \tag{4.5}$$

By summing up (4.4) and (4.5), dividing by $\mu(B_R^c)$, and taking into account (4.3), we get

$$-((R-r)^p - (2r)^p) + \frac{1}{\mu(B_R^c)} \int_{B_r} \left[f\left(u + \frac{\mu(B_R^c)}{|B_r|} \right) - f(u) \right] d\mathcal{L}^n \ge 0. \tag{4.6}$$

Yet, by passing to the limit as $R \to +\infty$ and $\mu(B_R^c) \to 0$, the first term in (4.6) tends to $-\infty$, while the second is decreasing as $R \to +\infty$. This last one tends to $\int_{B_r} f'(u) d\mathcal{L}^n$, provided it is finite for at least a value of R (which ensures the finiteness of the limit as well). To conclude, it is sufficient to prove that

$$\int_{B_r} \left[f\left(u + \frac{\mu(B_R^c)}{|B_r|} \right) - f(u) \right] d\mathcal{L}^n < +\infty.$$

This is quite easy in the case $f(z) = Az^q$ with $q > 1$, while for general f the assertion comes from the fact that u is continuous on $\overline{B_r}$ and hence bounded. If $u = 0$ a.e. in B_r, this is trivial; otherwise take the probability measures $\mu' = 1_{B_r}/\mu(B_r)\cdot\mu$ and $\nu' = t_\sharp \mu'$ for an optimal transport map t between μ and ν. It is clear that μ' minimizes $\mathfrak{F}^p_{\nu'}$ in the new domain $\Omega' = \overline{B_r}$. Then we may apply Theorem 3.8 and get the continuity of its density, which ensures the continuity of u on $\overline{B_r}$. ∎

To go through our proof we need to manage minimizing sequences, in the sense of the following lemma.

LEMMA 4.5. *It is possible to choose a minimizing sequence $((\mu_h,\nu_h))_h$ in $\mathcal{W}_p(\mathbb{R}^n)\times \mathcal{W}_p(\mathbb{R}^n)$ such that for every h the measure ν_h is finitely supported, and the density of μ_h is given by* (4.1)*, with disjoint balls centered at the atoms of ν_h.*

Proof. First we start from an arbitrary minimizing sequence $((\mu'_h,\nu'_h))_h$. Then we approximate each ν'_h in $\mathcal{W}_p$ by a finite support measure ν''_h. To do this we truncate the sequence of its atoms and move the mass in excess to the origin. In this way, we have $G(\nu''_h) \le G(\nu'_h)$, by the subadditivity of g, while the value of the transport term increases by an arbitrary small quantity. Consequently, $((\mu'_h,\nu''_h))_h$ is still a minimizing sequence. Then we replace μ'_h by μ''_h, chosen in such a way that it minimizes $\mathfrak{F}^p_{\nu''_h}$. By Lemma 4.4, each μ''_h has a compact support. Then we translate every atom of each ν''_h, together with its own set Ω_i, to some disjoint sets Ω_i^*. In this way we get new measures μ'''_h and ν'''_h. The value of the functional in this step has not changed. We may choose to place the atoms of each ν'''_h so far from each other that each distance between atoms is at least $2\overline{R}$. Then we minimize again in μ, getting a new sequence of pairs $((\mu''''_h,\nu'''_h))_h$, and we set $\nu_h = \nu'''_h$ and $\mu_h = \mu''''_h$. Thanks to Theorem 4.2 and Lemma 4.3 the requirements of the thesis are fulfilled. ∎

It is clear now that if one can obtain a uniform estimate on the number of atoms of the measures ν_h, the existence problem is easily solved. In fact we already know that each ball belonging to the support of μ_h is centered at an atom of ν_h and has a radius not larger than $\overline{R}$. Provided we are able to prove an estimate like $\sharp\,\{\text{atoms of } \nu_h\} \le N$, it would be sufficient to act by translation on the atoms and their corresponding balls, obtaining a new minimizing sequence (the value of $\mathfrak{F}^p$ does not change) with supports all contained in a same bounded set (for instance, the ball $B_{N\overline{R}}$).

We now try to give sufficient conditions in order to find minimizing sequences where the number of atoms stays bounded. Notice that on sequences of the form given by Lemma 4.5, the functional $\mathfrak{F}^p$ has the expression

$$\mathfrak{F}^p(\mu_h,\nu_h)=\sum_{i=1}^{k(h)} E(m_{i,h}),\quad \text{if } \nu_h=\sum_{i=1}^{k(h)} m_{i,h}\delta_{x_{i,h}}, \tag{4.7}$$

where the quantity $E(m)$ is the total contribute given by an atom with mass m to the functional. We may compute

$$E(m) \;=\; g(m)+\int_0^{R(m)}\left[f(k(R(m)^p-r^p))+k(R(m)^p-r^p)r^p\right]n\omega_n r^{n-1}dr, \tag{4.8}$$

taking into account the particular form of the density in the ball.

THEOREM 4.6. *Let us suppose $f\in C^2((0,+\infty))$, and $g\in C^2((0,1])\cap C^0([0,1])$, in addition to all previous assumptions. Then the minimization problem for $\mathfrak{F}^p$ in $\mathcal{W}_p(\mathbb{R}^n)\times\mathcal{W}_p(\mathbb{R}^n)$ has a solution, provided*

$$\limsup_{R\to 0^+} g''\left(\int_0^R k(R^p-r^p)n\omega_n r^{n-1}dr\right)\int_0^R k'(R^p-r^p)n\omega_n r^{n-1}dr<-1.$$

Proof. According to what has been previously proven, it is sufficient to produce a minimizing sequence of the form of Lemma 4.5 with a bounded number of atoms. We claim that it is enough to prove that the function E is subadditive on an interval $[0,m_0]$. In fact, having proven it, we start from a sequence $((\mu_h,\nu_h))_h$ built as in Lemma 4.5 and use the characterization of $\mathfrak{F}^p$ given in (4.7). Then we modify our sequence by replacing in each ν_h any pair of atoms of mass less than $m_0/2$ with a single atom with the sum of the masses. We keep atoms far away from each other in order to use (4.7). We may perform such a replacement as far as we find more than one atom whose mass is less than or equal to $m_0/2$. At the end we get a new pair $((\mu'_h,\nu'_h))_h$, where the number of atoms of ν'_h is less than $N=1+\lfloor 2/m_0\rfloor$. The value of the functional $\mathfrak{F}^p$ has not increased, thanks to the subadditivity of E on $[0,m_0]$.

Taking into account that $E(0)=0$ and that concave functions vanishing at 0 are subadditive, we look at concavity properties of the function E in an interval $[0,m_0]$. It is sufficient to compute the second derivative of E and find it negative in a neighborhood of the origin.

By means of the explicit formula (4.8) and also taking into account (4.2), setting $E(m)=g(m)+K(R(m))$, we start by computing dK/dr. Using the facts that $f'\circ k=id$ and that $k(0)=0$, we can obtain the formula

$$\frac{dK(R(m))}{dm}(m)=R(m)^p.$$

From another derivation and some standard computation we finally obtain

$$E''(m)=g''(m)+\frac{1}{\int_0^{R(m)}k'(R(m)^p-r^p)n\omega_n r^{n-1}dr}.$$

The assumption of this theorem ensures that such a quantity is negative for small m, and so the proof is achieved. $\square$

Remark 4.7. Notice that when the functions f and g are of the form $f(t) = at^q$, $q > 1$, $g(t) = bt^r$, $r < 1$, with a and b positive constants, it holds that

$$g''\left(\int_0^R k(R^p - r^p) n\omega_n r^{n-1} dr\right) \le -CR^{(n+\frac{p}{q-1})(r-2)},$$

$$\int_0^R k'(R^p - r^p) n\omega_n r^{n-1} dr \le CR^{n+p\frac{2-q}{q-1}},$$

and so the lim sup in Theorem 4.6 may be estimated from above by

$$\lim_{R\to 0^+} -CR^{\frac{p}{q-1}(r-q)+n(r-1)} = -\infty.$$

Consequently the assumption in Theorem 4.6 is always verified when f and g are power functions.

Remark 4.8. From the proof of the existence theorem it is clear that there exists a minimizing pair $(\mu, \nu) \in \mathcal{W}_p(\mathbb{R}^n) \times \mathcal{W}_p(\mathbb{R}^n)$ where ν has finitely many atoms and μ is supported in a finite, disjoint union of balls centered at the atoms of ν and contained in a bounded domain Ω_0, with a density given by Theorem 4.2. The same happens if we look for the minimizers in a bounded domain Ω, provided Ω is large enough to contain Ω_0, and hence a solution to the problem in $\mathbb{R}^n$. For example, all the open sets containing N balls of radius $\overline{R}$ admit a minimizing solution supported in disjoint balls.

We conclude by stressing the fact that in order to solve the problem in $\mathbb{R}^n$, we have only to look at the function E and find out the number of atoms and their respective masses $(m_i)_{i=1\dots k}$. The problem to solve is then

$$\min\left\{\sum_{i=1}^k E(m_i) \,:\, k \in \mathbb{N}, \sum_{i=1}^k m_i = 1\right\}. \tag{4.9}$$

Typically, for instance when f and g are power functions, the function E involved in (4.9) is a concave-convex function, as sketched in Figure 1. Due to such a concave-convex behavior, in general it is not clear whether the values of the numbers m_i solving (4.9) and representing subcities' sizes are all equal or may be different.

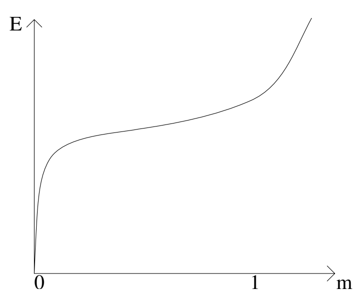


FIG. 1. *Typical behavior of E.*

Acknowledgments. The authors wish to thank an anonymous referee for pointing out the alternative proof sketched in section 3.

REFERENCES

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

[2] G. Bouchitté and G. Buttazzo, *New lower semicontinuity results for nonconvex functionals defined on measures*, Nonlinear Anal., 15 (1990), pp. 679–692.

[3] G. Carlier and I. Ekeland, *The structure of cities*, J. Global Optim., 29 (2004), pp. 371–376.

[4] G. Dal Maso, *An Introduction to Γ-Convergence*, Birkhäuser, Basel, Switzerland, 1992.

[5] L. C. Evans and W. Gangbo, *Differential equations methods for the Monge–Kantorovich mass transfer problems*, Mem. Amer. Math. Soc., 137 (1999), number 653.

[6] F. Santambrogio, *Misure ottime per costi di trasporto e funzionali locali*, Laurea Thesis, Università di Pisa, Pisa, 2003; also available online from http://www.unipi.it/etd and cvgmt.sns.it.

[7] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math., AMS, Providence, RI, 2003.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 531–548

STEADY STATES FOR A COAGULATION-FRAGMENTATION EQUATION WITH VOLUME SCATTERING*

PHILIPPE LAURENÇOT† AND CHRISTOPH WALKER‡

Abstract. A coagulation-fragmentation equation including volume scattering and collisional breakage is considered. We prove that the equation admits steady states of arbitrary mass provided that the kernels satisfy some suitable growth conditions. On the other hand, we also show that zero is the only steady state in particular cases.

Key words. coagulation-fragmentation equation, volume scattering, stationary solution

AMS subject classifications. 45G10

DOI. 10.1137/S0036141004444111

1. Introduction. The aim of this paper is to investigate the existence of steady states for a coagulation-fragmentation equation including a volume scattering effect. Recall that coagulation-fragmentation models describe the time evolution of a system consisting of a very large number of particles, which can either coalesce to form larger particles or split into smaller ones. Usually, these particles are supposed to be identified only by their size (mass, volume), which in the conventional continuous models might be any positive real number. In the model considered in the present paper, a critical particle size $y_0 \in (0,\infty)$ is introduced beyond which no particle can survive. This feature requires an additional mechanism, called *scattering* in what follows, preventing the occurrence of particles of size larger than the maximal size y_0 [6]. In addition to this scattering phenomenon, the subsequent model also includes the possibility of collisional breakage.

Denoting by $f(t,y) \geq 0$ the density of particles of size $y \in Y := (0,y_0)$ at time $t \geq 0$ (per unit volume), the evolution of the system of particles simultaneously undergoing coagulation and fragmentation can be described by the equation

$$\begin{aligned} \partial_t f &= L(f), \quad (t,y) \in (0,\infty)\times Y, \\ f(0,y) &= f_0(y), \quad y \in Y, \end{aligned} \tag{1.1}$$

where f_0 is a given initial distribution. The reaction terms $L(f) := L_{\mathsf{b}}(f) + L_{\mathsf{c}}(f) + L_{\mathsf{s}}(f)$ are defined by

$$\begin{aligned} L_{\mathsf{b}}(f)(y) &:= \int_y^{y_0} \gamma(y',y)\ f(y')\ \mathrm{d}y' - f(y)\ \int_0^y \frac{y'}{y}\ \gamma(y,y')\ \mathrm{d}y', \\ L_{\mathsf{c}}(f)(y) &:= \frac{1}{2}\ \int_0^y K(y',y-y')\ P(y',y-y')\ f(y-y')\ f(y')\ \mathrm{d}y' \\ &\quad + \frac{1}{2}\int_y^{y_0}\int_0^{y'} K(y'',y'-y'')\ Q(y'',y'-y'')\ \beta_{\mathsf{c}}(y',y)\ f(y'')\ f(y'-y'')\ \mathrm{d}y''\mathrm{d}y' \end{aligned}$$

*Received by the editors May 24, 2004; accepted for publication (in revised form) February 28, 2005; published electronically October 27, 2005.
http://www.siam.org/journals/sima/37-2/44411.html

†Mathématiques pour l'Industrie et la Physique, CNRS UMR 5640, Université Paul Sabatier–Toulouse 3, 118 route de Narbonne, F-31062 Toulouse cedex 4, France (laurenco@mip.ups-tlse.fr).
‡Department of Mathematics, Vanderbilt University, 1326 Stevenson Center, Nashville, TN 37240 (cwalker@math.vanderbilt.edu).

$$
\begin{aligned}
&\quad - f(y) \int_0^{y_0-y} K(y,y')\ \left\{P(y,y')+Q(y,y')\right\}\ f(y')\ \mathrm{d}y', \\
L_{\mathsf{s}}(f)(y) &:= \frac{1}{2} \int_{y_0}^{2y_0} \int_{y'-y_0}^{y_0} K(y'',y'-y'')\ \beta_{\mathsf{s}}(y',y)\ f(y'')\ f(y'-y'')\ \mathrm{d}y''\mathrm{d}y' \\
&\quad - f(y) \int_{y_0-y}^{y_0} K(y,y')\ f(y')\ \mathrm{d}y'
\end{aligned}
$$

for $y \in Y$ and describe the following reactions:

- The linear operator $L_{\mathsf{b}}(f)$ accounts for the gain and loss of particles of size y due to multiple spontaneous breakage, where $\gamma(y,y') \geq 0$ denotes the rate at which a particle of size $y \in Y$ decays into a particle of size $y' \in (0,y)$.
- Furthermore, two particles y and y' with cumulative size $y+y' < y_0$ can collide at a rate $K(y,y') \geq 0$ and either nothing happens—meaning that the involved particles remain unchanged, for instance in the case of grazing particles—or they merge with probability $P(y,y')$ or, in the case of high-energy collisions, shatter with probability $Q(y,y')$ into several particles according to the shattering distribution $\beta_{\mathsf{c}}(y+y',y'')$ (the latter process is also referred to as collisional breakage). Consistency of the model then demands

$$0 \leq P(y,y')+Q(y,y') \leq 1, \quad y+y' < y_0. \tag{1.2}$$

  These processes are reflected by the operator $L_{\mathsf{c}}(f)$.
- Finally, the scattering operator $L_{\mathsf{s}}(f)$ represents the interaction of two particles y and y' with cumulative size beyond the maximal size y_0. They can coalesce but the resulting particle instantaneously splits into particles all with size within the admissible range Y. The daughter particles are then distributed according to $\beta_{\mathsf{s}}(y+y',y'') \geq 0$. We refer to [6], where the volume scattering mechanism was introduced (see also [5] for a more detailed discussion on the modelling issue).

Since there is no particle inlet or outlet, one intuitively expects the total mass to be preserved during time, i.e.,

$$\int_0^{y_0} y\ f(t,y)\ \mathrm{d}y = \int_0^{y_0} y\ f_0(y)\ \mathrm{d}y, \quad t \geq 0. \tag{1.3}$$

Provided that the shattering and the scattering processes are mass preserving (see assumptions (1.10) and (1.12)), this is indeed the case.

From a mathematical viewpoint, some properties of the coagulation-fragmentation equation with volume scattering (1.1) have been investigated recently; in particular, results concerning the well-posedness of (1.1) are to be found in [3, 6, 14, 15], while results on the large time behavior of the solutions in some cases have been obtained in [16]. As for numerical simulations, we refer to [12]. We also mention at this point that, formally, the classical coagulation-fragmentation model usually contemplated in the literature can be derived from (1.1) by putting $y_0 := \infty$ and $P \equiv 1$ (implying $Q \equiv 0$ according to (1.2)). In particular, the shattering and scattering terms vanish in this case. A survey of the present state of knowledge on the classical coagulation-fragmentation equations and references to further literature for this case can be found in [2, 11].

In the present paper, we will focus on existence of nontrivial steady states to (1.1), that is, on nonzero solutions to the equation

$$L(f) = 0 \quad \text{in } Y. \tag{1.4}$$

Addressing this issue is mainly motivated by the study of the asymptotic behavior of solutions to (1.1). Note that the equality (1.3) entails a natural side condition, namely, to solve (1.4) subject to

$$\varrho = \int_0^{y_0} y\ f(y)\ \mathrm{d}y, \tag{1.5}$$

where $\varrho > 0$ is a given positive real number. Let us point out right now that (1.4), (1.5) do not always possess solutions for $\varrho > 0$. In particular, if $\gamma \equiv 0$ and merely binary shattering and binary scattering are taken into account, zero is the only steady state (see section 4). So far, existence of solutions to (1.4) apart from zero is known only if the kernels satisfy an extended version of the so-called detailed balance condition [15, 16], namely that there exists $H \in L^1(Y)$ such that

$$\gamma(y+y',y)\ H(y+y')\ =\ P(y,y')\ K(y,y')\ H(y)\ H(y')$$

for $0 < y+y' < y_0$, that

$$\begin{aligned}&\beta_{\mathsf{c}}(y,y')\ Q(y'',y-y'')\ K(y'',y-y'')\ H(y'')\ H(y-y'')\\ &\quad =\ \beta_{\mathsf{c}}(y,y'')\ Q(y',y-y')\ K(y',y-y')\ H(y')\ H(y-y')\end{aligned}$$

for $0 < y+y'\ ,\ y+y'' < y_0$, and that

$$\beta_{\mathsf{s}}(y,y')\ K(y'',y-y'')\ H(y'')\ H(y-y'')\ =\ \beta_{\mathsf{s}}(y,y'')\ K(y',y-y')\ H(y')\ H(y-y')$$

for $0 < y-y_0 < y',y'' < y_0$. In this case, each function $f_\alpha(y) := H(y)\,\alpha^y$, $y \in Y$, with $\alpha \geq 0$ satisfies (1.4). For the classical coagulation-fragmentation equation, that is, (1.1) with $y_0 = \infty$ and $P \equiv 1$, this condition has previously been used in various papers (for instance, see [1, 7, 9, 10]) and the long-time behavior of solutions has been investigated. That the latter equation admits nontrivial and smooth steady state solutions of arbitrary mass without assuming the detailed balance condition has been proven in [8] for constant fragmentation kernels γ and coagulation kernels

$$K(y,y') = a + b(y+y'),\quad y,y' > 0,$$

with $a, b \geq 0$. More recently, existence of nontrivial stationary solutions (in a weak sense) is shown for kernels of the form

$$\gamma(y,y') = y^\sigma B(y'/y),\quad K(y,y') = y^\alpha (y')^\nu + y^\nu (y')^\alpha$$

with $-1 \leq \alpha \leq 0 \leq \nu \leq 1$, $\alpha+\nu \in [0,1)$, $\sigma \geq -1$, and some suitable function B [4]. To the best of our knowledge, these are the only available results on existence of steady states for the classical coagulation-fragmentation equation ($y_0 = \infty$, $P \equiv 1$) in the absence of the detailed balance condition.

As for (1.4), (1.5), no result seems to be known but the existence of steady states is strongly supported by the numerical simulations in [12]. In this paper, we identify a class of data $(\gamma, K, \beta_{\mathsf{c}}, \beta_{\mathsf{s}})$ for which (1.4), (1.5) has at least one solution for every $\varrho > 0$. Before stating precisely our assumptions, let us first outline the approach we employ to solve (1.4), (1.5): in such a situation, a natural tool is the Schauder fixed point theorem, but its application requires some strong compactness which is not likely to be available here. Indeed, the operator L is an integral operator and does not seem to be compact. To overcome this difficulty, we consider the regularised problem $-\varepsilon\, f'' = L(f)$ in Y with suitable boundary conditions, where f'' denotes the second derivative of f and $\varepsilon \in (0,1)$. It is then possible to use the Schauder fixed point theorem to establish the existence of a solution f_ε to this problem which satisfies (1.5). The next step is to show that (f_ε) is relatively weakly sequentially compact in $L^1(Y)$ and that its cluster points for this topology solve (1.4), (1.5). Let us further mention that another way to remedy to the lack of strong compactness properties of L has been developed in [4] and relies on the Tikhonov fixed point theorem which only demands weak compactness.

The assumptions made throughout this paper are as follows. We suppose that the coagulation kernel K belongs to $L^\infty(Y\times Y)$ and satisfies

$$K_\star (yy')^\sigma \le K(y,y') = K(y',y) \le K^\star (yy')^\sigma, \quad (y,y') \in Y\times Y, \tag{1.6}$$

for some $K_\star, K^\star > 0$, $\sigma \in [0,1]$, and the monotonicity condition

$$K(y-y',y') \le K(y,y') \quad \text{for } \ 0<y'<y<y_0. \tag{1.7}$$

The probabilities P and Q are nonnegative symmetric functions defined on

$$\Xi := \{(y,y') \in Y\times Y;\, y+y' < y_0\}$$

obeying (1.2), and there is $P_\star \in (0,1)$ such that

$$P(y,y') + 2\, Q(y,y') \ge P_\star > 0 \quad \text{for a.a. } \ (y,y') \in \Xi. \tag{1.8}$$

In addition, P satisfies the monotonicity condition

$$P(y-y',y') \le P(y,y') \quad \text{for a.a. } \ (y,y') \in \Xi \quad \text{with } \ 0<y'<y<y_0. \tag{1.9}$$

The fragmentation kernel γ and the shattering distribution β_{c} are nonnegative measurable functions defined on

$$\Delta := \{(y,y') \in Y\times Y;\, 0<y'<y<y_0\},$$

and shattering is supposed to be a mass preserving process, that is,

$$Q(y,y') \left(\int_0^{y+y'} y''\ \beta_{\mathsf{c}}(y+y',y'')\,\mathrm{d}y'' - y - y' \right) = 0 \tag{1.10}$$

for a.a. $(y,y') \in \Xi$. We also assume that shattering is suitably dominated by coagulation, i.e., we assume that there exist $z_0 \in Y$ and $\kappa_0 > 0$ with

$$Q(y,y') \int_0^{y+y'} \beta_{\mathsf{c}}(y+y',y'')\,\mathrm{d}y'' \ \le\ P(y,y') + 2\, Q(y,y') - \kappa_0 \tag{1.11}$$

for a.a. $y+y' < z_0$. The scattering kernel β_{s} is a nonnegative measurable function defined on $(y_0,2y_0)\times(0,y_0)$ and satisfies

$$\int_0^{y_0} y'\ \beta_{\mathsf{s}}(y,y')\,\mathrm{d}y' = y \quad \text{for a.a. } \ y \in (y_0,2y_0). \tag{1.12}$$

Finally, we suppose that there are $p>1$ and μ_γ, μ_{c}, $\mu_{\mathsf{s}} > 0$ such that

$$\int_0^y (y')^{\sigma(1-2p)}\ \gamma(y,y')^p\ \mathrm{d}y' \le \mu_\gamma \quad \text{for a.a. } \ y \in Y, \tag{1.13}$$

$$Q(y,y')\int_0^{y+y'} (y'')^{\sigma(1-p)}\ \beta_{\mathsf{c}}(y+y',y'')^p\ \mathrm{d}y'' \le \mu_{\mathsf{c}} \quad \text{for a.a.} \quad (y,y')\in\Xi, \tag{1.14}$$

$$\int_0^{y_0} (y')^{\sigma(1-p)}\ \beta_{\mathsf{s}}(y,y')^p\ \mathrm{d}y' \le \mu_{\mathsf{s}} \quad \text{for a.a.} \quad y\in(y_0,2y_0). \tag{1.15}$$

Note that (1.13)–(1.15) and Hölder's inequality imply

$$\int_0^{y} \gamma(y,y')\ \mathrm{d}y' \le\ m_\gamma\ y^\sigma \quad \text{for a.a.} \quad y\in Y, \tag{1.16}$$

$$Q(y,y')\int_0^{y+y'} \beta_{\mathsf{c}}(y+y',y'')\ \mathrm{d}y'' \le\ m_{\mathsf{c}} \quad \text{for a.a.} \quad (y,y')\in\Xi, \tag{1.17}$$

$$\int_0^{y_0} \beta_{\mathsf{s}}(y,y')\ \mathrm{d}y' \le\ m_{\mathsf{s}} \quad \text{for a.a.} \quad y\in(y_0,2y_0) \tag{1.18}$$

for some constants $m_\gamma, m_{\mathsf{c}}, m_{\mathsf{s}} > 0$.

Possible (and reasonable) choices of kernels obeying all of the assumptions above are as follows: suppose that K is of the form

$$K(y,y') := A + B(yy')^\theta + C(y+y')^\mu$$

with $A>0$ and $B,C,\theta,\mu\ge 0$. In particular, the positivity of A and $y_0\in(0,\infty)$ warrant that (1.6) holds true with $\sigma := 0$. Let P and Q be nonnegative functions such that, for some $\tau, q>0$,

$$Q(y,y') := q(y+y)^\tau, \quad y+y'<y_0,$$

and such that (1.2), (1.8), and (1.9) hold. For $\bar\gamma,\alpha>0$ and $0\ge\zeta,\xi,\nu>-1$ define

$$\begin{aligned}\gamma(y,y') &:= \bar\gamma\ y^{\alpha-\zeta-1}\ (y')^\zeta,\\ \beta_{\mathsf{c}}(y,y') &:= (\xi+2)\ y^{-1-\xi}\ (y')^\xi,\\ \beta_{\mathsf{s}}(y,y') &:= (\nu+2)\ y_0^{-2-\nu}\ y\ (y')^\nu.\end{aligned}$$

We may then choose $p\in(1,1+\tau)$ with $-1/p<\min\{\nu,\xi,\zeta,\alpha-1\}$ so that all assumptions are satisfied with $\sigma=0$.

On the other hand, if $K(y,y') := \tilde K(yy')^\sigma$, $\sigma\in(0,1]$, $\tilde K>0$, and if $P,Q,\gamma,\beta_{\mathsf{c}},\beta_{\mathsf{s}}$ are as above with additionally $\alpha>\sigma$ and $1+\zeta>\sigma$, we again find $p>1$ which is sufficiently close to 1 and such that (1.13)–(1.15) hold.

Our main result is the following theorem.

THEOREM 1.1. *Let* (1.6)–(1.15) *hold. Then, given any* $\varrho>0$, *there exists a nonnegative function* $f\in L^1_+(Y,y^\sigma\mathrm{d}y)$ *satisfying* $L(f)=0$ *a.e. in* Y *and*

$$\int_0^{y_0} y\ f(y)\ \mathrm{d}y = \varrho,$$

where $L^1_+(Y,y^\sigma\mathrm{d}y)$ *denotes the positive cone of* $L^1(Y,y^\sigma\mathrm{d}y)$.

Recall that the positive cone $L^1_+(Y,y^\sigma\mathrm{d}y)$ of $L^1(Y,y^\sigma\mathrm{d}y)$ is the set of functions of $L^1(Y,y^\sigma\mathrm{d}y)$ which are nonnegative almost everywhere in Y.

The solution f to (1.4), (1.5) we construct in Theorem 1.1 actually belongs to $L^p(Y,y^\sigma\mathrm{d}y)$, but it is yet unclear whether f enjoys additional regularity properties. Also, uniqueness of a solution to (1.4), (1.5) does not seem to be obvious.

As already mentioned, the main idea in order to prove this theorem is to consider first a parameter-dependent regularized problem, which can be solved by a Schauder fixed point argument, and to show afterwards that the family of solutions is weakly compact in $L^p(Y, y^\sigma \mathrm{d}y)$. This then guarantees the existence of nontrivial steady states. In the next section we state and solve the regularized problem. Subsequently, we derive in section 3 some uniform estimates leading to the desired weak compactness. In the concluding section, we show that problem (1.4), (1.5) does not necessarily have a solution.

2. A regularized problem: Existence. Note that (1.6), (1.16), (1.17), and (1.18) imply, for $f \in L^1(Y, y^\sigma \,\mathrm{d}y)$, that the reaction terms $L_{\mathsf{b}}(f)$, $L_{\mathsf{c}}(f)$, and $L_{\mathsf{s}}(f)$ belong to $L^1(Y)$. In addition, given $\psi \in L^\infty(Y)$, we have the identities (see [14, Lemma 2.6] or [15, Lemma 2.7])

$$\int_0^{y_0} \psi(y)\ L_{\mathsf{b}}(f)(y)\ \mathrm{d}y = \int_0^{y_0} \int_0^{y} \left[\psi(y') - \frac{y'}{y}\ \psi(y)\right]\ \gamma(y,y')\ \mathrm{d}y'\ f(y)\ \mathrm{d}y, \tag{2.1}$$

$$\int_0^{y_0} \psi(y)\ L_{\mathsf{c}}(f)(y)\ \mathrm{d}y = \frac{1}{2}\ \int_0^{y_0} \int_0^{y_0-y} \psi_{\mathsf{c}}(y,y')\ K(y,y')\ f(y)\ f(y')\ \mathrm{d}y'\mathrm{d}y, \tag{2.2}$$

$$\int_0^{y_0} \psi(y)\ L_{\mathsf{s}}(f)(y)\ \mathrm{d}y = \frac{1}{2}\ \int_0^{y_0} \int_{y_0-y}^{y_0} \psi_{\mathsf{s}}(y,y')\ K(y,y')\ f(y)\ f(y')\ \mathrm{d}y'\mathrm{d}y, \tag{2.3}$$

where

$$\begin{aligned}\psi_{\mathsf{c}}(y,y') &:= P(y,y')\ \psi(y+y') - [P(y,y')+Q(y,y')]\ [\psi(y)+\psi(y')]\\ &\qquad +\ Q(y,y')\ \int_0^{y+y'} \psi(y'')\ \beta_{\mathsf{c}}(y+y',y'')\ \mathrm{d}y'',\\ \psi_{\mathsf{s}}(y,y') &:= \int_0^{y_0} \psi(y'')\ \beta_{\mathsf{s}}(y+y',y'')\ \mathrm{d}y'' - \psi(y) - \psi(y').\end{aligned}$$

If $f \in L^q(Y; y^k \mathrm{d}y)$ for some $q \in [1,\infty)$ and $k \in \mathbb{R}$, we define

$$M_{k,q}(f) := \int_0^{y_0} y^k\ |f|^q(y)\ \mathrm{d}y \quad \text{and} \quad M_k(f) := M_{k,1}(f).$$

Given $\delta \in (0,1)$, we set

$$K_\delta(y,y') := K(y,y') + \delta$$

and notice that

$$\|K_\delta\|_\infty \le \|K\|_\infty + 1.$$

Hereafter, we denote by $L_\delta(f)$ the reaction terms $L(f)$ but with K_δ instead of K. For $\varepsilon \in (0,\delta)$ and $\varrho \in (0,\infty)$ we define

$$\begin{aligned}&\omega^2 := \frac{\|K_\delta\|_\infty}{\varepsilon} + m_\gamma\ y_0^\sigma\\ &\text{and}\\ &R := \frac{1}{4\,\varepsilon\,y_0}\left(6\ \varepsilon\ \varrho + 5\ \omega^2\ y_0^2\ \varrho + 3\ \|K_\delta\|_\infty\ y_0\ \varrho^2\right).\end{aligned} \tag{2.4}$$

We next introduce

$$F(f) := \varphi_\varepsilon(f)\ L_\delta(f) + \omega^2 f$$

for $f \in L^1_+(Y)$, where

$$\varphi_\varepsilon(f) := \frac{1}{1+\varepsilon\, M_0(f)},$$

and observe that $F(f)$ belongs to $L^1(Y)$. We then denote by u_f the unique solution in $W^{2,1}(Y)$ to the boundary-value problem

$$-\varepsilon\ u_f'' + \omega^2\ u_f = F(f) \quad \text{in} \quad Y, \tag{2.5}$$

$$u_f(0) = y_0\ u_f'(y_0) - u_f(y_0) = 0. \tag{2.6}$$

Finally, let $\mathcal{C}$ be the subset of $L^1(Y)$ defined by

$$\mathcal{C} := \left\{ f \in L^1_+(Y);\ M_1(f) = \varrho,\ M_0(f) \le R \right\}. \tag{2.7}$$

Clearly, $\mathcal{C}$ is a nonempty, bounded, and closed convex subset of $L^1(Y)$. In addition, we have the following property.

LEMMA 2.1. *If $f \in \mathcal{C}$, then $u_f \in \mathcal{C}$.*

Proof. Since $f \ge 0$, it follows from (1.2), (1.6), and (1.16) that

$$\begin{aligned} F(f)(y) &\ge \omega^2\ f(y) - \varphi_\varepsilon(f)\ f(y)\ \left(\int_0^y \frac{y'}{y}\ \gamma(y,y')\ \mathrm{d}y' + \int_0^{y_0} K_\delta(y,y')\ f(y')\ \mathrm{d}y' \right) \\ &\ge \omega^2\ f(y) - \varphi_\varepsilon(f)\ f(y)\ \left(m_\gamma\ y_0^\sigma + \|K_\delta\|_\infty\ M_0(f)\right) \\ &\ge 0. \end{aligned}$$

The comparison principle then entails that $u_f \ge 0$. We next readily infer from (1.10), (1.12), and (2.1)–(2.3) that

$$\int_0^{y_0} y\ F(f)(y)\ \mathrm{d}y = \omega^2\ M_1(f),$$

while (2.6) yields that

$$-\varepsilon \int_0^{y_0} y\ u_f''(y)\ \mathrm{d}y = 0.$$

Consequently, we deduce from (2.5) after multiplication by y and integration over Y the equality $M_1(u_f) = M_1(f)$.

We now multiply (2.5) by y^3 and integrate over Y. Observe first that

$$-\int_0^{y_0} y^3\ u_f''(y)\ \mathrm{d}y = 2\ y_0^2\ u_f(y_0) - 6\ M_1(u_f)$$

by (2.6), while (1.2) and (1.10) entail that, for $(y,y') \in \Xi$, the function $\left(y^3\right)_{\mathsf{c}}$ (defined in (2.2) with $\psi(y) = y^3$) satisfies

$$\begin{aligned} \left(y^3\right)_{\mathsf{c}}(y,y') &= P(y,y')\ \left[y^3 + (y')^3 + 3\ y^2\ y' + 3\ y\ (y')^2 - y^3 - (y')^3\right] \\ &\quad +\ Q(y,y')\ \left[\int_0^{y+y'} (y'')^3\ \beta_{\mathsf{c}}(y+y',y'')\ \mathrm{d}y'' - y^3 - (y')^3\right] \\ &\le 3\ P(y,y')\ y\ y'\ (y+y') + Q(y+y')\ \left[(y+y')^2\ (y+y') - y^3 - (y')^3\right] \\ &\le 3\ (P+Q)(y,y')\ y\ y'\ (y+y') \\ &\le 3\ y\ y'\ (y+y'). \end{aligned}$$

We also infer from (1.12) that, for $(y,y') \in Y \times Y \setminus \Xi$, the function $\left(y^3\right)_{\mathsf{s}}$ (defined in (2.3) with $\psi(y) = y^3$) satisfies

$$\begin{aligned}\left(y^3\right)_{\mathsf{s}}(y,y') &\le y_0^2 \int_0^{y+y'} y''\ \beta_{\mathsf{s}}(y+y',y'')\ \mathrm{d}y'' - y^3 - (y')^3 \\ &\le y_0^2\ (y+y')\ - y^3 - (y')^3 \\ &\le (y+y')^3 - y^3 - (y')^3 \\ &\le 3\ y\ y'\ (y+y').\end{aligned}$$

At last, we notice that, for $(y,y') \in \Delta$,

$$(y')^3 - \frac{y'}{y}\ y^3 = y'\ \left((y')^2 - y^2\right) \le 0.$$

Now, since $f \in \mathcal{C}$, it follows from (2.1) to (2.3) and the previous upper bounds that

$$\begin{aligned}&\int_0^{y_0} y^3\ F(f)(y)\ \mathrm{d}y \\ &\le \omega^2\ M_3(f) + \frac{3}{2}\ \varphi_\varepsilon(f)\ \int_0^{y_0} \int_0^{y_0-y} \left(y^2\ y' + y\ (y')^2\right)\ K_\delta(y,y')\ f(y)\ f(y')\ \mathrm{d}y'\mathrm{d}y \\ &+\ \frac{3}{2}\ \varphi_\varepsilon(f)\ \int_0^{y_0} \int_{y_0-y}^{y_0} \left(y^2\ y' + y\ (y')^2\right)\ K_\delta(y,y')\ f(y)\ f(y')\ \mathrm{d}y'\mathrm{d}y \\ &\le \omega^2\ y_0^2\ M_1(f) + \frac{3\ \|K_\delta\|_\infty}{2}\ \varphi_\varepsilon(f)\ \int_0^{y_0} \int_0^{y_0} \left(y^2\ y' + y\ (y')^2\right)\ f(y)\ f(y')\ \mathrm{d}y'\mathrm{d}y \\ &\le \omega^2\ y_0^2\ \varrho + 3\ \|K_\delta\|_\infty\ \varphi_\varepsilon(f)\ \int_0^{y_0} \int_0^{y_0} y^2\ y'\ f(y)\ f(y')\ \mathrm{d}y'\mathrm{d}y \\ &\le \omega^2\ y_0^2\ \varrho + 3\ \|K_\delta\|_\infty\ y_0\ \varrho^2.\end{aligned}$$

Therefore, we obtain

$$2\ \varepsilon\ y_0^2\ u_f(y_0) \le 6\ \varepsilon\ \varrho + \omega^2\ y_0^2\ \varrho + 3\ \|K_\delta\|_\infty\ y_0\ \varrho^2. \tag{2.8}$$

On the other hand, due to

$$\begin{aligned}&\int_0^{y_0} y^2\ F(f)(y)\ \mathrm{d}y \\ &\ge\ -\varphi_\varepsilon(f)\ \int_0^{y_0} \int_0^{y} y\ y'\ \gamma(y,y')\ \mathrm{d}y'\ f(y)\ \mathrm{d}y \\ &-\frac{\varphi_\varepsilon(f)}{2}\ \int_0^{y_0} \int_0^{y_0-y} (y^2 + y'^2)\ K_\delta(y,y')\ \left\{P(y,y') + Q(y,y')\right\}\ f(y')\ f(y)\ \mathrm{d}y'\mathrm{d}y \\ &-\frac{\varphi_\varepsilon(f)}{2}\ \int_0^{y_0} \int_{y_0-y}^{y_0} (y^2 + y'^2)\ K_\delta(y,y')\ f(y')\ f(y)\ \mathrm{d}y'\mathrm{d}y \\ &\ge\ -m_\gamma\ y_0^{1+\sigma}\ M_1(f) - \varrho\ y_0\ \|K_\delta\|_\infty\ \frac{M_0(f)}{1+\varepsilon\, M_0(f)}\end{aligned}$$

and

$$-\int_0^{y_0} y^2\ u_f''(y)\ \mathrm{d}y\ =\ y_0\ u_f(y_0) - 2\ M_0(u_f),$$

we deduce from (2.5) that

$$\varepsilon\ y_0\ u_f(y_0) - 2\ \varepsilon\ M_0(u_f) + \omega^2\ M_2(u_f) \geq -m_\gamma\ y_0^{1+\sigma}\ \varrho - \varrho\ y_0\ \frac{\|K_\delta\|_\infty}{\varepsilon}.$$

Consequently, taking into account (2.8) and the definition of ω, we end up with

$$2\ \varepsilon\ M_0(u_f) \leq \varepsilon\ y_0\ u_f(y_0) + \omega^2\ y_0\ M_1(u_f) + m_\gamma\ y_0^{1+\sigma}\ \varrho + \varrho\ y_0\ \frac{\|K_\delta\|_\infty}{\varepsilon} \leq 2\ \varepsilon\ R,$$

and the proof is complete. ∎

PROPOSITION 2.2. *There is a function $f \in \mathcal{C} \cap W^{2,1}(Y)$ such that*

$$-\varepsilon\ f'' = \varphi_\varepsilon(f)\ L_\delta(f) \quad \text{in} \quad Y, \tag{2.9}$$

$$f(0) = y_0\ f'(y_0) - f(y_0) = 0. \tag{2.10}$$

Proof. By Lemma 2.1, the mapping $f \longmapsto u_f$ maps $\mathcal{C}$ into itself. In addition, it is clearly a continuous and compact mapping from $\mathcal{C}$ into itself for the norm-topology of $L^1(Y)$. Indeed, we recall that, for $f \in L^1(Y)$, u_f is given by

$$u_f(y) = \left(\lambda - \int_0^y e^{-\bar\omega y'}\ F(f)(y')\ \mathrm{d}y'\right)\ \frac{e^{\bar\omega y}}{2\,\varepsilon\,\bar\omega}\ -\ \left(\lambda - \int_0^y e^{\bar\omega y'}\ F(f)(y')\ \mathrm{d}y'\right)\ \frac{e^{-\bar\omega y}}{2\,\varepsilon\,\bar\omega}$$

for $y \in [0, y_0]$, where $\bar\omega := \omega\ \varepsilon^{-1/2}$ and

$$\lambda := \vartheta\ \int_0^{y_0} e^{-\bar\omega y'}\ F(f)(y')\ \mathrm{d}y'\ +\ (1-\vartheta)\ \int_0^{y_0} e^{\bar\omega y'}\ F(f)(y')\ \mathrm{d}y'$$

with

$$\vartheta := \frac{y_0\ \bar\omega - 1}{y_0\ \bar\omega - 1 + (y_0\ \bar\omega + 1)\ e^{-2\bar\omega y_0}}.$$

In particular, there is a constant Γ depending on y_0 and ε such that

$$\|u_f\|_{W^{1,\infty}(Y)} \leq \Gamma\ \|F(f)\|_{L^1(Y)}.$$

Since F is a locally Lipschitz continuous map from $L^1(Y)$ into $L^1(Y)$ (see [14, Lemma 2.1]), the claimed continuity and compactness of $f \longmapsto u_f$ follow.

Now, since $\mathcal{C}$ is a nonempty, closed, and convex subset of $L^1(Y)$, we are in a position to apply the Schauder fixed point theorem and conclude that there is $f \in \mathcal{C}$ such that $u_f = f$. Proposition 2.2 readily follows. ∎

3. A regularized problem: Uniform estimates. For $\delta \in (0,1)$, $\varepsilon \in (0,\delta)$, and $\varrho > 0$, we denote by $f_{\varepsilon,\delta}$ the solution to (2.9), (2.10) given by Proposition 2.2. In particular, we have

$$\int_0^{y_0} y\ f_{\varepsilon,\delta}(y)\ \mathrm{d}y = \varrho. \tag{3.1}$$

The aim of this section is to prove that $(f_{\varepsilon,\delta})$ is weakly sequentially compact first with respect to ε and subsequently with respect to δ. In the following, we denote by C various positive constants which do neither depend on ε nor on δ. Dependence on δ, for instance, will be indicated explicitly by writing $C(\delta)$.

We first proceed as in Lemma 2.1 to bound $f_{\varepsilon,\delta}(y_0)$.

LEMMA 3.1. *For $\delta \in (0,1)$ and $\varepsilon \in (0,\delta)$, we have*

$$f'_{\varepsilon,\delta}(0) \geq 0 \quad \text{and} \quad \varepsilon\, f_{\varepsilon,\delta}(y_0) \leq C.$$

Proof. Clearly, $f'_{\varepsilon,\delta}(0) \geq 0$ since $f_{\varepsilon,\delta}(0) = 0$ and $f_{\varepsilon,\delta}(y) \geq 0$ for $y \in Y$. We next multiply (2.9) by y^3 and integrate over Y. As in the proof of Lemma 2.1, we use (2.10) and (3.1) to obtain

$$\varepsilon \left(2\ y_0^2\ f_{\varepsilon,\delta}(y_0) - 6\ \varrho\right) \leq 3\ \|K_\delta\|_\infty\ y_0\ \varrho^2 \leq 3\ (\|K\|_\infty + 1)\ y_0\ \varrho^2. \qquad \square$$

We next estimate the L^1-norm of $f_{\varepsilon,\delta}$ using a different argument than in the previous section.

LEMMA 3.2. *For $\delta \in (0,1)$ and $\varepsilon \in (0,\delta)$, we have*

$$\delta^{1/2}\ M_0(f_{\varepsilon,\delta}) + M_\sigma(f_{\varepsilon,\delta}) \leq C. \tag{3.2}$$

Proof. We integrate (2.9) over Y. We first notice that

$$-\varepsilon \int_0^{y_0} f''_{\varepsilon,\delta}(y)\ \mathrm{d}y \ = \ -\varepsilon\ \left(f'_{\varepsilon,\delta}(y_0) - f'_{\varepsilon,\delta}(0)\right) \geq -\varepsilon\ \frac{f_{\varepsilon,\delta}(y_0)}{y_0} \geq -C$$

by (2.10) and Lemma 3.1, while the function $(1)_{\mathsf{c}}$ (defined in (2.2) with $\psi(y) = 1$) satisfies

$$\begin{aligned}
(1)_{\mathsf{c}}(y,y') &= P(y,y') - 2\ (P+Q)(y,y') + Q(y,y') \int_0^{y+y'} \beta_{\mathsf{c}}(y+y',y'')\ \mathrm{d}y'' \\
&\leq -\kappa_0\ \mathbf{1}_{[0,z_0]}(y+y') + Q(y,y') \int_0^{y+y'} \beta_{\mathsf{c}}(y+y',y'')\ \mathrm{d}y''\ \mathbf{1}_{[z_0,y_0]}(y+y') \\
&\leq -\kappa_0\ \mathbf{1}_{[0,z_0]}(y+y') + m_{\mathsf{c}}\ \mathbf{1}_{[z_0,y_0]}(y+y')
\end{aligned}$$

for $(y,y') \in \Xi$ by (1.11) and (1.17), since P and Q are nonnegative. Moreover, the function $(1)_{\mathsf{s}}$ (defined in (2.3) with $\psi(y) = 1$) satisfies

$$(1)_{\mathsf{s}}(y,y') = \int_0^{y_0} \beta_{\mathsf{s}}(y+y',y'')\ \mathrm{d}y'' - 2 \leq m_{\mathsf{s}} - 2$$

for $(y,y') \in Y \times Y \setminus \Xi$ by (1.18). Because of the above inequalities, we deduce from (1.16), (2.1)–(2.3), and (2.9) that

$$\begin{aligned}
-\frac{C}{\varphi_\varepsilon(f_{\varepsilon,\delta})} &\leq -\frac{\varepsilon}{\varphi_\varepsilon(f_{\varepsilon,\delta})} \int_0^{y_0} f''_{\varepsilon,\delta}(y)\ \mathrm{d}y \\
&\leq \int_0^{y_0} \int_0^y \left(1 - \frac{y'}{y}\right)\ \gamma(y,y')\ \mathrm{d}y'\ f_{\varepsilon,\delta}(y')\ \mathrm{d}y \\
&\quad - \frac{\kappa_0}{2} \int_0^{y_0} \int_0^{y_0-y} \mathbf{1}_{[0,z_0]}(y+y')\ K_\delta(y,y')\ f_{\varepsilon,\delta}(y')\ f_{\varepsilon,\delta}(y)\ \mathrm{d}y'\mathrm{d}y \\
&\quad + \frac{1}{2}\ m_{\mathsf{c}} \int_0^{y_0} \int_0^{y_0-y} \mathbf{1}_{[z_0,y_0]}(y+y')\ K_\delta(y,y')\ f_{\varepsilon,\delta}(y')\ f_{\varepsilon,\delta}(y)\ \mathrm{d}y'\mathrm{d}y \\
&\quad + \frac{1}{2}\ (m_{\mathsf{s}} - 2) \int_0^{y_0} \int_{y_0-y}^{y_0} K_\delta(y,y')\ f_{\varepsilon,\delta}(y')\ f_{\varepsilon,\delta}(y)\ \mathrm{d}y'\mathrm{d}y \\
&\leq m_\gamma\ M_\sigma(f_{\varepsilon,\delta}) - \frac{\kappa_0}{2} \int_0^{y_0} \int_0^{y_0} K_\delta(y,y')\ f_{\varepsilon,\delta}(y')\ f_{\varepsilon,\delta}(y)\ \mathrm{d}y'\mathrm{d}y \\
&\quad + C \int_0^{y_0} \int_0^{y_0} \mathbf{1}_{[z_0,y_0]}(y+y')\ K_\delta(y,y')\ f_{\varepsilon,\delta}(y')\ f_{\varepsilon,\delta}(y)\ \mathrm{d}y'\mathrm{d}y.
\end{aligned}$$

Owing to (1.6) and the definition of K_δ, we have

$$\mathbf{1}_{[z_0,y_0]}(y+y')\ K_\delta(y,y') \le \frac{y+y'}{z_0}\ \big(K^\star(yy')^\sigma+\delta\big), \quad (y,y')\in Y\times Y,$$

and thus, thanks to (3.1),

$$\begin{aligned}
-\frac{C}{\varphi_\varepsilon(f_{\varepsilon,\delta})} &\le m_\gamma\ M_\sigma(f_{\varepsilon,\delta}) - \frac{\kappa_0}{2}\ \big(K_\star\ M_\sigma(f_{\varepsilon,\delta})^2 + \delta\ M_0(f_{\varepsilon,\delta})^2\big) \\
&\quad + C\int_0^{y_0}\int_0^{y_0} y\ \left(K^\star\ (y\ y')^\sigma+\delta\right)\ f_{\varepsilon,\delta}(y')\ f_{\varepsilon,\delta}(y)\ \mathrm{d}y'\mathrm{d}y \\
&\le m_\gamma\ M_\sigma(f_{\varepsilon,\delta}) - \frac{\kappa_0}{2}\ \big(K_\star\ M_\sigma(f_{\varepsilon,\delta})^2 + \delta\ M_0(f_{\varepsilon,\delta})^2\big) \\
&\quad + C\ \left(K^\star\ \varrho\ y_0^\sigma\ M_\sigma(f_{\varepsilon,\delta}) + \varrho\ \delta\ M_0(f_{\varepsilon,\delta})\right) \\
&\le C - \frac{\kappa_0}{4}\ \big(K_\star\ M_\sigma(f_{\varepsilon,\delta})^2 + \delta\ M_0(f_{\varepsilon,\delta})^2\big)
\end{aligned}$$

by the Young inequality. Since $\varepsilon\le\delta$, a further application of the Young inequality entails that

$$K_\star\ M_\sigma(f_{\varepsilon,\delta})^2 + \delta\ M_0(f_{\varepsilon,\delta})^2 \le C\ \left(1+\delta\ M_0(f_{\varepsilon,\delta})\right) \le C + \frac{\delta}{2}\ M_0(f_{\varepsilon,\delta})^2,$$

whence (3.2). □

We next turn to the cornerstone of the proof, that is, the weak compactness of $(f_{\varepsilon,\delta})$ with respect to ε. More precisely, the following result is true.

Lemma 3.3. *For $\delta\in(0,1)$ and $\varepsilon\in(0,\delta)$, we have*

$$\int_0^{y_0} \big(f_{\varepsilon,\delta}(y)\big)^p\ \mathrm{d}y\ \le\ C(\delta). \tag{3.3}$$

Proof. Owing to (2.10) and the Hölder inequality, we first notice that

$$\begin{aligned}
\big(f_{\varepsilon,\delta}(y_0)\big)^{(1+p)/2} &= \int_0^{y_0} \frac{\mathrm{d}}{\mathrm{d}y}\,[f_{\varepsilon,\delta}(y)]^{(1+p)/2}\ \mathrm{d}y \\
&= \frac{1+p}{2}\ \int_0^{y_0} \big(f_{\varepsilon,\delta}(y)\big)^{1/2}\ \big(f_{\varepsilon,\delta}(y)\big)^{(p-2)/2}\ f'_{\varepsilon,\delta}(y)\ \mathrm{d}y \\
&\le \frac{1+p}{2}\ M_0(f_{\varepsilon,\delta})^{1/2}\ \left(\int_0^{y_0} \big(f_{\varepsilon,\delta}(y)\big)^{p-2}\ |f'_{\varepsilon,\delta}(y)|^2\ \mathrm{d}y\right)^{1/2}
\end{aligned}$$

and therefore

$$C(\delta)\ \big(f_{\varepsilon,\delta}(y_0)\big)^{1+p} \le \int_0^{y_0} \big(f_{\varepsilon,\delta}(y)\big)^{p-2}\ |f'_{\varepsilon,\delta}(y)|^2\ \mathrm{d}y \tag{3.4}$$

by Lemma 3.2. We now multiply (2.9) by $p\,(f_{\varepsilon,\delta}(y))^{p-1}$ and integrate over Y. From (2.10) and (3.4), we infer that

$$\begin{aligned}
-\varepsilon\ p\ \int_0^{y_0} \big(f_{\varepsilon,\delta}(y)\big)^{p-1}\ f''_{\varepsilon,\delta}(y)\ \mathrm{d}y &= -\ \varepsilon\ p\ \big(f_{\varepsilon,\delta}(y_0)\big)^{p-1}\ f'_{\varepsilon,\delta}(y_0) \\
&\quad + \varepsilon\ p\ (p-1)\ \int_0^{y_0} \big(f_{\varepsilon,\delta}(y)\big)^{p-2}\ |f'_{\varepsilon,\delta}(y)|^2\ \mathrm{d}y \\
&\ge -\ \frac{\varepsilon\,p}{y_0}\ \big(f_{\varepsilon,\delta}(y_0)\big)^p + \varepsilon\ p\ (p-1)C(\delta)\big(f_{\varepsilon,\delta}(y_0)\big)^{p+1}.
\end{aligned} \tag{3.5}$$

Since

$$r^p \le \frac{p}{p+1}\ \xi\ r^{p+1} + \frac{1}{p+1}\ \xi^{-p}$$

for $r \ge 0$ and $\xi \in (0,\infty)$ by the Young inequality, we use this inequality with $r = f_{\varepsilon,\delta}(y_0)$ and $\xi = (p^2-1)\ y_0\ C(\delta)/p$ to bound from below the right-hand side of (3.5) and obtain that

$$-\varepsilon\ p \int_0^{y_0} \big(f_{\varepsilon,\delta}(y)\big)^{p-1}\ f''_{\varepsilon,\delta}(y)\ \mathrm{d}y \ge -\frac{\varepsilon\ p\ \xi^{-p}}{(p+1)\ y_0} \ge -\varepsilon\ C(\delta).$$

Consequently,

$$\begin{aligned} -\ C(\delta)\ &\le\ -\ \varepsilon\ C(\delta)\ \le\ -\ \varepsilon\ p \int_0^{y_0} \big(f_{\varepsilon,\delta}(y)\big)^{p-1}\ f''_{\varepsilon,\delta}(y)\ \mathrm{d}y \\ -\ C(\delta)\ &\le\ \varphi_\varepsilon(f_{\varepsilon,\delta})\ \left(I_1+I_2+I_3+I_4-I_5-I_6\right), \end{aligned} \tag{3.6}$$

where we put

$$\begin{aligned} I_1 &:= \frac{p}{2} \int_0^{y_0}\int_{y_0}^{2y_0}\int_{y'-y_0}^{y_0} K_\delta(y'',y'-y'')\ \beta_{\mathsf{s}}(y',y) \\ &\qquad\qquad \times\ f_{\varepsilon,\delta}(y'-y'')\ f_{\varepsilon,\delta}(y'')\ \big(f_{\varepsilon,\delta}(y)\big)^{p-1}\ \mathrm{d}y''\mathrm{d}y'\mathrm{d}y, \\ I_2 &:= \frac{p}{2} \int_0^{y_0}\int_{y}^{y_0}\int_0^{y'} \beta_{\mathsf{c}}(y',y)\ K_\delta(y'',y'-y'')\ Q(y'',y'-y'') \\ &\qquad\qquad \times\ f_{\varepsilon,\delta}(y'-y'')\ f_{\varepsilon,\delta}(y'')\ \big(f_{\varepsilon,\delta}(y)\big)^{p-1}\ \mathrm{d}y''\mathrm{d}y'\mathrm{d}y, \\ I_3 &:= p \int_0^{y_0}\int_y^{y_0} \gamma(y',y)\ f_{\varepsilon,\delta}(y')\ \big(f_{\varepsilon,\delta}(y)\big)^{p-1}\ \mathrm{d}y'\mathrm{d}y, \\ I_4 &:= \frac{p}{2} \int_0^{y_0}\int_0^{y} K_\delta(y',y-y')\ P(y',y-y')\ f_{\varepsilon,\delta}(y-y')\ f_{\varepsilon,\delta}(y')\ (f_{\varepsilon,\delta}(y))^{p-1}\ \mathrm{d}y'\mathrm{d}y, \\ I_5 &:= p \int_0^{y_0}\int_0^{y_0-y} K_\delta(y,y')\ \big\{P(y,y')+Q(y,y')\big\}\ f_{\varepsilon,\delta}(y')\ (f_{\varepsilon,\delta}(y))^p\ \mathrm{d}y'\mathrm{d}y, \\ I_6 &:= p \int_0^{y_0}\int_{y_0-y}^{y_0} K_\delta(y,y')\ f_{\varepsilon,\delta}(y')\ (f_{\varepsilon,\delta}(y))^p\ \mathrm{d}y'\mathrm{d}y. \end{aligned}$$

Observe then that (1.15) and the Young inequality imply that, for $\xi \in (0,1)$, there is a constant $C_\xi > 0$ such that

$$\int_0^{y_0} \beta_{\mathsf{s}}(y',y)\ p\ f_{\varepsilon,\delta}(y)^{p-1}\ \mathrm{d}y\ \le\ C_\xi\ \mu_{\mathsf{s}} + \xi \int_0^{y_0} y^\sigma \big(f_{\varepsilon,\delta}(y)\big)^p\ \mathrm{d}y\ =\ C_\xi\ \mu_{\mathsf{s}} + \xi\ M_{\sigma,p}\big(f_{\varepsilon,\delta}\big).$$

Therefore, (1.6) and Lemma 3.2 entail that

$$\begin{aligned} I_1 &= \frac{1}{2}\int_{y_0}^{2y_0}\int_{y'-y_0}^{y_0}\int_0^{y_0} \beta_{\mathsf{s}}(y',y)\ p\ \big(f_{\varepsilon,\delta}(y)\big)^{p-1}\ \mathrm{d}y \\ &\qquad\qquad \times\ K_\delta(y'',y'-y'')\ f_{\varepsilon,\delta}(y'-y'')\ f_{\varepsilon,\delta}(y'')\ \mathrm{d}y''\mathrm{d}y' \\ &\le \big(C_\xi\ \mu_{\mathsf{s}} + \xi\ M_{\sigma,p}\big(f_{\varepsilon,\delta}\big)\big) \int_0^{y_0}\int_{y_0-y'}^{y_0} K_\delta(y'',y')\ f_{\varepsilon,\delta}(y')\ f_{\varepsilon,\delta}(y'')\ \mathrm{d}y''\mathrm{d}y' \\ &\le \big(C_\xi\ \mu_{\mathsf{s}} + \xi\ M_{\sigma,p}\big(f_{\varepsilon,\delta}\big)\big) \left(K^\star\ M_\sigma(f_{\varepsilon,\delta})^2 + \frac{2}{y_0}\ \varrho\ \delta\ M_0(f_{\varepsilon,\delta})\right), \end{aligned}$$

$$I_1 \;\le\; C\;\left(C_\xi + \xi\; M_{\sigma,p}\big(f_{\varepsilon,\delta}\big)\right). \tag{3.7}$$

We estimate I_2 analogously and thus obtain from (1.2), (1.6), (1.14), and Lemma 3.2 that, for $\xi \in (0,1)$,

$$\begin{aligned} I_2 &= \frac{1}{2}\;\int_0^{y_0}\int_0^{y'}\int_0^{y'} \beta_{\mathsf{c}}(y',y)\; p\;\left(f_{\varepsilon,\delta}(y)\right)^{p-1}\;\mathrm{d}y\; Q(y'',y'-y'') \\ &\qquad\qquad\qquad\qquad \times\; K_\delta(y'',y'-y'')\; f_{\varepsilon,\delta}(y'-y'')\; f_{\varepsilon,\delta}(y'')\;\mathrm{d}y''\mathrm{d}y' \\ &\le \left(C_\xi\;\mu_{\mathsf{c}} + \xi\; M_{\sigma,p}\big(f_{\varepsilon,\delta}\big)\right)\;\int_0^{y_0}\int_0^{y_0-y} K_\delta(y,y')\; f_{\varepsilon,\delta}(y)\; f_{\varepsilon,\delta}(y')\;\mathrm{d}y'\mathrm{d}y \\ &\le \left(C_\xi\;\mu_{\mathsf{c}} + \xi\; M_{\sigma,p}\big(f_{\varepsilon,\delta}\big)\right)\;\left(K^\star\; M_\sigma(f_{\varepsilon,\delta})^2 + \delta\; M_0(f_{\varepsilon,\delta})^2\right) \end{aligned}$$

$$I_2 \;\le\; C\;\left(C_\xi\; + \xi\; M_{\sigma,p}\big(f_{\varepsilon,\delta}\big)\right). \tag{3.8}$$

In a similar way, the Young inequality and (1.13) yield

$$\begin{aligned} I_3 &= \int_0^{y_0} f_{\varepsilon,\delta}(y')\;\int_0^{y'} \gamma(y',y)\; p\;\left(f_{\varepsilon,\delta}(y)\right)^{p-1}\;\mathrm{d}y\,\mathrm{d}y' \\ &\le \int_0^{y_0} y'^{\sigma}\; f_{\varepsilon,\delta}(y')\;\int_0^{y'} p\; y^{-\sigma}\;\gamma(y',y)\;\left(f_{\varepsilon,\delta}(y)\right)^{p-1}\;\mathrm{d}y\,\mathrm{d}y' \\ &\le \int_0^{y_0} y'^{\sigma}\; f_{\varepsilon,\delta}(y')\;\int_0^{y'} p\; y^{\sigma(1-2p)/p}\;\gamma(y',y)\;\left(y^{\sigma/p} f_{\varepsilon,\delta}(y)\right)^{p-1}\;\mathrm{d}y\,\mathrm{d}y' \\ &\le \int_0^{y_0} y'^{\sigma}\; f_{\varepsilon,\delta}(y')\;\left(C_\xi\;\mu_\gamma + \xi\; M_{\sigma,p}\big(f_{\varepsilon,\delta}\big)\right)\;\mathrm{d}y', \end{aligned}$$

whence, by Lemma 3.2,

$$I_3 \;\le\; C\;\left(C_\xi + \xi\; M_{\sigma,p}\big(f_{\varepsilon,\delta}\big)\right)\; M_\sigma(f_{\varepsilon,\delta}) \;\le\; C\;\left(C_\xi + \xi\; M_{\sigma,p}\big(f_{\varepsilon,\delta}\big)\right). \tag{3.9}$$

We now estimate $I_4 - I_5 - I_6$. For that purpose, observe first that, by the Young inequality,

$$\begin{aligned} I_4 &= \frac{1}{2}\;\int_0^{y_0}\int_0^{y} K_\delta(y',y-y')\; P(y',y-y')\; f_{\varepsilon,\delta}(y-y')\; f_{\varepsilon,\delta}(y')\; p\;\left(f_{\varepsilon,\delta}(y)\right)^{p-1}\mathrm{d}y'\mathrm{d}y \\ &\le \frac{1}{2}\;\int_0^{y_0}\int_0^{y_0-y'} K_\delta(y',y)\; P(y',y)\; f_{\varepsilon,\delta}(y)\;\left(f_{\varepsilon,\delta}(y')\right)^{p}\;\mathrm{d}y\;\mathrm{d}y' \\ &\quad + \frac{1}{2}\;(p-1)\;\int_0^{y_0}\int_0^{y} K_\delta(y-y',y')\; P(y-y',y')\; f_{\varepsilon,\delta}(y')\;\left(f_{\varepsilon,\delta}(y)\right)^{p}\;\mathrm{d}y'\,\mathrm{d}y. \end{aligned}$$

Therefore, we derive that

$$\begin{aligned} I_4 - I_5 - I_6 \le{}& \frac{1}{2}\;\int_0^{y_0}\int_0^{y_0-y} K_\delta(y,y')\; P(y,y')\; f_{\varepsilon,\delta}(y')\;\left(f_{\varepsilon,\delta}(y)\right)^{p}\;\mathrm{d}y'\mathrm{d}y \\ &+ \frac{1}{2}\;(p{-}1)\int_0^{y_0}\int_0^{y} K_\delta(y-y',y')\; P(y-y',y')\; f_{\varepsilon,\delta}(y')\big(f_{\varepsilon,\delta}(y)\big)^{p}\mathrm{d}y'\mathrm{d}y \\ &- p\;\int_0^{y_0}\int_0^{y_0-y} K_\delta(y,y')\;(P+Q)(y,y')\; f_{\varepsilon,\delta}(y')\;\left(f_{\varepsilon,\delta}(y)\right)^{p}\;\mathrm{d}y'\mathrm{d}y \end{aligned}$$

$$\begin{aligned}
& - p \int_0^{y_0} \int_{y_0-y}^{y_0} K_\delta(y,y')\ f_{\varepsilon,\delta}(y')\ \big(f_{\varepsilon,\delta}(y)\big)^p\ \mathrm{d}y'\mathrm{d}y \\
\le & -\frac{1}{2}\ p \int_0^{y_0} \int_0^{y_0-y} K_\delta(y,y')\ (P+2\ Q)(y,y')\ f_{\varepsilon,\delta}(y')\ \big(f_{\varepsilon,\delta}(y)\big)^p\ \mathrm{d}y'\mathrm{d}y \\
& -\frac{1}{2}\ (p+1) \int_0^{y_0} \int_{y_0-y}^{y_0} K_\delta(y,y')\ f_{\varepsilon,\delta}(y')\ \big(f_{\varepsilon,\delta}(y)\big)^p\ \mathrm{d}y'\mathrm{d}y \\
\le & -\frac{1}{2}\ p\ P_\star \int_0^{y_0} \int_0^{y_0} K_\delta(y,y')\ f_{\varepsilon,\delta}(y')\ \big(f_{\varepsilon,\delta}(y)\big)^p\ \mathrm{d}y'\mathrm{d}y,
\end{aligned}$$

where we have used the monotonicity conditions (1.7) and (1.9) to obtain the second inequality and (1.8) for the last inequality (recall that $P_\star \le 1$). Since $\varrho = M_1(f_{\varepsilon,\delta}) \le y_0\, M_0(f_{\varepsilon,\delta})$ and $\varrho = M_1(f_{\varepsilon,\delta}) \le y_0^{1-\sigma}\ M_\sigma(f_{\varepsilon,\delta})$, we deduce from (1.6) that

$$I_4 - I_5 - I_6 \le -\ \frac{1}{2}\ p\ P_\star\ \left(K_\star\ M_\sigma\big(f_{\varepsilon,\delta}\big)\ M_{\sigma,p}\big(f_{\varepsilon,\delta}\big) + \delta\ M_0\big(f_{\varepsilon,\delta}\big)\ M_{0,p}\big(f_{\varepsilon,\delta}\big)\right),$$

$$I_4 - I_5 - I_6 \le -\ \frac{1}{2}\ p\ \varrho\ P_\star\ \left(K_\star\ y_0^{\sigma-1}\ M_{\sigma,p}\big(f_{\varepsilon,\delta}\big) + \delta\ y_0^{-1}\ M_{0,p}\big(f_{\varepsilon,\delta}\big)\right). \tag{3.10}$$

Gathering (3.6)–(3.10) we end up with

$$-\frac{C(\delta)}{\varphi_\varepsilon(f_{\varepsilon,\delta})}\ \le\ C\ \left(C_\xi + \xi\ M_{\sigma,p}\big(f_{\varepsilon,\delta}\big)\right)\ -\ \frac{1}{2\,y_0}\ p\ P_\star\ \varrho\ \delta\ M_{0,p}\big(f_{\varepsilon,\delta}\big).$$

Then choosing $\xi \in (0,1)$ sufficiently small and noticing that $(\varphi_\varepsilon(f_{\varepsilon,\delta}))^{-1} \le (1 + \delta\, M_0(f_{\varepsilon,\delta})) \le C$ due to $\varepsilon \in (0,\delta)$ and Lemma 3.2, the assertion follows since $M_{\sigma,p}(f_{\varepsilon,\delta}) \le y_0^\sigma\ M_{0,p}(f_{\varepsilon,\delta})$. ∎

The fact that the monotonicity condition (1.7) on the coagulation kernel K yields L^p-estimates has already been used in [9] for the classical coagulation equation (see also [11] and the references therein). In addition, the weak compactness in $L^1(Y)$ of the trajectories of (1.1) established in [15] relies on a similar observation. We adapt here this strategy to estimate $I_4 - I_5 - I_6$ under more general assumptions than the one used in [15].

Now the proof of Theorem 1.1 is a consequence of the previous considerations.

Proof of Theorem 1.1. Keeping $\delta \in (0,1)$ fixed, the set $\{f_{\varepsilon,\delta};\, \varepsilon \in (0,\delta)\}$ is bounded in $L^p(Y)$ according to Lemma 3.3. Therefore, there are a sequence $(f_{\varepsilon_n,\delta})$ and $f_\delta \in L^p(Y)$ such that

$$f_{\varepsilon_n,\delta} \rightharpoonup f_\delta \quad \text{in} \quad L^p(Y) \quad \text{as} \quad \varepsilon_n \to 0. \tag{3.11}$$

Since $f_{\varepsilon_n,\delta}$ is nonnegative and satisfies $M_1(f_{\varepsilon_n,\delta}) = \varrho$ by (3.1) for each $n \ge 1$, it readily follows from (3.11) that

$$f_\delta \ge 0 \quad \text{a.e. in} \quad Y \quad \text{and} \quad M_1(f_\delta) = \varrho. \tag{3.12}$$

We then claim that $L_\delta(f_\delta) = 0$. Indeed, on the one hand, it is well known that L_δ is weakly continuous in $L^1(Y)$ (see either the pioneering work [13] or [15, Appendix A] for a complete proof), and the convergence (3.11) ensures that $L_\delta(f_{\varepsilon_n,\delta}) \rightharpoonup L_\delta(f_\delta)$ in $L^1(Y)$. On the other hand, by (3.2), $(-\varepsilon_n\, f''_{\varepsilon_n,\delta})$ converges to zero in $\mathcal{D}'(Y)$. Consequently,

$$\int_0^{y_0} L_\delta(f_\delta)(y)\ \psi(y)\ \mathrm{d}y = 0 \qquad \text{for each} \quad \psi \in C_0^\infty(Y), \tag{3.13}$$

whence,

$$L_\delta(f_\delta) = 0 \quad \text{a.e. in } Y. \tag{3.14}$$

We may now test (3.14) with $p\,\left(f_\delta\right)^{p-1}$ and obtain

$$0 \;=\; p \int_0^{y_0} \left(f_\delta(y)\right)^{p-1}\, L_\delta(f_\delta)(y)\, \mathrm{d}y \;\le\; I_1 + I_2 + I_3 + I_4 - I_5 - I_6, \tag{3.15}$$

where the I_k's are defined as in the proof of Lemma 3.3 but with $f_{\varepsilon,\delta}$ replaced by f_δ. Since (3.2) and (3.11) imply that

$$\delta^{1/2}\, M_0(f_\delta) \le C \quad \text{and} \quad M_\sigma(f_\delta) \le C, \tag{3.16}$$

we can proceed as in the proof of (3.7), (3.8), (3.9), and (3.10) in Lemma 3.3 to deduce that, for $\xi \in (0,1)$,

$$I_1 + I_2 + I_3 \;\le\; C\; \left(C_\xi + \xi\; M_{\sigma,p}(f_\delta)\right) \tag{3.17}$$

and

$$I_4 - I_5 - I_6 \;\le\; -\,\frac{1}{2}\; p\; \varrho\; P_\star\; K_\star\; y_0^{\sigma-1}\; M_{\sigma,p}\big(f_\delta\big). \tag{3.18}$$

Combining (3.15), (3.17), and (3.18) and choosing $\xi \in (0,1)$ sufficiently small, we finally obtain that

$$\int_0^{y_0} y^\sigma \left(f_\delta(y)\right)^p\, \mathrm{d}y \;\le C.$$

Therefore, we may extract a subsequence (f_{δ_n}) and find $f \in L^p(Y, y^\sigma\,\mathrm{d}y)$ such that

$$f_{\delta_n} \rightharpoonup f \quad \text{in} \quad L^p(Y, y^\sigma\,\mathrm{d}y) \quad \text{as} \quad \delta_n \to 0. \tag{3.19}$$

Clearly,

$$f \ge 0 \qquad \text{and} \qquad M_1(f) = \varrho \tag{3.20}$$

owing to (3.12) and $\sigma \le 1$. In particular, $f \not\equiv 0$ since $\varrho > 0$.

It therefore remains to prove that $L(f) = 0$ a.e. in Y. For that purpose, we observe that (3.14) also reads as

$$L(f_\delta) = L(f_\delta) - L_\delta(f_\delta) \quad \text{a.e. in} \quad Y. \tag{3.21}$$

On the one hand, we have $L(f_\delta) = \tilde{L}(g_\delta)$, where $g_\delta(y) := y^\sigma\; f_\delta(y)$, $y \in Y$, and $\tilde{L}$ is defined as L with $\tilde{\gamma}(y,y') := y^{-\sigma}\; \gamma(y,y')$ and $\tilde{K}(y,y') := (y\,y')^{-\sigma}\; K(y,y')$ instead of γ and K. Owing to (1.6) and (1.16), $\tilde{L}$ is weakly continuous in $L^1(Y)$ (see, e.g., [13, 15]). We then deduce from this property and the convergence (3.19) that

$$L(f_{\delta_n}) = \tilde{L}(g_{\delta_n}) \rightharpoonup \tilde{L}(g) = L(f) \quad \text{in} \quad L^1(Y) \tag{3.22}$$

with $g(y) := y^\sigma\; f(y)$, $y \in Y$.

On the other hand, let ψ be an arbitrary function in $C_0^\infty(Y)$ and choose $a > 0$ such that the support of ψ is contained in $[a, y_0 - a]$. Then, $\psi(y) \le (\|\psi\|_\infty\; y)/a$. This fact, together with (1.2), (1.10), and (1.12) allow us to deduce that the functions ψ_{c} and ψ_{s} defined in (2.2) and (2.3), respectively, satisfy

$$\begin{aligned}
|\psi_{\mathsf{c}}(y,y')| &\le |\psi(y+y')| + |\psi(y)| + |\psi(y')| \\
&\qquad + \frac{\|\psi\|_\infty\ Q(y,y')}{a}\ \int_0^{y+y'} y''\ \beta_{\mathsf{c}}(y+y',y'')\ \mathrm{d}y'' \\
&\le \frac{3\ \|\psi\|_\infty}{a}\ (y+y'), \\
|\psi_{\mathsf{s}}(y,y')| &\le \frac{\|\psi\|_\infty}{a}\ \int_0^{y_0} y''\ \beta_{\mathsf{s}}(y+y',y'')\ \mathrm{d}y'' + |\psi(y)| + |\psi(y')| \le \frac{2\ \|\psi\|_\infty}{a}\ (y+y').
\end{aligned}$$

We then infer from (2.2), (2.3), (3.12), and (3.16) that

$$\begin{aligned}
\left|\int_0^{y_0} \psi(y)\ (L(f_\delta) - L_\delta(f_\delta))(y)\ \mathrm{d}y\right| &\le \frac{\delta}{2}\ \int_0^{y_0}\int_0^{y_0-y} |\psi_{\mathsf{c}}(y,y')|\ f_\delta(y)\ f_\delta(y')\ \mathrm{d}y'\mathrm{d}y \\
&\quad + \frac{\delta}{2}\ \int_0^{y_0}\int_{y_0-y}^{y_0} |\psi_{\mathsf{s}}(y,y')|\ f_\delta(y)\ f_\delta(y')\ \mathrm{d}y'\mathrm{d}y \\
&\le \frac{5\ \|\psi\|_\infty\ \delta}{a}\ \int_0^{y_0}\int_0^{y_0} y\ f_\delta(y)\ f_\delta(y')\ \mathrm{d}y'\mathrm{d}y \\
&\le \frac{5\ \varrho\ \|\psi\|_\infty\ \delta^{1/2}}{a}\ \left(\delta^{1/2}\ M_0(f_\delta)\right) \\
&\le \frac{C\ \|\psi\|_\infty}{a}\ \delta^{1/2}.
\end{aligned}$$

Letting $\delta \to 0$ then implies that

$$L(f_\delta) - L_\delta(f_\delta) \to 0 \quad \text{in} \quad \mathcal{D}'(Y) \tag{3.23}$$

as $\delta \to 0$. As a consequence of (3.22) and (3.23), we may pass to the limit as $\delta_n \to 0$ in (3.21) and conclude that $L(f) = 0$ in $\mathcal{D}'(Y)$. Since $L(f)$ actually belongs to $L^1(Y)$, we conclude that $L(f) = 0$ a.e. in Y, which completes the proof of Theorem 1.1. ∎

4. Nonexistence of nonzero solutions. In this concluding section we show that the problem (1.4), (1.5) is not always well-posed if either (1.13)–(1.15) or (1.11) is violated. To this end, we first assume that there is no spontaneous breakage, i.e., $\gamma \equiv 0$, and that K and P are strictly positive a.e. on their domains. Furthermore, we suppose that shattering and scattering are mass preserving and binary processes, that is, that (1.10) and (1.12) are satisfied and additionally that

$$\beta_{\mathsf{c}}(y,y') = \beta_{\mathsf{c}}(y,y-y'), \qquad 0 < y' < y < y_0, \tag{4.1}$$

$$\beta_{\mathsf{s}}(y,y') = \beta_{\mathsf{s}}(y,y-y') > 0, \quad 0 < y - y_0 < y' < y_0 \tag{4.2}$$

and

$$\beta_{\mathsf{s}}(y,y') = 0, \quad 0 < y' < y - y_0 < y_0. \tag{4.3}$$

The latter assumption is due to consistency of the model since each of the daughter particles y' and $y - y'$ in (4.2) has to belong to Y. Note that (1.10), (1.12), and (4.1)–(4.3) imply

$$Q(y,y')\int_0^{y+y'} \beta_{\mathsf{c}}(y+y',y'')\ \mathrm{d}y'' = 2\ Q(y,y') \quad \text{for a.a. } (y,y') \in \Xi,$$

and

$$\int_{y-y_0}^{y_0} \beta_{\mathsf{s}}(y,y')\ \mathrm{d}y' = 2 \quad \text{for a.a. } y \in (y_0, 2y_0);$$

in particular, (1.15) is violated due to the Hölder inequality.

PROPOSITION 4.1. *If $\gamma \equiv 0$, K and P are strictly positive a.e. on their domains and if β_{c} and β_{s} satisfy* (4.1)–(4.3), *the only solution $f \in L^1_+(Y)$ to* (1.4) *is $f \equiv 0$.*

Proof. Let $u \in L^1_+(Y)$ be a solution to (1.4). Then we deduce from (2.1)–(2.3) with $\psi \equiv 1$ that

$$\begin{aligned} 0 = &-\int_0^{y_0} L(u)(y)\ \mathrm{d}y \ = \ \frac{1}{2}\int_0^{y_0}\int_0^{y_0-y} K(y,y')\ P(y,y')\ u(y)\ u(y')\ \mathrm{d}y'\mathrm{d}y \\ \geq &\ \frac{1}{2}\int_0^{y_0/2}\int_0^{y_0/2} K(y,y')\ P(y,y')\ u(y)\ u(y')\ \mathrm{d}y'\mathrm{d}y \ \geq \ 0, \end{aligned}$$

whence $u \equiv 0$ on $(0, y_0/2)$ and $L_{\mathsf{c}}(u)(y) = 0$ for a.e. $y \in Y$. Therefore,

$$\begin{aligned} 0 = L(u)(y) = &\ \frac{1}{2}\int_{y_0}^{y_0+y}\int_{y'-y_0}^{y_0} K(y'', y'-y'')\ \beta_{\mathsf{s}}(y',y)\ u(y'')\ u(y'-y'')\ \mathrm{d}y''\mathrm{d}y' \\ &- u(y)\int_{y_0-y}^{y_0} K(y,y')\ u(y')\ \mathrm{d}y' \end{aligned} \tag{4.4}$$

for a.e. $y \in Y$. We claim that this implies that $u \equiv 0$ on $(\xi \vee (y_0-\xi), (y_0+\xi)/2)$ for a.e. $\xi \in (y_0/3, y_0)$ such that $u(\xi) = 0$ (recall that $\xi \vee (y_0-\xi) := \max\{\xi, y_0-\xi\}$). Indeed, consider $\xi \in \big(y_0/3, y_0\big)$ such that $u(\xi) = 0$. Since

$$\big(\xi \vee (y_0-\xi), (y_0+\xi)/2\big)^2 \subset \big\{(y'', y')\ ;\ \xi < y'' < y_0,\ y_0 < y'+y'' < y_0+\xi\big\},$$

we infer from (4.4) that

$$\begin{aligned} 0 = &\int_{y_0}^{y_0+\xi}\int_{y'-y_0}^{y_0} K(y'', y'-y'')\ \beta_{\mathsf{s}}(y',\xi)\ u(y'')\ u(y'-y'')\ \mathrm{d}y''\mathrm{d}y' \\ = &\int_0^{y_0}\int_{y_0}^{y_0+(y''\wedge\xi)} K(y'', y'-y'')\ \beta_{\mathsf{s}}(y',\xi)\ u(y'')\ u(y'-y'')\ \mathrm{d}y'\mathrm{d}y'' \\ = &\int_\xi^{y_0}\int_{y_0-y''}^{y_0+\xi-y''} K(y'', y')\ \beta_{\mathsf{s}}(y'+y'',\xi)\ u(y'')\ u(y')\ \mathrm{d}y'\mathrm{d}y'' \\ \geq &\int_{\xi\vee(y_0-\xi)}^{(y_0+\xi)/2}\int_{\xi\vee(y_0-\xi)}^{(y_0+\xi)/2} K(y'', y')\ \beta_{\mathsf{s}}(y'+y'',\xi)\ u(y'')\ u(y')\ \mathrm{d}y'\mathrm{d}y'' \\ \geq &\ 0, \end{aligned}$$

whence $u \equiv 0$ on $\big((\xi \vee y_0-\xi), (y_0+\xi)/2\big)$. Defining $\xi_k := (1-2^{-k-1})y_0$, we inductively infer that $u \equiv 0$ on $(0, \xi_k)$ for $k \in \mathbb{N}$ by a density argument, whence $u \equiv 0$ on Y. ☐

On the other hand, integrating (1.4) over Y and recalling that scattering produces at least two daughter particles, i.e.,

$$\int_0^{y_0} \beta_{\mathsf{s}}(y,y')\ \mathrm{d}y' \ \geq\ 2, \quad y \in (y_0, 2y_0),$$

we easily see that zero is the only steady state provided that, in addition,

$$Q(y,y') \int_0^{y+y'} \beta_{\mathsf{c}}(y+y',y'')\ \mathrm{d}y'' \ \geq\ 2\ Q(y,y') + P(y,y'), \quad y+y' \in Y,$$

and

$$\int_0^y \left(1-\frac{y'}{y}\right)\ \gamma(y,y')\ \mathrm{d}y' > 0\ , \quad y \in Y.$$

Obviously, the former assumption contradicts (1.11).

Acknowledgments. We thank Prof. Herbert Amann for stimulating discussions. This work was initiated during a visit of the first author to the Institute of Mathematics of the University of Zürich. He gratefully thanks this institution for hospitality and support. We also thank the referees for useful comments.

REFERENCES

[1] M. Aizenman and T.A. Bak, *Convergence to equilibrium in a system of reacting polymers*, Comm. Math. Phys., 65 (1979), pp. 203–230.

[2] D.J. Aldous, *Deterministic and stochastic models for coalescence (aggregation, coagulation): A review of the mean-field theory for probabilists*, Bernoulli, 5 (1999), pp. 3–48.

[3] I. Borsi, *Dynamics of liquid-liquid dispersions with unbounded fragmentation kernel*, Adv. Math. Sci. Appl., 11 (2001), pp. 571–591.

[4] M. Escobedo, S. Mischler, and M. Rodriguez Ricard, *On self-similarity and stationary problem for fragmentation and coagulation models*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 22 (2005), pp. 99–125.

[5] A. Fasano, *The dynamics of two-phase liquid dispersions: Necessity of a new approach*, Milan J. Math., 70 (2002), pp. 245–264.

[6] A. Fasano and F. Rosso, *A new model for the dynamics of dispersions in a batch reactor: Theory and numerical simulations*, in Lectures on Applied Mathematics. Proceedings of the Symposium Organized by the Sonderforschungsbereich 438 on the Occasion of Karl-Heinz Hoffmann's 60th Birthday, H.J. Bungartz, R.H.W. Hoppe, and C. Zenger, eds., Springer-Verlag, Berlin, 2000, pp. 123–141.

[7] P. Dubovskii and I.W. Stewart, *Approach to equilibrium for the coagulation-fragmentation equation via a Lyapunov functional*, Math. Methods Appl. Sci., 19 (1996), pp. 171–183.

[8] P. Dubovskii and I.W. Stewart, *Trend to equilibrium for the coagulation-fragmentation equation*, Math. Methods Appl. Sci., 19 (1996), pp. 761–772.

[9] Ph. Laurençot and S. Mischler, *The continuous coagulation-fragmentation equations with diffusion*, Arch. Ration. Mech. Anal., 162 (2002), pp. 45–99.

[10] Ph. Laurençot and S. Mischler, *Convergence to equilibrium for the continuous coagulation-fragmentation equation*, Bull. Sci. Math., 127 (2003), pp. 179–190.

[11] Ph. Laurençot and S. Mischler, *On coalescence equations and related models*, in Modeling and Computational Methods for Kinetic Equations, P. Degond, L. Pareschi, and G. Russo, eds., Birkhäuser, Boston, 2004, pp. 321–356.

[12] A. Mancini and F. Rosso, *A new model for the dynamics of dispersions in a batch reactor: Numerical approach*, Meccanica, 37 (2002), pp. 221–237.

[13] I.W. Stewart, *A global existence theorem for the general coagulation-fragmentation equation with unbounded kernels*, Math. Methods Appl. Sci., 11 (1989), pp. 627–648.

[14] C. Walker, *Coalescence and breakage processes*, Math. Methods Appl. Sci., 25 (2002), pp. 729–748.

[15] C. Walker, *On Diffusive and Non-Diffusive Coalescence and Breakage Processes*, Ph.D. thesis, Universität Zürich, 2003.

[16] C. Walker, *Asymptotic behavior of liquid-liquid dispersions*, Proc. Roy. Soc. Edinburgh Sect. A, 134 (2004), pp. 753–772.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 549–597

MICROLOCAL DISPERSIVE SMOOTHING FOR THE NONLINEAR SCHRÖDINGER EQUATION*

JÉRÉMIE SZEFTEL†

Abstract. We prove a dispersive smoothing result for the nonlinear Schrödinger equation. We deal with the linear term by using the method of Craig, Kappeler, and Strauss. We rewrite the nonlinear term as the sum of a paradifferential operator and a remainder using Bony's theorem. We prove an interpolation result between weighted L^2 spaces and Sobolev spaces which enables us to deal with the remainder. Finally, we deal with the paradifferential operator using the symbolic calculus.

Key words. nonlinear Schrödinger equation, smoothing effect, interpolation, weighted L^2 space, paradifferential calculus

AMS subject classifications. 35Q55, 35A27, 35B65, 46B70, 35S50

DOI. 10.1137/S0036141003432109

1. Introduction.

Lascar [12] obtained the first propagation results for the linear Schrödinger equation (see also Boutet de Monvel [3]). He introduced a parabolic wave front set and proved its propagation along the flow of the Laplacian at any fixed time $t \neq 0$. However, we cannot use this result to link the initial data to the singularities of the solution at a positive time. The following example illustrates the link between u_0 and $u(t,.)$:

$$\begin{cases} i\dfrac{\partial u}{\partial t} + \Delta u = 0, \quad t > 0, \quad x \in \mathbb{R}^d, \\ u|_{t=0} = u_0. \end{cases} \tag{1.1}$$

If we take $u_0 = \delta$, then the solution $u(t,.)$ is in $C^\infty(\mathbb{R}^d)$ for all positive times. Conversely, if we take the smooth data $u_0 = e^{-i|x|^2}$, then the solution $u(t,.)$ is singular at $t = 1$. We notice that the smoothness of $u(t,.)$ for $t > 0$ depends on the behavior of u_0 at infinity. This idea has been extended (microlocalization, Laplacians which are flat perturbation at infinity of the constant coefficient case) in many recent works. See Craig, Kappeler, and Strauss [6]; Wunsch [17]; Robbiano and Zuily [14], [15]; and Doi [7]; we refer the reader to [6] for a more complete bibliography.

In the case of the nonlinear Schrödinger equation, there are several works proving the existence of a smoothing effect (see, for instance, Hayashi, Nakamitsu, Tsutsumi [8], Kenig, Ponce, Vega [11], and Chihara [5]).

To our knowledge, there is no result on the microlocal smoothing effect for the nonlinear Schrödinger equation, and this is the aim of this study. The present work consists of four parts.

*Received by the editors July 25, 2003; accepted for publication (in revised form) February 28, 2005; published electronically October 27, 2005.
http://www.siam.org/journals/sima/37-2/43210.html
†Department of Mathematics, Princeton University, Fine Hall, Washington Road, Princeton, NJ 08544-1000 (jszeftel@math.princeton.edu).

- We first state our result of microlocal smoothing effect for the equation

$$\begin{cases} i\dfrac{\partial u}{\partial t} + \dfrac{1}{2}\displaystyle\sum_{j,l=1}^{d} \partial_{x_j} a^{jl}(x)\partial_{x_l} u = f(u,\bar{u}), \quad 0 < t < T, \quad x \in \mathbb{R}^d, \\ u|_{t=0} = u_0, \end{cases} \tag{1.2}$$

  where $\sum_{j,l=1}^{d} \partial_{x_j} a^{jl}(x)\partial_{x_l}$ is an asymptotically constant elliptic self-adjoint operator and f is C^∞ and vanishes at 0. We also give an outline of the proof.
- We recall technical tools obtained by Craig, Kappeler, and Strauss [6]. We will use these tools to adapt the proof of [6] to our nonlinear equation.
- Then we prove technical lemmas which will be useful in what follows. Among others, we establish interpolation results between Sobolev spaces and weighted L^2 spaces and we construct a paradifferential algebra well-suited to our problem.
- Finally, we prove the microlocal smoothing effect for (1.2) stated in the first part. We paralinearize $f(u,\bar{u})$ using Bony's theorem [2]. We get a paradifferential operator evaluated at u, a paradifferential operator evaluated at $\bar{u}$, and a regular remainder. We treat the remainder like a right-hand side using the result of the second part. As we do not know how to deal with the paradifferential operator evaluated at $\bar{u}$, we introduce $v = u - Q\bar{u}$, where Q is chosen such that v satisfies the same kind of equation as u without term evaluated at $\bar{v}$. Then, we adapt the strategy of [6] to the operator evaluated at v which is the sum of $i\partial_t + 1/2\sum_{j,l=1}^{d} \partial_{x_j} a^{jl}(x)\partial_{x_l}$ and a paradifferential operator. Finally, the smoothing effect for v implies a smoothing effect for u.

2. Statements of the main results and outline of the proof.

2.1. Statements of the main results. We first recall some usual definitions and properties of microlocal analysis. A subset V of $T^*(\mathbb{R}^d)\setminus\{0\}$ is conic if $(x,\xi) \in V$ implies $(x,\lambda\xi) \in V$ for all $\lambda > 0$. To a symbol $a(x,\xi)$, we associate the operator $a(x,D)$ acting on functions u defined on $\mathbb{R}^d$ by

$$a(x,D)u(x) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{i(x-y)\xi} a(x,\xi)u(y)dy\, d\xi.$$

The usual pseudodifferential algebra S^m consists of the symbols $a(x,\xi)$ such that

$$\left|\partial_x^\alpha \partial_\xi^\beta a(x,\xi)\right| \le C_{\alpha\beta}\langle\xi\rangle^{m-|\beta|}$$

for all multi-indices α,β in $\mathbb{N}^d$, where $\langle\xi\rangle = (1+|\xi|^2)^{1/2}$. For $a(x,\xi)$ in S^m, $a(x,D)$ is bounded from H^s to H^{s-m} for all real numbers s. Finally, let u be in $\mathcal{S}'(\mathbb{R}^d)$. u is microlocally H^s at (x_0,ξ_0) if there is $a(x,\xi)$ in S^0 such that $\liminf_{\lambda\to+\infty} |a(x_0,\lambda\xi_0)| \ne 0$ and $a(x,D)u \in H^s$.

Let τ be an increasing function from $\mathbb{N}$ to $\mathbb{R}$ such that

$$\tau(m) > m+1 \text{ for } m \ge 0 \tag{2.1}$$

and let the C^∞ coefficients $a^{jl}(x)$ satisfy

$$\left|\partial_x^\alpha\left(a^{jl}(x) - a_0^{jl}\right)\right| \le \frac{C_\alpha}{\langle x\rangle^{\tau(|\alpha|)}} \quad \forall\alpha \in \mathbb{N}^d, \tag{2.2}$$

where a_0^{jl} are constants and where $\langle x\rangle = (1+|x|^2)^{1/2}$. The Schrödinger equation studied in [6] is

$$\begin{cases} i\dfrac{\partial u}{\partial t} + \dfrac{1}{2}\displaystyle\sum_{j,l=1}^{d} \partial_{x_j} a^{jl}(x)\partial_{x_l} u = 0, \quad 0<t<T, \quad x\in\mathbb{R}^d, \\ u|_{t=0} = u_0. \end{cases} \tag{2.3}$$

The principal symbol $a_2(x,\xi) = 1/2\sum_{j,l=1}^{d} a^{jl}(x)\xi_j\xi_l$ is elliptic:

$$a_2(x,\xi) \geq C|\xi|^2, \quad C>0. \tag{2.4}$$

The bicharacteristics are orbits of the flow $\varphi(s,x,\xi) = (X(s,x,\xi),\Xi(s,x,\xi)) : T^*(\mathbb{R}^d) \to T^*(\mathbb{R}^d)$ of the Hamiltonian system with Hamiltonian $a_2(x,\xi)$:

$$\frac{dX}{ds} = \partial_\Xi a_2(X,\Xi), \quad \frac{d\Xi}{ds} = -\partial_X a_2(X,\Xi). \tag{2.5}$$

We will say that a point $(x_0,\xi_0)\in T^*(\mathbb{R}^d)$, $\xi_0\neq 0$, is not trapped by the backward flow if

$$|X(s,x_0,\xi_0)| \to +\infty \quad \text{when} \quad s\to-\infty. \tag{2.6}$$

For such (x_0,ξ_0), we consider conic neighborhoods $\mathcal{E}$ of $\{\varphi(s,x_0,\xi_0), s\leq 0\}$ such that there is a $\Xi_1\in\mathbb{R}^d$ and the asymptotic behavior of the neighborhood $\mathcal{E}$ as $|x|\to+\infty$ is given by

$$\mathcal{E} \simeq \left\{(x,\xi)\Big/ \left|\frac{\xi.\Xi_1}{|\xi||\Xi_1|}-1\right|<\epsilon, \quad \left|\frac{x.\xi}{|x||\xi|}+1\right|<\epsilon\right\} \tag{2.7}$$

for some $\epsilon>0$.

Definition 2.1. *The C^∞ function $a(x,\xi)$ is a symbol of class $S(m,k)$ if*

$$\left|\partial_x^\alpha\partial_\xi^\beta a(x,\xi)\right| \leq C_{\alpha\beta}\langle\xi\rangle^{m-|\beta|}\langle x\rangle^{k-|\alpha|} \tag{2.8}$$

for all multi-indices α,β in $\mathbb{N}^d$.

We will adapt the results of [6] to the following equation:

$$\begin{cases} i\dfrac{\partial u}{\partial t} + \dfrac{1}{2}\displaystyle\sum_{j,l=1}^{d} \partial_{x_j} a^{jl}(x)\partial_{x_l} u = f(u,\bar u), \quad 0<t<T, \quad x\in\mathbb{R}^d, \\ u|_{t=0} = u_0, \end{cases} \tag{2.9}$$

where the coefficients satisfy (2.1), (2.2), and (2.4) and where f is in $C^\infty(\mathbb{C}^2)$ and vanishes at $(0,0)$. In that follows, we consider functions u belonging to spaces of the type $L^\infty(0,T,H^s)$. We choose $s>d/2$ such that $s-d/2$ is not an integer to be in the frame of Bony's paradifferential calculus. Moreover, to construct a function $v=u-Q\bar u$ satisfying an equation of type (2.9) without term evaluated at $\bar v$, we ask that for any integer k such that $2k<s-d/2+2$, $s-2k\geq 0$.

Lemma 2.2. *The set of real numbers s such that $s>d/2$, $s-d/2$ is not an integer, and $s-2k\geq 0$ for all integer k satisfying $2k<s-d/2+2$ is the subset I_d of $\mathbb{R}$ given by*

$$
\begin{aligned}
&\textit{if } d=1, \quad I_d=\bigcup_{k\geq 1}[2k,2k+1/2[,\\
&\textit{if } d=2, \quad I_d=\bigcup_{k\geq 1}]2k,2k+1[,\\
&\textit{if } d=3, \quad I_d=\bigcup_{k\geq 1}[2k,2k+1/2[\cup]2k+1/2,2k+3/2[,\\
&\textit{if } d\geq 4, \quad I_d=\bigcup_{k\geq 0}]k+d/2,k+1+d/2[.
\end{aligned}
\tag{2.10}
$$

Proof. If $2k < s-d/2+2$, $s-2k > d/2-2$, so $s-2k>0$ for $d\geq 4$. Therefore, when $d\geq 4$, I_d is the set of $s>d/2$ such that $s-d/2$ is not an integer, which is given by

$$I_d=\bigcup_{k\geq 0}]k+d/2,k+1+d/2[.$$

We suppose now that $1\leq d\leq 3$. It suffices to consider the biggest integer k such that $2k<s-d/2+2$. As $s-d/2$ is not an integer, k satisfies

$$2k<s-d/2+2<2(k+1),$$

which together with $s\geq 2k$ yields

$$2k\leq s<2k+d/2.$$

Thus, for $1\leq d\leq 3$, we have

$$I_d=\{s>d/2\,/\,s-d/2\notin\mathbb{N}\}\cap\bigcup_{k\geq 1}[2k,2k+d/2[,$$

which yields (2.10) for $1\leq d\leq 3$. ∎

The following theorems describe the microlocal smoothing effect for the nonlinear Schrödinger equation. We will prove them in section 5.

THEOREM 2.3. *Let (x_0,ξ_0) not be trapped backwards. Let there exist a conic neighborhood $\mathcal{E}$ of $\{\varphi(s,x_0,\xi_0), s\leq 0\}$ satisfying* (2.7). *Let $T>0$, s in I_d, and u in $C(0,T,L^2(\mathbb{R}^d))$ be a solution of* (2.9). *Assume that there exists $\theta(x)$ in $S(0,0)$ equal to 1 on $\mathcal{E}$ such that θu is in $L^\infty(0,T,H^s)$ and $\langle x\rangle^s\theta u$ is in $L^\infty(0,T,L^2)$. Moreover, assume that there exists $s(x,\xi)$ in $S(0,2s-d/2)$ such that $\langle x\rangle^{2(2s-d/2)}\leq s^2(x,\xi)$ on the set $\mathcal{E}$. Assume that u_0 satisfies*

$$\langle s(x,D)u_0, s(x,D)u_0\rangle<+\infty, \tag{2.11}$$

$$\langle s(x,D)\overline{u_0}, s(x,D)\overline{u_0}\rangle<+\infty. \tag{2.12}$$

If $u(t,.)$ is microlocally H^σ at $(x_0,-\xi_0)$ for a time t, $0<t\leq T$, then $u(t,.)$ is microlocally $H^{\min(\sigma+2,2s-d/2)}$ at (x_0,ξ_0).

If there exists a pseudodifferential operator with a symbol $c(x,\xi)$ in S^0 elliptic at $(x_0,-\xi_0)$ such that

$$\int_0^T t^{2(2s-d/2)+\delta}\|c(x,D)u(t,.)\|^2_{H^\sigma}dt<+\infty \tag{2.13}$$

for a $\delta > 0$, then there exists a pseudodifferential operator with a symbol $c_1(x,\xi)$ in S^0 elliptic at (x_0,ξ_0) such that

$$\int_0^T t^{2(2s-d/2)+\delta}\|c_1(x,D)u(t,.)\|^2_{H^{\min(\sigma+2,2s-d/2+1/2)}}dt < +\infty. \tag{2.14}$$

Finally, assume that $\partial_{\bar{u}}f(0,0) = 0$. Then, we do not need assumption (2.12)*.*

Remark. The assumptions (2.11) and (2.12) mean that u_0 and $\overline{u_0}$ are decreasing like $\langle x\rangle^{-(2s-d/2)}$ along $\{\varphi(s,x_0,\xi_0), s \le 0\}$ when $|x|$ converges towards infinity.

THEOREM 2.4. *Let (x_0,ξ_0) be such that (x_0,ξ_0) and $(x_0,-\xi_0)$ are not trapped backwards. Let a conic neighborhood $\mathcal{E}_1$ of $\{\varphi(s,x_0,\xi_0), s \le 0\}$ and a conic neighborhood $\mathcal{E}_2$ of $\{\varphi(s,x_0,-\xi_0), s \le 0\}$ be such that $\mathcal{E}_1$ and $\mathcal{E}_2$ satisfy* (2.7)*. Let $T > 0$, s in I_d, and u in $C(0,T,L^2(\mathbb{R}^d))$ be a solution of* (2.9)*. Assume there exists $\theta(x)$ in $S(0,0)$ equal to 1 on $\mathcal{E}_1 \cup \mathcal{E}_2$ such that θu is in $L^\infty(0,T,H^s)$ and $\langle x\rangle^s\theta u$ is in $L^\infty(0,T,L^2)$. Moreover, assume that for $j = 1,2$, there exists $s_j(x,\xi)$ in $S(0,2s-d/2)$ such that $\langle x\rangle^{2(2s-d/2)} \le s_j^2(x,\xi)$ on the set $\mathcal{E}_j$. Assume that u_0 satisfies*

$$\langle s_j(x,D)u_0, s_j(x,D)u_0\rangle < +\infty, \quad j = 1,2, \tag{2.15}$$

$$\langle s_j(x,D)\overline{u_0}, s_j(x,D)\overline{u_0}\rangle < +\infty, \quad j = 1,2. \tag{2.16}$$

Then, $u(t,.)$ is microlocally $H^{2s-d/2}$ at (x_0,ξ_0) for all $0 < t \le T$. Moreover, there exists a pseudodifferential operator with symbol $c(x,\xi)$ in S^0 elliptic at (x_0,ξ_0) such that

$$\int_0^T t^{2(2s-d/2)+\delta}\|c(x,D)u(t,.)\|^2_{H^{2s-d/2+1/2}}dt < +\infty \tag{2.17}$$

for all $\delta > 0$. Finally, suppose that $\partial_{\bar{u}}f(0,0) = 0$. Then, we do not need assumption (2.16)*.*

Theorem 2.4 implies, in particular, the following corollary.

COROLLARY 2.5. *Assume that no point of $T^*(\mathbb{R}^d)\setminus\{0\}$ is trapped backwards. Let $T > 0$, s in I_d, and u in $C(0,T,L^2(\mathbb{R}^d))$ be a solution of* (2.9)*. Assume that u is in $L^\infty(0,T,H^s)$ and $\langle x\rangle^s u$ is in $L^\infty(0,T,L^2)$. Moreover, assume that $\langle x\rangle^{2s-d/2}u_0$ is in L^2. Then, $u(t,.)$ is in $H^{2s-d/2}_{loc}$ for all $0 < t \le T$. Moreover, for any ϕ in $C_0^\infty(\mathbb{R}^d)$,*

$$\int_0^T t^{2(2s-d/2)+\delta}\|\phi u(t,.)\|^2_{H^{2s-d/2+1/2}}dt < +\infty \tag{2.18}$$

for all $\delta > 0$.

The reader interested in other applications of Theorem 2.4 is referred to [16], where we combine it with a reflection of singularities result to compute the Dirichlet-to-Neumann map of the nonlinear Schrödinger equation with the flat Laplacian outside a convex obstacle.

2.2. Outline of the proof. For the sake of simplicity, we assume here that u is in $L^\infty(0,T,H^s)$ and $\langle x\rangle^s u$ is in $L^\infty(0,T,L^2)$. The proof contains three steps.

- *First step.* We define

$$\lambda_1 = \partial_u f(0,0), \quad \lambda_2 = \partial_{\bar{u}}f(0,0), \quad g(u,\bar{u}) = f(u,\bar{u}) - \lambda_1 u - \lambda_2\bar{u}, \tag{2.19}$$

g vanishes to the second order at 0, and u is a solution of

(2.20)
$$i\frac{\partial u}{\partial t}+\frac{1}{2}\sum_{j,l=1}^{d}\partial_{x_j}a^{jl}(x)\partial_{x_l}u=\lambda_1 u+\lambda_2\bar{u}+g(u,\bar{u}),\quad 0<t<T,\quad x\in\mathbb{R}^d.$$

We paralinearize $g(u,\bar{u})$ using Bony's theorem [2]:

$$g(u,\bar{u})=T_{\frac{\partial g}{\partial u}(u,\bar{u})}u+T_{\frac{\partial g}{\partial \bar{u}}(u,\bar{u})}\bar{u}+r(u,\bar{u}).$$

$T_{\frac{\partial g}{\partial u}(u,\bar{u})}$ and $T_{\frac{\partial g}{\partial \bar{u}}(u,\bar{u})}$ are linear operators and $r(u,\bar{u})$ is a smooth remainder in the sense that it belongs to $L^\infty(0,T,H^{2s-d/2})$. Moreover, as $\langle x\rangle^s u$ is in $L^\infty(0,T,L^2)$ and as g vanishes to the second order at 0, we prove that $\langle x\rangle^{2s-d/2}r(u,\bar{u})$ is in $L^\infty(0,T,L^2)$. Finally, u is a solution of

(2.21)
$$i\frac{\partial u}{\partial t}+\frac{1}{2}\sum_{j,l=1}^{d}\partial_{x_j}a^{jl}(x)\partial_{x_l}u=\lambda_1 u+\lambda_2\bar{u}+T_{\frac{\partial g}{\partial u}(u,\bar{u})}u+T_{\frac{\partial g}{\partial \bar{u}}(u,\bar{u})}\bar{u}+r(u,\bar{u}).$$

- *Second step.* We want to adapt the strategy of [6] to (2.21). However, we do not know how to deal with the terms $\lambda_2\bar{u}$ and $T_{\frac{\partial g}{\partial \bar{u}}(u,\bar{u})}\bar{u}$. In fact a microlocal information on u at (x_0,ξ_0) implies for $\bar{u}$ a microlocal information at $(x_0,-\xi_0)$, but not at (x_0,ξ_0). Therefore, we introduce $v=u-Q\bar{u}$ and we choose the operator Q such that v satisfies the same kind of equation as u without term evaluated at $\bar{v}$:

$$i\frac{\partial v}{\partial t}+\frac{1}{2}\sum_{j,l=1}^{d}\partial_{x_j}a^{jl}(x)\partial_{x_l}v=Lv+r(u,\bar{u}), \tag{2.22}$$

  where L is a paradifferential operator.
- *Third step.* We adapt the strategy of [6], which relies on energy estimates, to (2.22). We have to check that the terms Lv and $r(u,\bar{u})$ coming from the nonlinearity do not perturb the energy estimates of [6] to obtain a smoothing effect for v. Finally, using the smoothing effect for v, the fact that $u=v+Q\bar{u}$ and the fact that Q is bounded from H^σ to $H^{\sigma+2}$ for any σ implies a smoothing effect for u.

Remark. Kenig, Ponce, and Vega [11] and Chihara [5] prove a smoothing effect for the nonlinear Schrödinger equation. Unlike Theorems 2.3 and 2.4, these results are not microlocal, but we would like to compare them with Corollary 2.5. In [11], the authors consider the second order operator $\sum_{j\le k}\partial_{x_j}^2-\sum_{j>k}\partial_{x_j}^2, 1\le k\le d$ (i.e., a constant coefficient not necessarily elliptic operator), and the nonlinearity contains also gradient terms, but they prove a gain of $1/2$ derivative, whereas we prove a gain of $s-d/2$ derivatives. The results in [5] give a gain of m derivatives for any integer m and the nonlinearity contains also gradient terms, but the results apply to the flat Laplacian with a nonlinearity vanishing at least at order 3 at 0, which in addition satisfies a gauge invariance. We rely on Bony's theorem, whereas these works rely on sharp estimates for linear nonhomogeneous Schrödinger equations. However, there are some similarities in these methods. For instance, Chihara [5] uses an argument similar to the second step of our proof to get rid of the term in $\bar{u}$, and Kenig, Ponce, and Vega [11] use a construction of symbols which has a link with that of Craig, Kappeler, and Strauss [6], reviewed in the following section, to obtain their sharp estimates.

3. Review of the results of Craig, Kappeler, and Strauss. The results in [6] rely on the construction of symbol pairs $(b(x,\xi), c(x,\xi))$ satisfying

$$\{a_2, b\} = \sum_{j=1}^{d} \partial_{\xi_j} a_2 \partial_{x_j} b - \partial_{x_j} a_2 \partial_{\xi_j} b = -c \tag{3.1}$$

($\{\,,\}$ is the Poisson bracket) and

$$b(x,\xi) \geq 0, \quad c(x,\xi) \geq 0. \tag{3.2}$$

The following proposition is proved in [6] and gives the existence of pairs of symbols $(b(x,\xi), c(x,\xi))$ satisfying (3.1) and (3.2).

PROPOSITION 3.1. *Let m and k be real and positive and let (x_0, ξ_0) not be trapped backwards. Let there be a conic neighborhood $\mathcal{E}$ of $\{\varphi(s, x_0, \xi_0), s \leq 0\}$ satisfying* (2.7). *There is a conic neighborhood $\mathcal{E}^0 \subset \mathcal{E}$ of $\{\varphi(s, x_0, \xi_0), s \leq 0\}$ satisfying* (2.7) *with a smaller $\epsilon_0 < \epsilon$ and a pair of symbols $b(x,\xi) \in S(m,k)$ and $c(x,\xi) \in S(m+1, k-1)$ with support in $\mathcal{E}$ such that* (3.1) *and* (3.2) *hold and $b(x_0,\xi_0) = c(x_0,\xi_0) = 1$. We can choose b and c such that $\sqrt{b} \in S(m/2, k/2)$ and $\sqrt{c} \in S((m+1)/2, (k-1)/2)$. Moreover, there is a constant C such that on the set $\mathcal{E}^0 \cap \{|\xi| \geq 1\}$ they satisfy*

$$\langle \xi \rangle^m \langle x \rangle^k \leq b(x,\xi), \quad \frac{1}{C} \langle \xi \rangle^{m+1} \langle x \rangle^{k-1} \leq c(x,\xi). \tag{3.3}$$

If $k = 0$, then there is the exception that for any $\nu < -1$ we may choose $c(x,\xi)$ in $S(m+1,-1)$ such that $1/C\langle \xi \rangle^{m+1} \langle x \rangle^{\nu} \leq c(x,\xi)$.

We recall some properties of the algebra $S(m,k)$ that we use in what follows (see [6] or [9] for a more general treatment). As $S(m,0) \subset S^m$, where S^m is the usual pseudodifferential algebra, $a(x,D)$ is bounded from H^t to H^{t-m} and from C^t to C^{t-m} for all $a(x,\xi)$ in $S(m,0)$ and for all t (when t is an integer, C^t is the Zygmund class C^t_* [10]). The following proposition is on the composition of operators.

PROPOSITION 3.2. *Let a in $S(m,k)$ and b in $S(m_1,k_1)$ be two symbols. Let N be an integer. Then*

$$a(x,D)b(x,D) = \sum_{|\alpha|<N} \frac{1}{i^{|\alpha|}\alpha!} \big(\partial_\xi^\alpha a \partial_x^\alpha b\big)(x,D) + r_N(x,D), \tag{3.4}$$

where r_N is in $S(m_1+m_2-N, k_1+k_2-N)$. Moreover, the seminorms of r_N can be estimated by a sum of products of seminorms of $\partial_\xi^\alpha a$ in $S(m_1-N,k_1)$ and of $\partial_x^\alpha b$ in $S(m_2, k_2-N)$, where $|\alpha| = N$.

We prove this proposition as it is more precise than the result established in [6].

Proof. From the usual procedures in the pseudodifferential calculus, we have

$$\begin{aligned} a(x,D)b(x,D)u &= \int\!\!\int e^{i\eta(x-z)} \int\!\!\int e^{i\xi(z-y)} a(x,\eta) b(z,\xi) u(y)dy\, d\xi\, dz\, d\eta \\ &= \sum_{|\alpha|<N} \frac{1}{i^{|\alpha|}\alpha!} \int\!\!\int e^{i(x-y)\xi} \partial_\xi^\alpha a(x,\xi) \partial_x^\alpha b(x,\xi) u(y) dy\, d\xi \\ &\quad + r_N(x,D)u, \end{aligned} \tag{3.5}$$

where

$$r_N[a,b](x,\xi) = r_N(x,\xi) = \sum_{|\alpha|=N} \frac{N}{i^{|\alpha|}\alpha!}\int_0^1\int\int e^{i(\xi-\eta)(z-x)}t^{N-1}\partial_\xi^\alpha a(x,\xi+t(\eta-\xi))\partial_x^\alpha b(z,\xi)dz\,d\eta\,dt. \tag{3.6}$$

As $\partial_x r_N[a,b] = r_N[\partial_x a,b] + r_N[a,\partial_x b]$ and $\partial_\xi r_N[a,b] = r_N[\partial_\xi a,b] + r_N[a,\partial_\xi b]$, it suffices to prove that

$$|r_N[a,b](x,\xi)| \le C\langle\xi\rangle^{m_1+m_2-N}\langle x\rangle^{k_1+k_2-N}, \tag{3.7}$$

where C is estimated by sums of products of seminorms of $\partial_\xi^\alpha a$ in $S(m_1-N,k_1)$ and of $\partial_x^\alpha b$ in $S(m_2,k_2-N)$ with $|\alpha|=N$. Then let χ be in $C_0^\infty(\mathbb{R}^d)$ such that $\chi(s)=1$ if $|s|\le 1/9$ and $\chi(s)=0$ if $|s|\ge 1/4$. Using

$$\begin{aligned}1 = {}&\chi(|\xi-\eta|^2/\langle\xi\rangle^2)\chi(|x-z|^2/\langle x\rangle^2)\\ &+\chi(|\xi-\eta|^2/\langle\xi\rangle^2)(1-\chi(|x-z|^2/\langle x\rangle^2))\\ &+(1-\chi(|\xi-\eta|^2/\langle\xi\rangle^2))\chi(|x-z|^2/\langle x\rangle^2)\\ &+(1-\chi(|\xi-\eta|^2/\langle\xi\rangle^2))(1-\chi(|x-z|^2/\langle x\rangle^2)),\end{aligned}$$

it suffices to prove (3.7) for each of the following cases:

$$|\xi-\eta|\le 1/2\langle\xi\rangle \quad\text{and}\quad |x-z|\le 1/2\langle x\rangle, \tag{3.8}$$

$$|\xi-\eta|\le 1/2\langle\xi\rangle \quad\text{and}\quad |x-z|\ge 1/3\langle x\rangle, \tag{3.9}$$

$$|\xi-\eta|\ge 1/3\langle\xi\rangle \quad\text{and}\quad |x-z|\le 1/2\langle x\rangle, \tag{3.10}$$

$$|\xi-\eta|\ge 1/3\langle\xi\rangle \quad\text{and}\quad |x-z|\ge 1/3\langle x\rangle. \tag{3.11}$$

When $|x-z|\le 1/2\langle x\rangle$, we have $\langle z\rangle\simeq\langle x\rangle$. Let P be an integer, $P>d/2$. We use $(1+|x-z|^{2P})^{-1}(1+(-\Delta_\eta)^P)e^{i(\xi-\eta)(z-x)} = e^{i(\xi-\eta)(z-x)}$, integrations by parts, and the fact that $\int(1+|x-z|^{2P})^{-1}dz<+\infty$.

When $|x-z|\ge 1/3\langle x\rangle$, we have $\langle z\rangle\le\langle x\rangle+\langle z-x\rangle$. Let $P>d/2$. We use $|x-z|^{-2P}(-\Delta_\eta)^P e^{i(\xi-\eta)(z-x)} = e^{i(\xi-\eta)(z-x)}$, integrations by parts, and the fact that

$$\int_{|x-z|\ge 1/3\langle x\rangle}|x-z|^{-2P}dz\le C\langle x\rangle^{d-2P}.$$

We deal with the integrations in η in the same way, replacing x by ξ and z by η. Moreover, for $0\le t\le 1$, $\langle\xi+t(\eta-\xi)\rangle\le\langle\xi\rangle+\langle\eta-\xi\rangle$ and $\langle\xi+t(\eta-\xi)\rangle\simeq\langle\xi\rangle$ when $|\xi-\eta|\le 1/2\langle\xi\rangle$. Finally, for an integer $P>d/2+\max(|k_2-N|,|m_1-N|)/2$,

$$\begin{aligned}&|r_N[a,b](x,\xi)|\le C\langle\xi\rangle^{m_1+m_2-N}\langle x\rangle^{k_1+k_2-N} && \text{in (3.8)},\\ &|r_N[a,b](x,\xi)|\le C\langle\xi\rangle^{m_1+m_2-N}\langle x\rangle^{k_1+2|k_2-N|+d-2P} && \text{in (3.9)},\\ &|r_N[a,b](x,\xi)|\le C\langle\xi\rangle^{2|m_1-N|+d-2P+m_2}\langle x\rangle^{k_1+k_2-N} && \text{in (3.10)},\\ &|r_N[a,b](x,\xi)|\le C\langle\xi\rangle^{2|m_1-N|+d-2P+m_2}\langle x\rangle^{k_1+2|k_2-N|+d-2P} && \text{in (3.11)},\end{aligned}$$

which yields (3.7) by taking $P\ge d/2+3/2\max(|k_2-N|,|m_1-N|)$. ∎

PROPOSITION 3.3. *Let a be in $S(m,k)$. The adjoint operator satisfies*

$$a(x,D)^* = \sum_{|\alpha|<N} \frac{1}{i^{|\alpha|}\alpha!}\big(\partial_\xi^\alpha \partial_x^\alpha \bar{a}\big)(x,D) + R_N, \tag{3.12}$$

where N is an integer such that $N \geq \max(k,m)$ and R_N is a bounded operator on L^2.

The following lemmas are proved in [6].

LEMMA 3.4. *Let $b(x,\xi)$ be in $S(m,k)$ such that $b(x,\xi) = s^2(x,\xi)$ with $s(x,\xi)$ in $S(m/2,k/2)$. Then there exists $e(x,\xi)$ in $S(m-1,k-1)$ with $\operatorname{supp}(e) \subset \operatorname{supp}(b)$ such that*

$$\operatorname{Re}\langle u, b(x,D)u\rangle \geq \|s(x,D)u\|_{L^2}^2 + \operatorname{Re}\langle u, e(x,D)u\rangle - C\|u\|_{L^2}^2,$$

where $\langle , \rangle$ is the scalar product in $L^2(\mathbb{R}^d)$.

LEMMA 3.5. *Let $b_0(x,\xi)$ be in $S(m,k)$ such that $0 \leq b_0(x,\xi) = s_0^2(x,\xi)$ with s_0 in $S(m/2,k/2)$. Define the set*

$$\mathcal{E}^0 = \{(x,\xi)/\langle x\rangle^{-k}\langle \xi\rangle^{-m} b_0(x,\xi) \geq 1\}.$$

Suppose that b in $S(m,k)$ has support in $\mathcal{E}^0$. Then there exists $c(x,\xi)$ in $S(0,0)$ and R bounded on L^2 such that

$$b(x,D) = s_0^*(x,D)c(x,D)s_0(x,D) + R.$$

4. Technical lemmas. In this section, we prove, among other things, Lemma 4.7 and Proposition 4.12, which are useful for bounding the terms coming from the nonlinearity in the energy estimates of the third step of the proof. Moreover, in order to construct the operator Q in the second step of the proof, we define a paradifferential algebra and study its properties in Proposition 4.11. In what follows, the various spaces are all defined on $\mathbb{R}^d$.

The following proposition will be used when proving Proposition 4.3, Lemma 4.7, and Proposition 4.8.

PROPOSITION 4.1. *Let k and l be two real numbers with $l \geq 0$ and let p be an integer. Let $\theta \in C_0^\infty(\mathbb{R}^d)$, let $a(x)$ be in $S(0,k)$ and let j be an integer such that $j \geq l$. Then*

$$\theta(2^{-p}D)a(x) = \sum_{0\leq|\alpha|<j} \frac{2^{-p|\alpha|}}{i^{|\alpha|}\alpha!}\partial_x^\alpha a(x)\big(\partial_\xi^\alpha\theta\big)(2^{-p}D) + \frac{\langle x\rangle^{k-j}}{2^{p(j-l)}}R_1 \tag{4.1}$$

and

$$a(x)\theta(2^{-p}D) = \sum_{0\leq|\alpha|<j} \frac{2^{-p|\alpha|}i^{|\alpha|}}{\alpha!}\big(\partial_\xi^\alpha\theta\big)(2^{-p}D)\partial_x^\alpha a(x) + \frac{\langle x\rangle^{k-j}}{2^{p(j-l)}}R_2, \tag{4.2}$$

where R_1 and R_2 are continuous from H^t to H^{t+l} and from C^t to C^{t+l} for all t with norms independent of p.

Proof. $2^{p(j-l)}\langle x\rangle^{j-k}\theta(2^{-p}\xi)$ is in $S(j-l,j-k)$ and $a(x)$ is in $S(0,k)$. Therefore, Proposition 3.2 implies

$$\begin{aligned} &2^{p(j-l)}\langle x\rangle^{j-k}\theta(2^{-p}D)a(x) \\ &\quad= \sum_{0\leq|\alpha|<j} \frac{2^{-p|\alpha|}}{i^{|\alpha|}\alpha!}2^{p(j-l)}\langle x\rangle^{j-k}\partial_x^\alpha a(x)\big(\partial_\xi^\alpha\theta\big)(2^{-p}D) + r(x,D), \end{aligned} \tag{4.3}$$

where r is in $S(-l,0)$ with its seminorms bounded by sums of products of seminorms of $\partial_\xi^\alpha(2^{p(j-l)}\langle x\rangle^{j-k}\theta(2^{-p}\xi))$ in $S(-l,j-k)$ and of $\partial_x^\alpha a(x)$ in $S(0,k-j)$ with $|\alpha|=j$. For $|\alpha|=j$, $\partial_\xi^\alpha(2^{p(j-l)}\langle x\rangle^{j-k}\theta(2^{-p}\xi)) = 2^{-pl}\langle x\rangle^{j-k}(\partial_\xi^\alpha\theta)(2^{-p}\xi)$ has its seminorms in $S(-l,j-k)$ bounded independently of p. Therefore, $r(x,\xi)$ has its seminorms in $S(-l,0)$ (and hence in S^{-l}) bounded independently of p. Multiplying (4.3) by $2^{p(l-j)}\langle x\rangle^{k-j}$,

$$\theta(2^{-p}D)a(x) = \sum_{0\le|\alpha|<j}\frac{2^{-p|\alpha|}}{i^{|\alpha|}\alpha!}\partial_x^\alpha a(x)\big(\partial_\xi^\alpha\theta\big)(2^{-p}D) + \frac{\langle x\rangle^{k-j}}{2^{p(j-l)}}r(x,D). \tag{4.4}$$

Thus, (4.1) is satisfied with $R_1 = r(x,D)$. Moreover, R_1 is continuous from H^t to H^{t+l} and from C^t to C^{t+l} for all t with norms independent of p since the seminorms of $r(x,\xi)$ in S^{-l} are bounded independently of p.

Equation (4.4) with $\bar a$ and $\overline\theta$ gives

$$\overline\theta(2^{-p}D)\bar a(x) = \sum_{0\le|\alpha|<j}\frac{2^{-p|\alpha|}}{i^{|\alpha|}\alpha!}\partial_x^\alpha \bar a(x)\big(\partial_\xi^\alpha\overline\theta\big)(2^{-p}D) + \frac{\langle x\rangle^{k-j}}{2^{p(j-l)}}r_1(x,D), \tag{4.5}$$

where $r_1(x,\xi)$ has seminorms in $S(-l,0)$ bounded independently of p. Using Proposition 3.2, $\langle x\rangle^{k-j}r_1(x,D)\langle x\rangle^{j-k} = r_2(x,D)$ and $r_2(x,\xi)$ has seminorms in $S(-l,0)$ bounded independently of p. Equation (4.5) becomes

$$\overline\theta(2^{-p}D)\bar a(x) = \sum_{0\le|\alpha|<j}\frac{2^{-p|\alpha|}}{i^{|\alpha|}\alpha!}\partial_x^\alpha \bar a(x)\big(\partial_\xi^\alpha\overline\theta\big)(2^{-p}D) + 2^{-p(j-l)}r_2(x,D)\langle x\rangle^{k-j} \tag{4.6}$$

and, taking the adjoint,

$$a(x)\theta(2^{-p}D) = \sum_{0\le|\alpha|<j}\frac{2^{-p|\alpha|}i^{|\alpha|}}{\alpha!}\big(\partial_\xi^\alpha\theta\big)(2^{-p}D)\partial_x^\alpha a(x) + \frac{\langle x\rangle^{k-j}}{2^{p(j-l)}}r_2(x,D)^*. \tag{4.7}$$

Thus, (4.2) is satisfied with $R_2 = r_2(x,D)^*$. The properties of S^{-l} imply $r_2(x,D)^* = r_3(x,D)$ with r_3 in S^{-l}. R_2 is continuous from H^t to H^{t+l} and from C^t to C^{t+l} for all t with norms independent of p since $r_3(x,\xi)$ has seminorms in S^{-l} bounded independently of p. $\square$

Following Meyer [13], we introduce a positive function φ in $C_0^\infty(\mathbb{R}^d)$ equal to 1 for $|\xi|\le 1/2$ and to 0 for $|\xi|\ge 1$. For all integers p we define S_p by $\widehat{S_pu}(\xi)=\varphi(2^{-p}\xi)\hat u(\xi)$. We also define Δ_p by $\Delta_p = S_{p+1}-S_p$, i.e., by $\widehat{\Delta_pu}(\xi)=\psi(2^{-p}\xi)\hat u(\xi)$, where $\psi(\xi)=\varphi(\xi/2)-\varphi(\xi)$. Littlewood–Paley decomposition of $u\in\mathcal{S}'(\mathbb{R}^d)$ is

$$u=\sum_{p\ge -1}\Delta_p(u),$$

where the sum converges for the topology of $\mathcal{S}'(\mathbb{R}^d)$ and $\Delta_{-1}=S_0$. We have a simple characterization of the spaces H^s and C^s (see, for example, [4]),

$$H^s = \left\{u\in\mathcal{S}'(\mathbb{R}^d)\Big/\sum_{p\ge-1}4^{ps}\|\Delta_pu\|_{L^2}^2<+\infty\right\} \tag{4.8}$$

and $(\sum_{p\geq -1} 4^{ps}\|\Delta_p u\|_{L^2}^2)^{1/2}$ is equivalent to the usual norm on H^s.

$$C^s = \left\{ u \in \mathcal{S}'(\mathbb{R}^d) / \sup_{p\geq -1} 2^{ps}\|\Delta_p u\|_{L^\infty} < +\infty \right\}, \tag{4.9}$$

and $\sup_{p\geq -1} 2^{ps}\|\Delta_p u\|_{L^\infty}$ is equivalent to the usual norm on C^s. Moreover, let $0 < r_1 < r_2$ and suppose $(u_q)_{q\in\mathbb{N}}$ is a sequence in $\mathcal{S}'(\mathbb{R}^d)$. Then

$$\text{supp } \hat{u}_q \subset \{r_1 2^q \leq |\xi| \leq r_2 2^q\} \text{ and } \sum_{q\geq 0} 4^{qs}\|u_q\|_{L^2}^2 < +\infty \Rightarrow u = \sum_{q\geq 0} u_q \in H^s, \tag{4.10}$$

and if $s > 0$,

$$\text{supp } \hat{u}_q \subset \{|\xi| \leq r_2 2^q\} \text{ and } \sum_{q\geq 0} 4^{qs}\|u_q\|_{L^2}^2 < +\infty \Rightarrow u = \sum_{q\geq 0} u_q \in H^s. \tag{4.11}$$

The following definition will be useful in defining a paradifferential algebra well-suited to our problem.

Definition 4.2. *Let s and s' be two real numbers such that $s > 0$. The space $H^s_{s'}$ is defined by*

$$H^s_{s'} = \{u \in H^{s+s'} / \langle x\rangle^s u \in H^{s'}\}.$$

The following proposition, among others, will be used in the third step of the proof to bound the remainder term in the energy estimates.

Proposition 4.3. *Let s and s' be two real numbers such that $s > 0$. Let u in $H^s_{s'}$. Then for all $0 \leq \theta \leq 1$, $\langle x\rangle^{\theta s} u$ is in $H^{(1-\theta)s+s'}$.*

Proof. Using (4.1) with $k = \theta s$, $l = (s+s')_-$, where $(s+s')_- = \max(-(s+s'), 0)$, $j = [\max(s, |s+s'|)] + 1$, $a(x) = \langle x\rangle^{\theta s}$, and $\chi = \psi$ yields

$$\Delta_p \langle x\rangle^{\theta s} = \sum_{0\leq|\alpha|<j} \frac{2^{-p|\alpha|}}{i^{|\alpha|}\alpha!} \partial_x^\alpha \langle x\rangle^{\theta s} \big(\partial_\xi^\alpha \psi\big)(2^{-p}D) + \frac{\langle x\rangle^{\theta s - j}}{2^{p(j-l)}} R, \tag{4.12}$$

where R is bounded from $H^{s+s'}$ to L^2 with a norm independent of p. Thus

$$\begin{aligned} &2^{p(1-\theta)s+ps'}\|\Delta_p \langle x\rangle^{\theta s} u\|_{L^2} \\ &\leq \sum_{0\leq|\alpha|<j} \frac{2^{p((1-\theta)s+s'-|\alpha|)}}{\alpha!} \|\partial_x^\alpha \langle x\rangle^{\theta s} \big(\partial_\xi^\alpha \psi\big)(2^{-p}D) u\|_{L^2} \\ &\quad + 2^{-p(j-l-(1-\theta)s-s')}\|\langle x\rangle^{\theta s - j} R u\|_{L^2}. \end{aligned} \tag{4.13}$$

As $\theta s - [s] - 1 \leq 0$,

$$2^{-p(j-l-(1-\theta)s-s')}\|\langle x\rangle^{\theta s-j} R u\|_{L^2} \leq 2^{p(s+s'-[|s+s'|]-1)} C\|Ru\|_{L^2} \leq C\varepsilon_p \|u\|_{H^{s+s'}}, \tag{4.14}$$

where $(\epsilon_p)_{p\geq -1}$ is in $l^2(\mathbb{N})$ because $s + s' - [|s+s'|] - 1 < 0$. Moreover, by Hölder and Young inequalities

$$\begin{aligned} &2^{p((1-\theta)s+s')}\|\partial_x^\alpha \langle x\rangle^{\theta s} \big(\partial_\xi^\alpha \psi\big)(2^{-p}D) u\|_{L^2} \\ &\leq 2^{p((1-\theta)s+s')}\|\big(\partial_x^\alpha \langle x\rangle^{\theta s}\big)^{1/\theta} \big(\partial_\xi^\alpha \psi\big)(2^{-p}D) u\|_{L^2}^\theta \|\big(\partial_\xi^\alpha \psi\big)(2^{-p}D) u\|_{L^2}^{1-\theta} \\ &\leq C 2^{ps'}\|\langle x\rangle^{s-|\alpha|/\theta} \big(\partial_\xi^\alpha \psi\big)(2^{-p}D) u\|_{L^2} + C 2^{ps+ps'}\|\big(\partial_\xi^\alpha \psi\big)(2^{-p}D) u\|_{L^2}. \end{aligned} \tag{4.15}$$

Looking at the support of $\partial_\xi^\alpha\psi$ we have

$$\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)=\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)(\Delta_{p-1}+\Delta_p+\Delta_{p+1}).$$

Thus, as u is in $H^{s+s'}$,

$$\begin{aligned}(4.16)\qquad 2^{p(s+s')}\big\|\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)u\big\|_{L^2}\le C2^{p(s+s')}(&\|\Delta_{p-1}u\|_{L^2}\\ &+\|\Delta_p u\|_{L^2}+\|\Delta_{p+1}u\|_{L^2})\le\varepsilon_p,\end{aligned}$$

where $(\epsilon_p)_{p\ge-1}$ is in $l^2(\mathbb{N})$ by (4.8).

$$\big\|\langle x\rangle^{s-|\alpha|/\theta}\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)u\big\|_{L^2}\le C\big\|\langle x\rangle^{s}\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)u\big\|_{L^2}.$$

Using (4.8), (4.13), (4.14), (4.15), and (4.16), it suffices to prove

$$(4.17)\qquad 2^{ps'}\big\|\langle x\rangle^{s}\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)u\big\|_{L^2}\le\epsilon_p,$$

where $(\epsilon_p)_{p\ge-1}$ is in $l^2(\mathbb{N})$. Using (4.2) with $k=s$, $l=s'_-$, $j=[\max(s,|s'|)]+1$, $a(x)=\langle x\rangle^s$, and $\chi=\partial_\xi^\alpha\psi$,

$$\begin{aligned}(4.18)\qquad \langle x\rangle^{s}\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)=\sum_{0\le|\beta|<j}&\frac{2^{-p|\beta|}i^{|\beta|}}{\beta!}\big(\partial_\xi^{\alpha+\beta}\psi\big)(2^{-p}D)\partial_x^\beta\langle x\rangle^s\\ &+2^{-p(j-l)}\langle x\rangle^{s-j}R,\end{aligned}$$

where R is bounded from $H^{s'}$ to L^2 with a norm independent of p. Thus

$$\begin{aligned}(4.19)\qquad &2^{ps'}\big\|\langle x\rangle^{s}\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)u\big\|_{L^2}\\ &\le\sum_{0\le|\beta|<j}\frac{2^{-p|\beta|+ps'}}{\beta!}\big\|\big(\partial_\xi^{\alpha+\beta}\psi\big)(2^{-p}D)\partial_x^\beta(\langle x\rangle^s)u\big\|_{L^2}\\ &\quad+2^{-p(j-l-s')}\|\langle x\rangle^{s-j}Ru\|_{L^2}.\end{aligned}$$

As $s-j\le0$,

$$\begin{aligned}(4.20)\qquad 2^{-p(j-l-s')}\|\langle x\rangle^{s-j}Ru\|_{L^2}&\le C2^{-p([|s'|]+1-|s'|)}\|Ru\|_{L^2}\\ &\le C2^{-p([|s'|]+1-|s'|)}\|u\|_{H^{s'}}\le\varepsilon_p,\end{aligned}$$

where $(\epsilon_p)_{p\ge-1}$ is in $l^2(\mathbb{N})$. Looking at the support of $\partial_\xi^{\alpha+\beta}\psi$ we have

$$\begin{aligned}(4.21)\qquad 2^{ps'}\big\|\big(\partial_\xi^{\alpha+\beta}\psi\big)(2^{-p}D)\partial_x^\beta(\langle x\rangle^s)u\big\|_{L^2}&\le C2^{ps'}\big(\big\|\Delta_{p-1}\partial_x^\beta(\langle x\rangle^s)u\big\|_{L^2}\\ &+\big\|\Delta_{p}\partial_x^\beta(\langle x\rangle^s)u\big\|_{L^2}+\big\|\Delta_{p+1}\partial_x^\beta(\langle x\rangle^s)u\big\|_{L^2}\big).\end{aligned}$$

Since $\langle x\rangle^s u$ is in $H^{s'}$ and the multiplication by $(\partial_x^\beta\langle x\rangle^s)\langle x\rangle^{-s}$ is bounded on L^2, we have from (4.8) and (4.21) that

$$(4.22)\qquad 2^{ps'}\big\|\big(\partial_\xi^{\alpha+\beta}\psi\big)(2^{-p}D)\partial_x^\beta(\langle x\rangle^s)u\big\|_{L^2}\le\epsilon_p,$$

where $(\epsilon_p)_{p\ge-1}$ is in $l^2(\mathbb{N})$. Finally, (4.19), (4.20), and (4.22) imply (4.17). ∎

We study the action of operators with symbol in $S(m,k)$ on the spaces $H^s_{s'}$.

LEMMA 4.4. *Let m and k with $k \le 0$ be two real numbers, and let $b(x,\xi)$ be in $S(m,k)$. Then $b(x,D)$ is bounded from $H^s_{s'}$ to $H^{s-k}_{s'-m+k}$.*

Proof. As $k \le 0$, $S(m,k)$ is included in S^m. So $b(x,D)$ is bounded from $H^{s+s'}$ to $H^{s+s'-m}$. As $H^{s+s'-m}$ is included in $H^{s+s'-m+k}$, $b(x,D)$ is bounded from $H^{s+s'}$ to $H^{s+s'-m+k}$. Moreover, $\langle x\rangle^{s-k} b(x,D)\langle x\rangle^{-s} = b_1(x,D)$ with $b_1(x,\xi)$ in $S(m,0)$ by using Proposition 3.2. Thus, $\langle x\rangle^{s-k} b(x,D)\langle x\rangle^{-s}$ is bounded from $H^{s'}$ to $H^{s'-m}$. As $H^{s'-m}$ is included in $H^{s'-m+k}$, $\langle x\rangle^{s-k} b(x,D)\langle x\rangle^{-s}$ is bounded from $H^{s'}$ to $H^{s'-m+k}$. $\square$

The following corollary gives an example of a solution u of the nonlinear Schrödinger equation (4.23) such that u is in $L^\infty(0,T,H^s)$ and $\langle x\rangle^s u$ is in $L^\infty(0,T,L^2)$, therefore satisfying the assumptions of Corollary 2.5.

COROLLARY 4.5. *Let $s > d/2$ and let u_0 be in $H^s(\mathbb{R}^d)$. Then, for $T > 0$ sufficiently small, there exists a unique solution u in $C(0,T,H^s)$ of*

$$\begin{cases} i\dfrac{\partial u}{\partial t} + \dfrac{1}{2}\displaystyle\sum_{j,l=1}^{d} \partial_{x_j} a^{jl}(x)\partial_{x_l} u = f(u,\bar u), \quad 0 < t < T, \quad x \in \mathbb{R}^d, \\ u|_{t=0} = u_0, \end{cases} \tag{4.23}$$

where $A = -1/2\sum_{j,l=1}^{d} \partial_{x_j} a^{jl}(x)\partial_{x_l}$ is an elliptic self-adjoint operator and f is C^∞ and vanishes at 0. Moreover, assume the symbol of A is in $S(2,0)$ and $\langle x\rangle^s u_0$ is in L^2. $\langle x\rangle^s u$ is then in $L^\infty(0,T,L^2)$.

Proof. The local existence and uniqueness in $C(0,T,H^s)$ of the u solution of (4.23) is well known. We recall the proof. Since A is an unbounded self-adjoint operator, $-iA$ is the generator of a semigroup of isometry on L^2. Let L be the operator on $C(0,T,H^s)$ defined by

$$Lu(t) = e^{-itA}u_0 - i\int_0^t e^{-i(t-\tau)A} f(u(\tau),\bar u(\tau))d\tau$$

for $0 \le t \le T$. As e^{-iAt} commutes with $(I+A)^{s/2}$ on H^s and as

$$C_1\|u\|_{H^s} \le \left\|(I+A)^{\frac{s}{2}}u\right\|_{L^2} \le C_2\|u\|_{H^s}$$

for u in H^s,

$$\begin{aligned} &\|Lu\|_{C(0,T,H^s)} \le C\big(\|u_0\|_{H^s} + T\|f(u,\bar u)\|_{C(0,T,H^s)}\big), \\ &\|Lu - Lv\|_{C(0,T,H^s)} \le CT\|f(u,\bar u) - f(v,\bar v)\|_{C(0,T,H^s)}. \end{aligned} \tag{4.24}$$

Moreover, f is C^∞ and vanishes at 0 and H^s is embedded in L^∞. Thus (see, for example, [1])

$$\|f(u,\bar u)\|_{H^s} \le \theta(\|u\|_{H^s}), \tag{4.25}$$

where θ is an increasing function. We define the C^∞ function g by

$$g(u,\bar u) = f(u,\bar u) - \partial_u f(0,0)u - \partial_{\bar u} f(0,0)\bar u.$$

H^s is an algebra and $\partial_u g$ and $\partial_{\bar u} g$ vanish at 0. Thus, by (4.25),

$$\begin{aligned} \|g(u,\bar u) - g(v,\bar v)\|_{H^s} &\le C\int_0^1 (\|\partial_u g(ru+(1-r)v, r\bar u + (1-r)\bar v)\|_{H^s} \\ &\quad + \|\partial_{\bar u} g(ru+(1-r)v, r\bar u+(1-r)\bar v)\|_{H^s})dr\|u-v\|_{H^s} \\ &\le \theta(\|u\|_{H^s} + \|v\|_{H^s})\|u-v\|_{H^s}, \end{aligned} \tag{4.26}$$

where θ is an increasing function. We have

$$\|\partial_u f(0,0)u + \partial_{\bar u} f(0,0)\bar u - \partial_u f(0,0)v - \partial_{\bar u} f(0,0)\bar v\|_{H^s} \le C\|u-v\|_{H^s},$$

which together with (4.26) yields

$$\|f(u,\bar u) - f(v,\bar v)\|_{H^s} \le \theta(\|u\|_{H^s} + \|v\|_{H^s})\|u-v\|_{H^s}, \tag{4.27}$$

where θ is an increasing function. Inequalities (4.24), (4.25), and (4.27) imply

$$\begin{aligned}&\|Lu\|_{C(0,T,H^s)} \le C\big(\|u_0\|_{H^s} + T\theta(\|u\|_{C(0,T,H^s)})\big),\\ &\|Lu - Lv\|_{C(0,T,H^s)} \le CT\theta\big(\|u\|_{C(0,T,H^s)} + \|v\|_{C(0,T,H^s)}\big)\|u-v\|_{C(0,T,H^s)}.\end{aligned} \tag{4.28}$$

We choose

$$0 < T \le \min(\|u_0\|_{H^s}(1+\theta(2C\|u_0\|_{H^s}))^{-1}, (1+C\theta(4C\|u_0\|_{H^s}))^{-1}),$$

and we use the contraction mapping principle with the operator L and the function space

$$\big\{v \in C(0,T,H^s)\,/\,\|v\|_{C(0,T,H^s)} \le 2C\|u_0\|_{H^s}\big\}.$$

So, we have the existence and uniqueness in $C(0,T,H^s)$ of the solution u of (4.23).

Moreover, suppose that the symbol of A is in $S(2,0)$ and $\langle x\rangle^s u_0$ is in L^2. For any operator B,

$$\begin{aligned}\partial_t\langle Bu,u\rangle &= \langle B\partial_t u,u\rangle + \langle Bu,\partial_t u\rangle + \left\langle \frac{\partial B}{\partial t}u,u\right\rangle\\ &= \langle B(-iAu - if(u,\bar u)),u\rangle + \langle Bu,-iAu-if(u,\bar u)\rangle + \left\langle \frac{\partial B}{\partial t}u,u\right\rangle\\ &= -i\langle [B,A]u,u\rangle - i\langle Bf(u,\bar u),u\rangle + i\langle Bu,f(u,\bar u)\rangle + \left\langle \frac{\partial B}{\partial t}u,u\right\rangle.\end{aligned} \tag{4.29}$$

Choosing $Bu = \langle x\rangle^{2s}u$ and integrating (4.29) between 0 and t, we obtain

$$\begin{aligned}\|\langle x\rangle^s u(t)\|_{L^2}^2 \le \|\langle x\rangle^s u_0\|_{L^2}^2 &+ \int_0^t |\langle [B,A]u(\tau),u(\tau)\rangle| d\tau\\ +2\int_0^t \|\langle x\rangle^s f(u(\tau),\bar u(\tau))\|_{L^2}&\|\langle x\rangle^s u(\tau)\|_{L^2} d\tau.\end{aligned} \tag{4.30}$$

The mean value theorem and the fact that f vanishes at 0 imply

$$|f(u(t,x),\bar u(t,x))| \le \theta\big(\|u\|_{L^\infty(0,T,L^\infty(\mathbb{R}^d))}\big)|u(t,x)|.$$

Thus

$$\begin{aligned}&\int_0^t \|\langle x\rangle^s f(u(\tau),\bar u(\tau))\|_{L^2}\|\langle x\rangle^s u(\tau)\|_{L^2} d\tau\\ &\qquad\le \theta(\|u\|_{L^\infty})\int_0^t \|\langle x\rangle^s u(\tau)\|_{L^2}^2 d\tau.\end{aligned} \tag{4.31}$$

As A is in $Op(S(2,0))$ and B is in $Op(S(0,2s))$, $[A,B]$ is in $Op(S(1,2s-1))$ by Proposition 3.2. Therefore, by Proposition 3.2, $\langle x\rangle^{-s+1/2}[A,B]\langle x\rangle^{-s+1/2}=R$, where R is in $Op(S(1,0))$. Thus,

$$\begin{aligned} |\langle [B,A]u(\tau),u(\tau)\rangle| &= |\langle R\langle x\rangle^{s-1/2}u(\tau),\langle x\rangle^{s-1/2}u(\tau)\rangle| \\ &\le \|R\langle x\rangle^{s-1/2}u(\tau)\|_{H^{-1/2}}\|\langle x\rangle^{s-1/2}u(\tau)\|_{H^{1/2}} \\ &\le C\|\langle x\rangle^{s-1/2}u(\tau)\|^2_{H^{1/2}}. \end{aligned} \tag{4.32}$$

Moreover, as $s>d/2\ge 1/2$, Proposition 4.3 implies

$$\big\|\langle x\rangle^{s-1/2}u(\tau)\big\|^2_{H^{1/2}} \le C\big(\|\langle x\rangle^{s}u(\tau)\|^2_{L^2}+\|u(\tau)\|^2_{H^s}\big). \tag{4.33}$$

$\langle x\rangle^s u_0$ is in L^2 and u is in $C(0,T,H^s)$ and in $C(0,T,L^\infty)$ as $s>d/2$. Thus, (4.30), (4.31), (4.32), and (4.33) imply

$$\|\langle x\rangle^s u(t)\|^2_{L^2}\le C+C\int_0^t \|\langle x\rangle^s u(\tau)\|^2_{L^2}d\tau. \tag{4.34}$$

Therefore, $\langle x\rangle^s u$ is in $L^\infty(0,T,L^2)$ by Gronwall's lemma. $\square$

The following corollary will be useful to construct the operator Q in the second step of the proof.

COROLLARY 4.6. *Let $s>d/2$ and $T>0$ be two real numbers, and let there be a solution u of* (4.23). *We assume that the symbol of A is in $S(2,0)$ and that there exists two real functions $\theta_1(x)$ and $\theta_2(x)$ in $S(0,0)$ such that $\theta_2\equiv 1$ in a neighborhood of the support of θ_1, and that $\theta_2 u$ is in $L^\infty(0,T,H^s_0)$. Then, $\theta_1\partial_t^k u$ is in $L^\infty(0,T,H^s_{-2k})$ for all integers k such that $2k<s-d/2+2$. Moreover, assume g is in $C^\infty(\mathbb{C}^2)$ and vanishes at 0 and s in I_d defined by* (2.10). *Then $\partial_t^k g(\theta_1 u,\overline{\theta_1 u})$ is in $L^\infty(0,T,H^s_{-2k})$ for all integers k such that $2k<s-d/2+2$.*

Proof. As u is a solution of (4.23), $\partial_t^k u$ satisfies for $k\ge 1$ at least formally

$$\begin{aligned} \partial_t^k u &= i^k A^k u \\ &\quad + \sum_{|\alpha_1|+\cdots+|\alpha_p|+|\beta_1|+\cdots+|\beta_q|\le 2(k-1)} g_{\alpha_1,\dots,\beta_q}(x)\partial_x^{\alpha_1}u\cdots\partial_x^{\alpha_p}u\partial_x^{\beta_1}\bar u\cdots \\ &\quad \times\partial_x^{\beta_q}\bar u f_{\alpha_1,\dots,\beta_q}(u,\bar u), \end{aligned} \tag{4.35}$$

where the $g_{\alpha_1,\dots,\beta_q}(x)$ are in $S(0,0)$ and where $f_{\alpha_1,\dots,\beta_q}$ are C^∞ functions. As $\theta_2\equiv 1$ in a neighborhood of the support of θ_1, (4.35) implies

$$\begin{aligned} \theta_1\partial_t^k u &= i^k\theta_1 A^k\theta_2 u \\ &\quad + \sum_{|\alpha_1|+\cdots+|\alpha_p|+|\beta_1|+\cdots+|\beta_q|\le 2(k-1)} g_{\alpha_1,\dots,\beta_q}(x)\partial_x^{\alpha_1}\theta_2 u\cdots \\ &\quad \times\partial_x^{\beta_q}\overline{\theta_2 u} f_{\alpha_1,\dots,\beta_q}\big(\theta_2 u,\overline{\theta_2 u}\big). \end{aligned} \tag{4.36}$$

For all $|\alpha|\le 2(k-1)$, $\partial_x^\alpha\theta_2 u$ is in $L^\infty(0,T,H^{s-2(k-1)})$. As $s-2(k-1)>d/2$, we get

(4.37)

$$\sum_{|\alpha_1|+\cdots+|\alpha_p|+|\beta_1|+\cdots+|\beta_q|\le 2(k-1)} g_{\alpha_1,\dots,\beta_q}(x)\partial_x^{\alpha_1}\theta_2 u\cdots\partial_x^{\beta_q}\overline{\theta_2 u} f_{\alpha_1,\dots,\beta_q}\big(\theta_2 u,\overline{\theta_2 u}\big)$$

is in $L^\infty(0,T,H^{s-2(k-1)})$

(see, for example, [13]). A^k has its symbol in $S(2k,0)$ by Proposition 3.2. This implies that $A^k\theta_2 u$ is in $L^\infty(0,T,H^s_{-2k})$ by Lemma 4.4. Thus, $A^k\theta_2 u$ is in $L^\infty(0,T,H^{s-2k})$, and (4.36) and (4.37) imply that $\theta_1\partial_t^k u$ is in $L^\infty(0,T,H^{s-2k})$.

We prove now that $\langle x\rangle^s\theta_1\partial_t^k u$ is in $L^\infty(0,T,H^{-2k})$.

$$g_{\alpha_1,\dots,\beta_q}(x)\partial_x^{\alpha_2}\theta_2 u\cdots\partial_x^{\beta_q}\overline{\theta_2 u}f_{\alpha_1,\dots,\beta_q}(\theta_2 u,\overline{\theta_2 u}) \text{ is in } L^\infty\big(0,T,H^{s-2(k-1)+|\alpha_1|}\big), \tag{4.38}$$

since $|\alpha_2|+\cdots+|\beta_q|\le 2(k-1)-|\alpha_1|$. Moreover, Bony's decomposition in the paraproduct and remainder of uv yields

$$\|uv\|_{H^{r_1}}\le C\|u\|_{H^{r_1}}\|v\|_{H^{r_2}}\quad \forall r_2>d/2 \text{ and } -r_2\le r_1\le r_2 \tag{4.39}$$

(see, for example, [4] for the properties of the paraproduct and of the remainder). As $\langle x\rangle^s\partial^{\alpha_1}\theta_2 u$ is in $L^\infty(0,T,H^{-|\alpha_1|})$ by Lemma 4.4, (4.38) and (4.39) with $r_1=-|\alpha_1|$ and $r_2=s-2(k-1)+|\alpha_1|$ imply that

$$\langle x\rangle^s g_{\alpha_1,\dots,\beta_q}(x)\partial_x^{\alpha_1}\theta_2 u\cdots\partial_x^{\beta_q}\overline{\theta_2 u}f_{\alpha_1,\dots,\beta_q}\big(\theta_2 u,\overline{\theta_2 u}\big)\quad \text{is in } L^\infty(0,T,H^{-|\alpha_1|}). \tag{4.40}$$

As $\langle x\rangle^s A^k\theta_2 u$ is in $L^\infty(0,T,H^{-2k})$, (4.36) and (4.40) imply that $\langle x\rangle^s\theta_1\partial_t^k u$ is in $L^\infty(0,T,H^{-2k})$.

We assume in addition that g is in $C^\infty(\mathbb{C}^2)$ and vanishes at 0, that s is in I_d, and that k is an integer satisfying $2k<s-d/2+2$. Lemma 2.2 shows that $s-2k\ge 0$. Moreover, we have

$$\begin{aligned}\partial_t^k g(\theta_1 u,\overline{\theta_1 u})\\ =\sum_{j_1+\cdots+j_p+l_1+\cdots+l_q=k}\partial_t^{j_1}\theta_1 u\cdots\partial_t^{j_p}\theta_1 u\partial_t^{l_1}\overline{\theta_1 u}\cdots\partial_t^{l_q}\overline{\theta_1 u}\frac{\partial^{p+q}g}{\partial u^p\partial\bar u^q}(\theta_1 u,\overline{\theta_1 u}).\end{aligned} \tag{4.41}$$

As $\theta_1 u$ is in $L^\infty(0,T,H^s)$, and as $s>d/2$,

$$\frac{\partial^{p+q}g}{\partial u^p\partial\bar u^q}(\theta_1 u,\overline{\theta_1 u})-\frac{\partial^{p+q}g}{\partial u^p\partial\bar u^q}(0,0) \text{ is in } L^\infty(0,T,H^s).$$

As $\partial_t^{j_2}\theta_1 u,\dots,\partial_t^{l_q}\overline{\theta_1 u}$ are in $L^\infty(0,T,H^{s-2(k-j_1)})$, and as $s-2(k-j_1)>d/2$, the product

$$\partial_t^{j_2}\theta_1 u\cdots\partial_t^{j_p}\theta_1 u\partial_t^{l_1}\overline{\theta_1 u}\cdots\partial_t^{l_q}\overline{\theta_1 u}\frac{\partial^{p+q}g}{\partial u^p\partial\bar u^q}\big(\theta_1 u,\overline{\theta_1 u}\big) \text{ is in } L^\infty\big(0,T,H^{s-2(k-j_1)}\big). \tag{4.42}$$

Moreover, $\partial_t^{j_1}\theta_1 u$ is in $L^\infty(0,T,H^{s-2j_1})$. Therefore, $\partial_t^k g(\theta_1 u,\overline{\theta_1 u})$ is in $L^\infty(0,T,H^{s-2k})$ by using (4.39) with $r_2=s-2(k-j_1)$ and $r_1=s-2j_1$ and by using (4.41). In the same way, (4.42) and the fact that $\langle x\rangle^s\partial_t^{j_1}\theta_1 u$ belongs to $L^\infty(0,T,H^{-2j_1})$ imply that $\langle x\rangle^s\partial_t^k g(\theta_1 u,\overline{\theta_1 u})$ is in $L^\infty(0,T,H^{-2k})$ by using (4.39) with $r_2=s-2(k-j_1)$ and $r_1=-2j_1$. So, $\partial_t^k g(\theta_1 u,\overline{\theta_1 u})$ is in $L^\infty(0,T,H^s_{-2k})$ for all integers k such that $2k<s-d/2+2$. ◻

Bony [2] defines the paraproduct of v by u as the following bilinear operator:

$$T_u v=\sum_{q\ge 2}S_{q-2}(u)\Delta_q(v). \tag{4.43}$$

Let $R(u,v)$ be the following bilinear operator:

$$R(u,v)=\sum_{|p-q|\leq 2}\Delta_p(u)\Delta_q(v). \tag{4.44}$$

We have $uv = T_u v + T_v u + R(u,v)$. Let $s > d/2$. Then for all u in H^s and all functions f in $C^\infty(\mathbb{R})$ such that $f(0)=0$, Bony's paralinearization theorem with the improvement of Meyer [13] is

$$f(u)=T_{f'(u)}u+r(u), \tag{4.45}$$

where $r(u)\in H^{2s-d/2}$. The generalization to several variables of (4.45) implies, for $f\in C^\infty(\mathbb{C}^2)$ such that $f(0,0)=0$,

$$f(u,\bar u)=T_{\frac{\partial f}{\partial u}(u,\bar u)}u+T_{\frac{\partial f}{\partial \bar u}(u,\bar u)}\bar u+r(u,\bar u), \tag{4.46}$$

where $r(u,\bar u)\in H^{2s-d/2}$.

The following lemma will be used in the first step of the proof to show moreover that the remainder satisfies that $\langle x\rangle^{2s-d/2}r(u,\bar u)$ is in L^2.

LEMMA 4.7. *Let $s > d/2$ and let u be in H_0^s. Let $f\in C^\infty(\mathbb{C}^2)$ such that f vanishes to the second order at* 0 *and let $r(u,\bar u)$ be defined by* (4.46). *$\langle x\rangle^{2s-d/2}f(u,\bar u)$ and $\langle x\rangle^{2s-d/2}r(u,\bar u)$ are in L^2.*

Proof. Since f vanishes to the second order at 0 and u is in L^∞, we have

$$|f(u(x),\bar u(x))|\leq C|u(x)|^2,$$

which, using the Sobolev embedding of $H^{d/4}$ in L^4, yields

$$\|\langle x\rangle^{2s-d/2}f(u,\bar u)\|_{L^2}\leq C\big\|\langle x\rangle^{s-d/4}u\big\|_{L^4}^2\leq C\big\|\langle x\rangle^{s-d/4}u\big\|_{H^{d/4}}^2. \tag{4.47}$$

Since u is in H_0^s, $\langle x\rangle^{s-d/4}u$ is in $H^{d/4}$ by Proposition 4.3. This implies that $\langle x\rangle^{2s-d/2}f(u,\bar u)$ is in L^2 using (4.47).

By (4.46), $r(u,\bar u)$ is equal to

$$f(u,\bar u)-T_{\frac{\partial f}{\partial u}(u,\bar u)}u-T_{\frac{\partial f}{\partial \bar u}(u,\bar u)}\bar u. \tag{4.48}$$

In order to prove that $\langle x\rangle^{2s-d/2}r(u,\bar u)$ is in L^2, it remains to establish

$$\langle x\rangle^{2s-d/2}\Big(T_{\frac{\partial f}{\partial u}(u,\bar u)}u+T_{\frac{\partial f}{\partial \bar u}(u,\bar u)}\bar u\Big)\in L^2. \tag{4.49}$$

It suffices to prove

$$\langle x\rangle^{2s-d/2}T_{\frac{\partial f}{\partial u}(u,\bar u)}u\in L^2, \tag{4.50}$$

since we can use the same argument for $T_{\partial_{\bar u}f(u,\bar u)}\bar u$. As $\partial_u f$ vanishes at 0 and u is in L^∞,

$$\left|\frac{\partial f}{\partial u}(u(x),\bar u(x))\right|\leq C|u(x)|.$$

This implies that $\langle x\rangle^s\partial_u f(u,\bar u)$ is in L^2. Moreover, as u is in H^s and $\partial_u f$ is C^∞ and vanishes at 0, $\partial_u f(u,\bar u)$ is in H^s. Therefore, $\langle x\rangle^{s-d/4}\partial_u f(u,\bar u)$ is in $H^{d/4}$ by

Proposition 4.3. Let $b = \partial_u f(u,\bar{u})$. Thus, $\langle x\rangle^{s-d/4} b$ is in $C^{-d/4}$ using the Sobolev embedding from $H^{d/4}$ to $C^{-d/4}$. To prove (4.50), it suffices to establish that

$$\langle x\rangle^{2s-d/2} T_b u \in L^2 \quad \text{if } \langle x\rangle^{s-d/4} b \in C^{-d/4}. \tag{4.51}$$

Using (4.2) with $k = s - d/4$, $l = 0$, $j = [s - d/4] + 2$, $a(x) = \langle x\rangle^{s-d/4}$, and $\chi = \psi$,

$$\langle x\rangle^{s-d/4}\Delta_p = \sum_{0\le|\alpha|<j} \frac{2^{-p|\alpha|} i^{|\alpha|}}{\alpha!}\big(\partial_\xi^\alpha \psi\big)(2^{-p}D)\partial_x^\alpha \langle x\rangle^{s-d/4} + \frac{\langle x\rangle^{s-d/4-j}}{2^{pj}} R, \tag{4.52}$$

where R is bounded on L^2 with a norm independent of p. Using (4.2) with $k = s - d/4$, $l_1 = d/4 + 1$, $j_1 = [s - d/4] + 3$, $a(x) = \langle x\rangle^{s-d/4}$, and $\chi = \varphi$,

$$\langle x\rangle^{s-d/4} S_p = \sum_{0\le|\beta|<j_1} \frac{2^{-p|\beta|} i^{|\beta|}}{\beta!}\big(\partial_\xi^\beta \varphi\big)(2^{-p}D)\partial_x^\beta \langle x\rangle^{s-d/4} + \frac{\langle x\rangle^{s-d/4-j_1}}{2^{p(j_1-l_1)}} R_1, \tag{4.53}$$

where R_1 is bounded from $C^{-d/4}$ to L^∞ with a norm independent of p since $C^1 \subset L^\infty$ and R_1 is bounded from $C^{-d/4}$ to C^1. Then, we get

$$\begin{aligned}
\langle x\rangle^{2s-d/2} T_b u(x) &= \sum_{p\ge2} \langle x\rangle^{s-d/4} S_{p-2} b(x) \langle x\rangle^{s-d/4}\Delta_p u(x)\\
&= \sum_{0\le|\alpha|<j}\ \sum_{0\le|\beta|<j_1} \frac{i^{|\alpha|+|\beta|}}{\alpha!\beta!} 4^{|\beta|} \sum_{p\ge2} a^p_{\alpha,\beta}(x)\\
&\quad + \sum_{0\le|\alpha|<j} \frac{i^{|\alpha|}}{\alpha!} 4^{j_1-l_1} \langle x\rangle^{s-d/4-j_1} \sum_{p\ge2} a^p_\alpha(x)\\
&\quad + \sum_{0\le|\beta|<j_1} \frac{i^{|\beta|}}{\beta!} 4^{|\beta|} \langle x\rangle^{s-d/4-j} \sum_{p\ge2} a^p_\beta(x)\\
&\quad + \sum_{p\ge2} 4^{j_1-l_1} \langle x\rangle^{2s-d/2-j-j_1} a^p(x),
\end{aligned} \tag{4.54}$$

where

$$\begin{aligned}
a^p_{\alpha,\beta} &= 2^{-p|\beta|}\big(\partial_\xi^\beta\varphi\big)\big(2^{-(p-2)}D\big)\partial_x^\beta(\langle x\rangle^{s-d/4}) b 2^{-p|\alpha|}\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)\partial_x^\alpha(\langle x\rangle^{s-d/4})u,\\
a^p_\alpha &= 2^{-p(j_1-l_1+|\alpha|)} R_1 b\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)\partial_x^\alpha(\langle x\rangle^{s-d/4})u,\\
a^p_\beta &= 2^{-p(|\beta|+j)}\big(\partial_\xi^\beta\varphi\big)\big(2^{-(p-2)}D\big)\partial_x^\beta(\langle x\rangle^{s-d/4}) bRu,\\
a^p &= 2^{-p(j_1-l_1+j)} R_1 bRu.
\end{aligned}$$

We have

$$\begin{aligned}
\sum_{p\ge2}\|a^p\|_{L^2} &\le \sum_{p\ge2} 2^{-p(j_1-l_1+j)}\|R_1 b\|_{L^\infty}\|Ru\|_{L^2}\\
&\le C\sum_{p\ge2} 2^{-2p}\|b\|_{C^{-d/4}}\|u\|_{L^2} < +\infty,
\end{aligned} \tag{4.55}$$

$$\begin{aligned}
\sum_{p\ge2}\|a^p_\alpha\|_{L^2} &\le \sum_{p\ge2} 2^{-p(j_1-l_1+|\alpha|)}\|R_1 b\|_{L^\infty}\|\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)\partial_x^\alpha(\langle x\rangle^{s-d/4})u\|_{L^2}\\
&\le C\sum_{p\ge2} 2^{-p}\|b\|_{C^{-d/4}}\|\partial_x^\alpha(\langle x\rangle^{s-d/4})u\|_{L^2} \le C\|\langle x\rangle^{s-d/4}u\|_{L^2} < +\infty,
\end{aligned} \tag{4.56}$$

and

$$
\begin{aligned}
(4.57)\qquad \sum_{p\geq 2}\|a^p_\beta\|_{L^2} &\leq \sum_{p\geq 2} 2^{-p(|\beta|+j)}\|\big(\partial^\beta_\xi\varphi\big)(2^{-(p-2)}D)\partial^\beta_x(\langle x\rangle^{s-d/4})b\|_{L^\infty}\|Ru\|_{L^2}\\
&\leq C\sum_{p\geq 2} 2^{-p(|\beta|+j-d/4)}\|\partial^\beta_x(\langle x\rangle^{s-d/4})b\|_{C^{-d/4}}\|u\|_{L^2}\\
&\leq C\sum_{p\geq 2} 2^{-p}\|\langle x\rangle^{s-d/4}b\|_{C^{-d/4}} < +\infty.
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
(4.58)\qquad \|a^p_{\alpha,\beta}\|_{L^2} &\leq 2^{-p(|\beta|+|\alpha|)}\|\big(\partial^\beta_\xi\varphi\big)(2^{-(p-2)}D)\partial^\beta_x(\langle x\rangle^{s-d/4})b\|_{L^\infty}\\
&\quad\ \|(\partial^\alpha_\xi\psi)(2^{-p}D)\partial^\alpha_x(\langle x\rangle^{s-d/4})u\|_{L^2}\\
&\leq C2^{pd/4}\|\partial^\beta_x(\langle x\rangle^{s-d/4})b\|_{C^{-d/4}}(\|\Delta_{p-1}\partial^\alpha_x(\langle x\rangle^{s-d/4})u\|_{L^2}\\
&\quad +\|\Delta_p\partial^\alpha_x(\langle x\rangle^{s-d/4})u\|_{L^2}+\|\Delta_{p+1}\partial^\alpha_x(\langle x\rangle^{s-d/4})u\|_{L^2})\\
&\leq C\|\langle x\rangle^{s-d/4}b\|_{C^{-d/4}}\|\partial^\alpha_x(\langle x\rangle^{s-d/4})u\|_{H^{d/4}}\varepsilon_p \leq C\varepsilon_p,
\end{aligned}
$$

where (ε_p) is in $l^2(\mathbb{N})$. Let $\mathcal{F}$ be the Fourier transform on $\mathbb{R}^d$. As $\mathcal{F}a^p_{\alpha,\beta}$ has support in $\{2^{p-2}\leq|\xi|\leq 9.2^{p-2}\}$, (4.10) and (4.58) imply that $\sum_{p\geq 2}a^p_{\alpha,\beta}$ is in L^2, which together with (4.54), (4.55), (4.56), and (4.57) yield (4.51). ∎

When $l(x,\xi)$ is homogeneous of degree m in ξ, C^∞ in ξ for $\xi\neq 0$, and C^ρ in x (i.e., all the derivatives $\partial^\alpha_\xi l(x,\xi)$ are C^ρ in x), Bony [2] defines the operator T_l by

$$T_l(x,D)u=\sum_{p\geq 2}S_{p-2}(\phi l)(x,D)\Delta_p u,$$

where ϕ is a C^∞ function of ξ vanishing in a neighborhood of 0 and equal to 1 outside a compact and where the Fourier multiplier S_p acting on ϕl with respect to x gives $S_p(\phi l)(x,\xi)$.

The following proposition will be used when proving Propositions 4.11 and 4.12.

PROPOSITION 4.8. *Let m, k, l, and ρ be real numbers such that $l\geq 0$ and $\rho\neq 0$. Let $b(x,\xi)$ be such that $\langle x\rangle^l b$ is homogeneous of degree m in ξ, C^∞ in ξ for $\xi\neq 0$, and C^ρ in x.*

When $\rho<0$, $\langle x\rangle^k T_b\langle x\rangle^{l-k}$ is bounded from H^t to $H^{t+\rho-m}$ for all t.

When $\rho>0$, $\langle x\rangle^k T_b\langle x\rangle^{l-k}$ is bounded from H^t to H^{t-m} for all t. Moreover, we have

$$\langle x\rangle^k T_b\langle x\rangle^{l-k}=\sum_{|\alpha|\leq\rho}\frac{1}{i^{|\alpha|}\alpha!}T_{\partial^\alpha_x(\langle x\rangle^{l-k})\langle x\rangle^k\partial^\alpha_\xi b}+R, \tag{4.59}$$

where R is bounded from H^t to $H^{t+\rho-m}$ for $t>-\rho+m$.

Proof. We start with the case $m=0$ and $b(x,\xi)=b(x)$, b in C^ρ. Using (4.2) with $k=k-l$, $l=(t+\rho)_+-t$, where $(t+\rho)_+=\max(t+\rho,0)$, $j=[\max(k-l,|\rho+t|)]+2$, $a(x)=\langle x\rangle^{k-l}$, and $\chi=\psi$,

$$\langle x\rangle^{k-l}\Delta_p=\sum_{0\leq|\alpha|<j}\frac{2^{-p|\alpha|}i^{|\alpha|}}{\alpha!}\big(\partial^\alpha_\xi\psi\big)(2^{-p}D)\partial^\alpha_x\langle x\rangle^{k-l}+\frac{\langle x\rangle^{k-l-j}}{2^{p(j-(t+\rho)_++t)}}R_1, \tag{4.60}$$

where R_1 is bounded from H^t to $H^{(t+\rho)_+}$ with a norm independent of p. Using (4.2) with $k=l$, $l=(t+\rho)_++1-\rho$, $j_1=[\max(l,|\rho+t|)]+2$, $a(x)=\langle x\rangle^l$, and $\chi=\varphi$,

$$\langle x\rangle^l S_p=\sum_{0\le|\beta|<j_1}\frac{2^{-p|\beta|}i^{|\beta|}}{\beta!}\big(\partial_\xi^\beta\varphi\big)(2^{-p}D)\partial_x^\beta\langle x\rangle^l+\frac{\langle x\rangle^{l-j_1}}{2^{p(j_1-(t+\rho)_+-1+\rho)}}R_2, \tag{4.61}$$

where R_2 is bounded from C^ρ to $C^{(t+\rho)_++1}$ with a norm independent of p. We have

$$\begin{aligned}\langle x\rangle^k T_b\langle x\rangle^{l-k}u(x)&=\sum_{p\ge2}\langle x\rangle^l S_{p-2}b(x)\langle x\rangle^{k-l}\Delta_p\langle x\rangle^{l-k}u(x)\\&=\sum_{0\le|\alpha|<j}\sum_{0\le|\beta|<j_1}\frac{i^{|\alpha|+|\beta|}}{\alpha!\beta!}4^{|\beta|}\sum_{p\ge2}a^p_{\alpha,\beta}(x)\\&\quad+\sum_{0\le|\alpha|<j}\frac{i^{|\alpha|}}{\alpha!}4^{j_1-(t+\rho)_+-1+\rho}\langle x\rangle^{l-j_1}\sum_{p\ge2}a^p_\alpha(x)\\&\quad+\sum_{0\le|\beta|<j_1}\frac{i^{|\beta|}}{\beta!}4^{|\beta|}\langle x\rangle^{k-l-j}\sum_{p\ge2}a^p_\beta(x)\\&\quad+\sum_{p\ge2}4^{j_1-(t+\rho)_+-1+\rho}\langle x\rangle^{k-j-j_1}a^p(x),\end{aligned}\tag{4.62}$$

where

$$\begin{aligned}a^p_{\alpha,\beta}&=2^{-p|\beta|}\big(\partial_\xi^\beta\varphi\big)(2^{-(p-2)}D)\partial_x^\beta(\langle x\rangle^l)b2^{-p|\alpha|}(\partial_\xi^\alpha\psi)(2^{-p}D)\partial_x^\alpha(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u,\\a^p_\alpha&=2^{-p(j_1-(t+\rho)_+-1+\rho+|\alpha|)}R_2b\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)\partial_x^\alpha(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u,\\a^p_\beta&=2^{-p(|\beta|+j-(t+\rho)_++t)}\big(\partial_\xi^\beta\varphi\big)(2^{-(p-2)}D)\partial_x^\beta(\langle x\rangle^l)bR_1u,\\a^p&=2^{-p(j_1-(t+\rho)_+-1+\rho+j-(t+\rho)_++t)}R_2bR_1u.\end{aligned}$$

The operator T defined by (4.43) is bounded from $L^\infty\times H^{t+\rho}$ to $H^{t+\rho}$ and the operator R defined by (4.44) is bounded from $C^1\times H^{(t+\rho)_+}$ to $H^{(t+\rho)_+}$ since we have $1+(t+\rho)_+>0$ (see, for example, [4]). Moreover, (4.9) implies

$$\begin{aligned}\|T_uv\|_{H^{t+\rho}}&\le\sum_{p\ge2}\|S_{p-2}u\Delta_pv\|_{H^{t+\rho}}\le C\sum_{p\ge2}2^{p(t+\rho)}\|S_{p-2}u\|_{L^2}\|\Delta_pv\|_{L^\infty}\\&\le C\sum_{p\ge2}2^{-p}\|u\|_{L^2}\|v\|_{C^{t+\rho+1}}\le C\|u\|_{L^2}\|v\|_{C^{t+\rho+1}}.\end{aligned}$$

Thus, T is bounded from $C^{t+\rho+1}\times L^2$ to $H^{t+\rho}$. As $uv=T_uv+T_vu+R(u,v)$, we get

$$\|uv\|_{H^{t+\rho}}\le C\|u\|_{H^{(t+\rho)_+}}\|v\|_{C^{(t+\rho)_++1}},$$

which implies

$$\begin{aligned}\sum_{p\ge2}\|a_p\|_{H^{t+\rho}}&\le\sum_{p\ge2}2^{-p(j_1+j-1-|t+\rho|)}\|R_2b\|_{C^{(t+\rho)_++1}}\|R_1u\|_{H^{(t+\rho)_+}}\\&\le C\sum_{p\ge2}2^{-p(1+[|t+\rho|]-|t+\rho|)}\|b\|_{C^\rho}\|u\|_{H^t}<+\infty,\end{aligned}\tag{4.63}$$

$$
\begin{aligned}
\sum_{p\geq 2}\|a^p_\alpha\|_{H^{t+\rho}} &\leq \sum_{p\geq 2} 2^{-p(j_1-(t+\rho)_+-1+\rho+|\alpha|)}\|R_2 b\|_{C^{(t+\rho)_++1}}\\
&\quad\times 2^{p((t+\rho)_+-t)}\|\big(\partial^\alpha_\xi\psi\big)(2^{-p}D)\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u\|_{H^t}\\
&\leq C\sum_{p\geq 2}2^{-p([|\rho+t|]+1-|\rho+t|)}\|b\|_{C^\rho}\|\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u\|_{H^t}\\
&\leq C\|u\|_{H^t}<+\infty,
\end{aligned}
\tag{4.64}
$$

and

$$
\begin{aligned}
\sum_{p\geq 2}\|a^p_\beta\|_{H^{t+\rho}} &\leq \sum_{p\geq 2}2^{-p(|\beta|+j-(t+\rho)_++t)}\|\big(\partial^\beta_\xi\varphi\big)(2^{-(p-2)}D)\\
&\quad\times\partial^\beta_x(\langle x\rangle^l)b\|_{C^{(t+\rho)_++1}}\|R_1u\|_{H^{(t+\rho)_+}}\\
&\leq C\sum_{p\geq 2}2^{-p(|\beta|+j-(t+\rho)_++t)}2^{p((t+\rho)_++1-\rho)}\|\partial^\beta_x(\langle x\rangle^l)b\|_{C^\rho}\|u\|_{H^t}\\
&\leq C\sum_{p\geq 2}2^{-p([|t+\rho|]+1-|t+\rho|)}\|\langle x\rangle^l b\|_{C^\rho}<+\infty.
\end{aligned}
\tag{4.65}
$$

If $\rho<0$,

$$
\begin{aligned}
2^{p(t+\rho)}\|a^p_{\alpha,\beta}\|_{L^2} &\leq 2^{p(t+\rho)}2^{-p(|\beta|+|\alpha|)}\|\big(\partial^\beta_\xi\varphi\big)(2^{-(p-2)}D)\partial^\beta_x(\langle x\rangle^l)b\|_{L^\infty}\\
&\quad\|\big(\partial^\alpha_\xi\psi\big)(2^{-p}D)\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u\|_{L^2}\\
&\leq C2^{p(t+\rho)}2^{-p\rho}\|\partial^\beta_x(\langle x\rangle^l)b\|_{C^\rho}2^{-pt}\|\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u\|_{H^t}\varepsilon_p\\
&\leq C\|\langle x\rangle^l b\|_{C^\rho}\|u\|_{H^t}\varepsilon_p\leq C\varepsilon_p,
\end{aligned}
\tag{4.66}
$$

where (ε_p) is in $l^2(\mathbb{N})$. As $\mathcal{F}a^p_{\alpha,\beta}$ has its support in $\{2^{p-2}\leq|\xi|\leq 9.2^{p-2}\}$, (4.10) and (4.66) imply that

$$
\sum_{p\geq 2}a^p_{\alpha,\beta}\quad\text{is in } H^{t+\rho}\quad\forall t\text{ and }\forall(\alpha,\beta).
\tag{4.67}
$$

Then (4.62), (4.63), (4.64), (4.65), and (4.67) show that $\langle x\rangle^kT_b\langle x\rangle^{l-k}$ is bounded from H^t to $H^{t+\rho}$ for all t if $\rho<0$.

If $\rho>0$,

$$
\begin{aligned}
2^{pt}\|a^p_{\alpha,\beta}\|_{L^2} &\leq 2^{pt}2^{-p(|\beta|+|\alpha|)}\|\big(\partial^\beta_\xi\varphi\big)(2^{-(p-2)}D)\partial^\beta_x(\langle x\rangle^l)b\|_{L^\infty}\\
&\quad\|\big(\partial^\alpha_\xi\psi\big)(2^{-p}D)\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u\|_{L^2}\\
&\leq C2^{pt}\|\partial^\beta_x(\langle x\rangle^l)b\|_{C^\rho}2^{-pt}\|\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u\|_{H^t}\varepsilon_p\\
&\leq C\|\langle x\rangle^l b\|_{C^\rho}\|u\|_{H^t}\varepsilon_p\leq C\varepsilon_p,
\end{aligned}
\tag{4.68}
$$

where (ε_p) is in $l^2(\mathbb{N})$. As $\mathcal{F}a^p_{\alpha,\beta}$ has its support in $\{2^{p-2}\leq|\xi|\leq 9.2^{p-2}\}$, (4.10) and (4.68) imply

$$
\sum_{p\geq 2}a^p_{\alpha,\beta}\quad\text{is in } H^{t}\quad\forall t\text{ and }\forall(\alpha,\beta).
\tag{4.69}
$$

Then (4.62), (4.63), (4.64), (4.65), and (4.69) show that $\langle x\rangle^kT_b\langle x\rangle^{l-k}$ is bounded from H^t to H^t for all t if $\rho>0$.

Assume now that $\rho > 0$ and $t + \rho > 0$. From

$$\sum_{p\geq 2} a^p_{0,0} = \sum_{p\geq 2} \varphi(2^{-(p-2)}D)(\langle x\rangle^l b)\psi(2^{-p}D)u = T_{\langle x\rangle^l b}u,$$

(4.62), (4.63), (4.64), and (4.65), we get (4.59) if we can prove

$$\sum_{p\geq 2} a^p_{\alpha,\beta} \in H^{t+\rho}, \quad 0 \leq |\alpha| \leq j,\ 0 \leq |\beta| \leq j_1, \text{ and } (\alpha,\beta) \neq 0. \tag{4.70}$$

When $|\beta| \geq 1$, $\partial^\beta_\xi \varphi$ has its support in $\{1/2 \leq |\xi| \leq 1\}$. Thus, for $|\beta| \geq 1$,

$$\begin{aligned}
2^{p(t+\rho)}\|a^p_{\alpha,\beta}\|_{L^2} &\leq 2^{p(t+\rho)}2^{-p(|\beta|+|\alpha|)}\|\big(\partial^\beta_\xi\varphi\big)(2^{-(p-2)}D)\partial^\beta_x(\langle x\rangle^l)b\|_{L^\infty} \\
&\quad \|\big(\partial^\alpha_\xi\psi\big)(2^{-p}D)\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u\|_{L^2} \\
&\leq C2^{p(t+\rho)}2^{-p\rho}\|\partial^\beta_x(\langle x\rangle^l)b\|_{C^\rho}2^{-pt}\|\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u\|_{H^t}\varepsilon_p \\
&\leq C\|\langle x\rangle^l b\|_{C^\rho}\|u\|_{H^t}\varepsilon_p \leq C\varepsilon_p,
\end{aligned} \tag{4.71}$$

where (ε_p) is in $l^2(\mathbb{N})$. As $\mathcal{F}a^p_{\alpha,\beta}$ has its support in $\{2^{p-2} \leq |\xi| \leq 9.2^{p-2}\}$, (4.10) and (4.71) imply that $\sum_{p\geq 2} a^p_{\alpha,\beta}$ is in $H^{t+\rho}$ for $|\beta| \geq 1$. In order to prove (4.70), it remains to show the case $\beta = 0$ and $|\alpha| \geq 1$. For $|\alpha| \geq 1$,

$$(\partial^\alpha_\xi\varphi)(\xi) + \sum_{p\geq 0} 2^{-p|\alpha|}\big(\partial^\alpha_\xi\psi\big)(2^{-p}\xi) = \partial^\alpha_\xi\left(\varphi(\xi) + \sum_{p\geq 0}\psi(2^{-p}\xi)\right) = \partial^\alpha_\xi 1 = 0.$$

Thus,

$$\big(\partial^\alpha_\xi\varphi\big)(D) + \sum_{p\geq 0} 2^{-p|\alpha|}\big(\partial^\alpha_\xi\psi\big)(2^{-p}D) = 0,$$

which yields

$$\begin{aligned}
0 &= \sum_{q\geq -1}\Delta_q(\langle x\rangle^l b)\Big(\big(\partial^\alpha_\xi\varphi\big)(D)(\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u) \\
&\quad + \sum_{p\geq 0} 2^{-p|\alpha|}\big(\partial^\alpha_\xi\psi\big)(2^{-p}D)(\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u)\Big) \\
&= \sum_{p\geq 2} S_{p-2}(\langle x\rangle^l b)2^{-p|\alpha|}\big(\partial^\alpha_\xi\psi\big)(2^{-p}D)\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u \\
&\quad + \sum_{q\geq 2}\Delta_q(\langle x\rangle^l b)2^{-(q-2)|\alpha|}\big(\partial^\alpha_\xi\varphi\big)(2^{-(q-2)}D)\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u \\
&\quad + \sum_{|p-q|\leq 2}\Delta_q(\langle x\rangle^l b)2^{-p|\alpha|}\big(\partial^\alpha_\xi\psi\big)(2^{-p}D)\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u.
\end{aligned}$$

Finally, for $|\alpha| \geq 1$,

$$\begin{aligned}
\sum_{p\geq 2} a^p_{\alpha,0} &= \sum_{p\geq 2} S_{p-2}(\langle x\rangle^l b)2^{-p|\alpha|}\big(\partial^\alpha_\xi\psi\big)(2^{-p}D)\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u \\
&= -\sum_{q\geq 2}\Delta_q(\langle x\rangle^l b)2^{-(q-2)|\alpha|} \\
&\quad \times\big(\partial^\alpha_\xi\varphi\big)(2^{-(q-2)}D)\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u \\
&\quad - \sum_{|p-q|\leq 2}\Delta_q(\langle x\rangle^l b)2^{-p|\alpha|}\big(\partial^\alpha_\xi\psi\big)(2^{-p}D)\partial^\alpha_x(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u.
\end{aligned} \tag{4.72}$$

As $\partial_\xi^\alpha \varphi$ has its support in $\{1/2 \leq |\xi| \leq 1\}$,

$$\begin{aligned}
\tag{4.73} &\|\Delta_q(\langle x\rangle^l b)2^{-(q-2)|\alpha|}\big(\partial_\xi^\alpha\varphi\big)(2^{-(q-2)}D)\partial_x^\alpha(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u\|_{L^2}\\
&\leq C2^{-q|\alpha|}\|\Delta_q(\langle x\rangle^l b)\|_{L^\infty}\\
&\|\big(\partial_\xi^\alpha\varphi\big)(2^{-(q-2)}D)\partial_x^\alpha(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u\|_{L^2}\\
&\leq C2^{-q\rho}\|\langle x\rangle^l b\|_{C^\rho}2^{-qt}\|u\|_{H^t}\varepsilon_q \leq C2^{-q(t+\rho)}\varepsilon_q,
\end{aligned}$$

where (ε_q) is in $l^2(\mathbb{N})$. Moreover, for $|q-p| \leq 2$,

$$\begin{aligned}
\tag{4.74} &\|\Delta_q(\langle x\rangle^l b)2^{-p|\alpha|}\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)\partial_x^\alpha(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u\|_{L^2}\\
&\leq C2^{-p|\alpha|}\|\Delta_q(\langle x\rangle^l b)\|_{L^\infty}\\
&\|\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)\partial_x^\alpha(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u\|_{L^2}\\
&\leq C2^{-q\rho}\|\langle x\rangle^l b\|_{C^\rho}2^{-pt}\|u\|_{H^t}\varepsilon_p \leq C2^{-p(t+\rho)}\varepsilon_p,
\end{aligned}$$

where (ε_q) is in $l^2(\mathbb{N})$.

The Fourier transform of the term from which we take the L^2 norm in (4.73) (resp., in (4.74)) has support in $\{|\xi| \leq 9.2^{q-2}\}$ (resp., in $\{|\xi| \leq 10.2^p\}$). Therefore, (4.11) yields

$$\begin{aligned}
&\sum_{q\geq 2}\Delta_q(\langle x\rangle^l b)2^{-(q-2)|\alpha|}\big(\partial_\xi^\alpha\varphi\big)(2^{-(q-2)}D)\partial_x^\alpha(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u \in H^{t+\rho},\\
&\sum_{|p-q|\leq 2}\Delta_q(\langle x\rangle^l b)2^{-p|\alpha|}\big(\partial_\xi^\alpha\psi\big)(2^{-p}D)\partial_x^\alpha(\langle x\rangle^{k-l})\langle x\rangle^{l-k}u \in H^{t+\rho}
\end{aligned}$$

as $t+\rho > 0$. This implies that $\sum_{p\geq 2} a^p_{\alpha,0}$ is in $H^{t+\rho}$ by (4.72). Finally, we have (4.70), which ends the proof of (4.59) when $m=0$ and $b(x,\xi)=b(x)$ with b in C^ρ.

We prove now the general case. Let $b(x,\xi)$ be such that $\langle x\rangle^l b$ is homogeneous of degree m in ξ, C^∞ in ξ for $\xi \neq 0$, and C^ρ in x. We decompose $b(x,\xi)$ in spherical harmonics

$$b(x,\xi) = \sum_\nu b_\nu(x)h_\nu(\xi),$$

where $\langle x\rangle^l b_\nu(x)$ belong to C^ρ with $\|\langle .\rangle^l b_\nu\|_{C^\rho} \leq 1$ and where h_ν are homogeneous of degree m, C^∞ for $\xi \neq 0$, and the sequence $\|h_\nu\|_{C^M(\mathcal{S}^{d-1})}$ is rapidly decreasing for any M. Then,

$$\begin{aligned}
\tag{4.75} \langle x\rangle^k T_b\langle x\rangle^{l-k} &= \sum_\nu \langle x\rangle^k T_{b_\nu}(\phi h_\nu)(D)\langle x\rangle^{l-k}\\
&= \sum_\nu \langle x\rangle^k T_{b_\nu}\langle x\rangle^{l-k}\langle x\rangle^{k-l}(\phi h_\nu)(D)\langle x\rangle^{l-k}.
\end{aligned}$$

If $\rho < 0$, the first part of the proof yields

$$\tag{4.76} \|\langle x\rangle^k T_{b_\nu}\langle x\rangle^{l-k}\|_{\mathcal{L}(H^t,H^{t+\rho})} \leq C \quad \forall t \quad \forall \nu$$

since $\|\langle x\rangle^l b_\nu\|_{C^\rho} \leq 1$. Moreover, Proposition 3.2 yields $\langle x\rangle^{k-l}(\phi h_\nu)(D)\langle x\rangle^{l-k} = h^1_\nu(x,D)$ with $h^1_\nu(x,D)$ in $S(m,0)$. As $\|h_\nu\|_{C^M(\mathcal{S}^{d-1})}$ is rapidly decreasing for any

M, the seminorms of $h^1_\nu(x,\xi)$ in $S(m,0)$ are rapidly decreasing. Therefore, the sequence

$$\|\langle x\rangle^{k-l}(\phi h_\nu)(D)\langle x\rangle^{l-k}\|_{\mathcal{L}(H^t,H^{t-m})} \text{ is rapidly decreasing.} \tag{4.77}$$

Formulas (4.75), (4.76), and (4.77) show that $\langle x\rangle^k T_b\langle x\rangle^{l-k}$ is bounded from H^t to $H^{t+\rho-m}$ for all t.

Assume now $\rho>0$. The first part of the proof yields

$$\|\langle x\rangle^k T_{b_\nu}\langle x\rangle^{l-k}\|_{\mathcal{L}(H^t)} \leq C \quad \forall t \quad \forall \nu. \tag{4.78}$$

So, (4.75), (4.78), and (4.77) imply that $\langle x\rangle^k T_b\langle x\rangle^{l-k}$ is bounded from H^t to H^{t-m} for all t. From (4.59) we get

$$\langle x\rangle^k T_{b_\nu}\langle x\rangle^{l-k} = T_{\langle x\rangle^l b_\nu} + R^1_\nu, \tag{4.79}$$

where R^1_ν is continuous from H^t to $H^{t+\rho}$ for all $t>-\rho$ with a norm bounded by a constant independent of ν since $\|\langle x\rangle^l b_\nu\|_{C^\rho}\leq 1$. Moreover, Proposition 3.2 and the fact that ϕ is equal to 1 outside a compact imply that

$$\begin{aligned}&\langle x\rangle^{k-l}(\phi h_\nu)(D)\langle x\rangle^{l-k}\\ &\quad=\sum_{|\alpha|\leq\rho}\frac{1}{i^{|\alpha|}\alpha!}\partial^\alpha_x(\langle x\rangle^{l-k})\langle x\rangle^{k-l}((\partial^\alpha_\xi h_\nu)\phi)(D)+R^2_\nu,\end{aligned} \tag{4.80}$$

where $R^2_\nu = r^2_\nu(x,D)$ with $r^2_\nu(x,\xi)$ in $S(m-[\rho]-1,-[\rho]-1)$. The seminorms of $r^2_\nu(x,\xi)$ in $S(m-[\rho],-[\rho])$ are rapidly decreasing. Therefore, $\|R^2_\nu\|_{\mathcal{L}(H^t,H^{t-m+[\rho]+1})}$ is rapidly decreasing.

$\langle x\rangle^l b_\nu(x)$ belongs to C^ρ, which implies that $T_{b_\nu}(\phi h_\nu)(D)$ is bounded from H^t to H^t (see [2]). As $\|\langle .\rangle^l b_\nu\|_{C^\rho}\leq 1$, the norms of $T_{b_\nu}(\phi h_\nu)(D)$ are bounded by a constant independent of ν. Moreover, the sequence of operators

$$\sum_{|\alpha|\leq\rho}\frac{1}{i^{|\alpha|}\alpha!}\partial^\alpha_x(\langle x\rangle^{l-k})\langle x\rangle^{k-l}((\partial^\alpha_\xi h_\nu)\phi)(D)$$

is bounded from H^t to H^{t-m} with rapidly decreasing norms. Together with the properties of R^1_ν and R^2_ν, this implies that

$$\begin{aligned}&\sum_\nu T_{\langle x\rangle^l b_\nu}R^2_\nu + R^1_\nu\left(\sum_{|\alpha|\leq\rho}\frac{1}{i^{|\alpha|}\alpha!}\partial^\alpha_x(\langle x\rangle^{l-k})\langle x\rangle^{k-l}((\partial^\alpha_\xi h_\nu)\phi)(D)+R^2_\nu\right)\\ &\text{is bounded from } H^t \text{ to } H^{t-m+\rho} \text{ for } t>-\rho+m.\end{aligned} \tag{4.81}$$

Therefore, (4.75), (4.79), (4.80), and (4.81) yield

$$\begin{aligned}&\langle x\rangle^k T_b\langle x\rangle^{l-k}\\ &\quad=\sum_\nu T_{\langle x\rangle^l b_\nu}\sum_{|\alpha|\leq\rho}\frac{1}{i^{|\alpha|}\alpha!}\partial^\alpha_x(\langle x\rangle^{l-k})\langle x\rangle^{k-l}((\partial^\alpha_\xi h_\nu)\phi)(D)+R_3\\ &\quad=T_{\langle x\rangle^l b}+\sum_\nu T_{\langle x\rangle^l b_\nu}\sum_{0<|\alpha|\leq\rho}\frac{1}{i^{|\alpha|}\alpha!}\partial^\alpha_x(\langle x\rangle^{l-k})\langle x\rangle^{k-l}((\partial^\alpha_\xi h_\nu)\phi)(D)+R_3,\end{aligned} \tag{4.82}$$

where R_3 is bounded from H^t to $H^{t-m+\rho}$ for $t > -\rho + m$. As $\partial_x^\alpha(\langle x\rangle^{l-k})\langle x\rangle^{k-l}$ is in S^0, $\partial_x^\alpha(\langle x\rangle^{l-k})\langle x\rangle^{k-l} - T_{\partial_x^\alpha(\langle x\rangle^{l-k})\langle x\rangle^{k-l}}$ is bounded from H^t to $H^{+\infty}$ for all t in $\mathbb{R}$ (see, for example, [10]). Together with (4.82), this yields

$$\begin{aligned}
&\sum_\nu T_{\langle x\rangle^l b_\nu} \sum_{0<|\alpha|\le\rho} \frac{1}{i^{|\alpha|}\alpha!}\partial_x^\alpha(\langle x\rangle^{l-k})\langle x\rangle^{k-l}((\partial_\xi^\alpha h_\nu)\phi)(D)\\
&= \sum_\nu \sum_{0<|\alpha|\le\rho} \frac{1}{i^{|\alpha|}\alpha!} T_{\langle x\rangle^l b_\nu} T_{\partial_x^\alpha(\langle x\rangle^{l-k})\langle x\rangle^{k-l}}((\partial_\xi^\alpha h_\nu)\phi)(D) + R_4,
\end{aligned} \tag{4.83}$$

where R_4 is continuous from H^t to $H^{+\infty}$. As $\langle x\rangle^l b$ is in C^ρ and the symbol ∂_x^α $(\langle x\rangle^{l-k})\langle x\rangle^{k-l}$ is in C^ρ, the symbolic calculus for the paradifferential operators (see, for example, [2]) implies

$$T_{\langle x\rangle^l b_\nu} T_{\partial_x^\alpha(\langle x\rangle^{l-k})\langle x\rangle^{k-l}} = T_{\partial_x^\alpha(\langle x\rangle^{l-k})\langle x\rangle^k b_\nu} + R_5, \tag{4.84}$$

where R_5 is bounded from H^t to $H^{t+\rho}$ for all t with a norm bounded by a constant independent of ν since $\|\langle.\rangle^l b_\nu\|_{C^\rho} \le 1$. Equations (4.82), (4.83), and (4.84) yield (4.59). ∎

When $\rho > 0$ is not an integer, Bony [2] defines the class Σ_ρ^m of symbols $l(x,\xi) = \sum_{j<\rho} l_{m-j}(x,\xi)$, where $l_{m-j}(x,\xi)$ is homogeneous of degree $m-j$ in ξ, C^∞ in ξ for $\xi \ne 0$, and $C^{\rho-j}$ in x. We define two other classes of symbols allowing us to construct the operator Q in the second step of the proof.

Definition 4.9. *Let there exist two real numbers m and s. We call $\Sigma_{s,s'}^m$ the class symbols $l(x,\xi)$, where $l(x,\xi)$ is homogeneous of degree m in ξ, C^∞ in ξ for $\xi \ne 0$, and $H_{s'}^s$ in x.*

Let there exist two real numbers m and $s > d/2$ such that $s-d/2$ is not an integer. We call $\widetilde{\Sigma}_s^m$ the class of symbols $l(x,\xi) = \sum_{j<s-d/2} l_{m-j}(x,\xi)$, where $l_{m-j}(x,\xi)$ is in $\Sigma_{s,-j}^{m-j}$.

Lemma 4.10.

(a) *Let m and s be two real numbers. Assume $l(x,\xi)$ is in $\Sigma_{s,s'}^m$. Then, $\partial_x^\alpha l(x,\xi)$ is in $\Sigma_{s,s'-|\alpha|}^m$ for all $|\alpha|$.*

(b) *Let m, m', s, s', and s'' be real numbers such that $s' \le 0$, $s'' \le 0$, $s + \max(s', s'') > d/2$, and $s + s' + s'' \ge 0$. Let p be in $\Sigma_{s,s'}^m$ and q in $\Sigma_{s,s''}^{m'}$. Then, pq is in $\Sigma_{s,s'+s''}^{m+m'}$.*

(c) *Let m, m', s, and s' be two real numbers. Let p be in $\Sigma_{s,s'}^m$ and q in $S(m',0)$. Then, pq is in $\Sigma_{s,s'}^{m+m'}$.*

Proof. (a) ∂_x^α has its symbol in $S(|\alpha|,0)$ and $l(x,\xi)$ is $H_{s'}^s$ in x. So, $\partial_x^\alpha l(x,\xi)$ is $H_{s'-|\alpha|}^s$ in x by Lemma 4.4. Thus, $\partial_x^\alpha l(x,\xi)$ is in $\Sigma_{s,s'-|\alpha|}^m$.

(b) We may assume that $s'' \ge s'$, which yields $s + s'' > d/2$. So, as p is $H^{s+s'}$ in x and q is $H^{s+s''}$ in x, pq is $H^{s+s'}$ in x using (4.39) with $r_1 = s + s'$ and $r_2 = s + s''$. Moreover, as $\langle x\rangle^s p$ is $H^{s'}$ in x and q is $H^{s+s''}$ in x, $\langle x\rangle^s pq$ is $H^{s'}$ in x using (4.39) with $r_1 = s'$ and $r_2 = s + s''$. So, pq is $H_{s'+s''}^s$ in x, which yields that pq is in $\Sigma_{s,s'+s''}^{m+m'}$.

(c) q is C^∞ in x with bounded derivatives. So, pq is in $H^{s+s'}$ and $\langle x\rangle^s pq$ is in $H^{s'}$. Therefore, pq is in $\Sigma_{s,s'}^{m+m'}$. ∎

The following proposition states the symbolic calculus properties needed to construct the operator Q in the second step of the proof.

Proposition 4.11. *Let m, m', and s be real numbers such that $s > d/2$ and $s - d/2$ is not an integer.*

(i) *Let p be in $\widetilde{\Sigma}^m_s$. Then, T_p is bounded from $H^t_{t'}$ to $H^t_{t'-m}$ for all t and t'. Moreover, $\langle x\rangle^{s-d/4+k}T_p\langle x\rangle^{-k}$ is bounded from H^t to $H^{t-d/4-m}$ for all t.*

(ii) *Let p be in $\widetilde{\Sigma}^m_s$ and q in $\widetilde{\Sigma}^{m'}_s$. Then*

$$T_p\circ T_q=\sum_{j+k+|\alpha|<s-d/2}\frac{1}{i^\alpha\alpha!}T_{\partial^\alpha_\xi p_{m-j}\partial^\alpha_x q_{m'-k}}+R,$$

where R is bounded from $H^t_{t'}$ to $H^{t+s-d/2}_{t'-m-m'}$ for all t and t'.

(iii) *Let p be in $\widetilde{\Sigma}^m_s$ and let $q(x,\xi)=\sum_{j<s-d/2}q_{m'-j}(x,\xi)$ with $q_{m'-j}(x,\xi)$ in $S(m'-j,0)$ homogeneous of degree $m'-j$ in ξ. Then*

$$T_p\circ q(x,D)=\sum_{j+k+|\alpha|<s-d/2}\frac{1}{i^\alpha\alpha!}T_{\partial^\alpha_\xi p_{m-j}\partial^\alpha_x q_{m'-k}}+R,$$

where R is bounded from $H^t_{t'}$ to $H^{t+s-d/2}_{t'-m-m'}$ for all t and t'.

(iv) *Let $p=\sum_{j<s-d/2}p_{m-j}(x,\xi)$ with $p_{m-j}(x,\xi)$ in $S(m-j,0)$ homogeneous of degree $m-j$ in ξ, and q is in $\widetilde{\Sigma}^{m'}_s$. Then*

$$p(x,D)\circ T_q=\sum_{j+k+|\alpha|<s-d/2}\frac{1}{i^\alpha\alpha!}T_{\partial^\alpha_\xi p_{m-j}\partial^\alpha_x q_{m'-k}}+R,$$

where R is bounded from $H^t_{t'}$ to $H^{t+s-d/2}_{t'-m-m'}$ for all t and t'.

Proof. (i) As $\widetilde{\Sigma}^m_s\subset\Sigma^m_{s-d/2}$, T_p is bounded from H^t to H^{t-m} using the symbolic calculus of paradifferential operators (see [2]). Moreover, $\langle x\rangle^kT_p\langle x\rangle^{-k}$ is bounded from H^t to H^{t-m} for all t and k by using Proposition 4.8 for each p_{m-j}. Thus, T_p is bounded from $H^t_{t'}$ to $H^t_{t'-m}$ for all t and t'. Finally, as $\langle x\rangle^{s-d/4}p_{m-j}$ is in $C^{-d/4-j}$, $\langle x\rangle^{s-d/4+k}T_p\langle x\rangle^{-k}$ is bounded from H^t to $H^{t-d/4-m}$ for all t, using Proposition 4.8 for each p_{m-j}.

(ii) As $\widetilde{\Sigma}^m_s\subset\Sigma^m_{s-d/2}$ and $\widetilde{\Sigma}^{m'}_s\subset\Sigma^{m'}_{s-d/2}$, R is $s-d/2-m-m'$-regularizing by using the symbolic calculus of paradifferential operators (see [2]).

It remains to prove that $\langle x\rangle^{t+s-d/2}Ru$ is in $H^{t'-m-m'}$ for u in $H^t_{t'}$. By Proposition 4.3, it suffices to prove that $\langle x\rangle^{t+s-d/2}R\langle x\rangle^{-t+d/4}$ is bounded from $H^{t'+d/4}$ to $H^{t'-m-m'}$. As $\sum_{j+k+|\alpha|<s-d/2}\frac{1}{i^\alpha\alpha!}\partial^\alpha_\xi p_{m-j}\partial^\alpha_x q_{m'-k}$ is in $\widetilde{\Sigma}^{m+m'}_s$ by (b) of Lemma 4.10, $\langle x\rangle^{t+s-d/2}\sum_{j+k+|\alpha|<s-d/2}\frac{1}{i^\alpha\alpha!}T_{\partial^\alpha_\xi p_{m-j}\partial^\alpha_x q_{m'-k}}\langle x\rangle^{-t+d/4}$ is bounded from $H^{t'+d/4}$ to $H^{t'-m-m'}$ by (i). It remains to show that the operator $\langle x\rangle^{t+s-d/2}T_pT_q\langle x\rangle^{-t+d/4}$ is bounded from $H^{t'+d/4}$ to $H^{t'-m-m'}$.

As

$$\langle x\rangle^{t+s-d/2}T_pT_q\langle x\rangle^{-t+d/4}=\langle x\rangle^{t+s-d/2}T_p\langle x\rangle^{-t+d/4}\langle x\rangle^{t-d/4}T_q\langle x\rangle^{-t+d/4},$$

$\langle x\rangle^{t+s-d/2}T_pT_q\langle x\rangle^{-t+d/4}$ is bounded from $H^{t'+d/4}$ to $H^{t'-m-m'}$ by using (i) with $\langle x\rangle^{t+s-d/2}T_p\langle x\rangle^{-t+d/4}$ and with $\langle x\rangle^{t-d/4}T_q\langle x\rangle^{-t+d/4}$.

(iii) As $q(x,\xi)$ is in $S^{m'}$, $q(x,D)-T_q$ is bounded from H^t to $H^{+\infty}$ (see, for example, [10]). As $\widetilde{\Sigma}^m_s\subset\Sigma^m_{s-d/2}$ and $S^{m'}\subset\Sigma^{m'}_{s-d/2}$, the symbolic calculus for the paradifferential operators (see [2]) implies that

$$T_p\circ T_q-\sum_{j+k+|\alpha|<s-d/2}\frac{1}{i^\alpha\alpha!}T_{\partial^\alpha_\xi p_{m-j}\partial^\alpha_x q_{m'-k}}$$

is $s-d/2-m-m'$-regularizing. Thus, Ru is in $H^{t+s-d/2-m-m'}$.

It remains to prove that $\langle x\rangle^{t+s-d/2}Ru$ is in $H^{t'-m-m'}$ for u in $H^t_{t'}$. By Proposition 4.3, it suffices to show that $\langle x\rangle^{t+s-d/2}R\langle x\rangle^{-t+d/4}$ is bounded from $H^{t'+d/4}$ to $H^{t'-m-m'}$. As $\sum_{j+k+|\alpha|<s-d/2}\frac{1}{i^\alpha\alpha!}\partial_\xi^\alpha p_{m-j}\partial_x^\alpha q_{m'-k}$ is in $\widetilde{\Sigma}_s^{m+m'}$ by (c) of Lemma 4.10, $\langle x\rangle^{t+s-d/2}\sum_{j+k+|\alpha|<s-d/2}\frac{1}{i^\alpha\alpha!}T_{\partial_\xi^\alpha p_{m-j}\partial_x^\alpha q_{m'-k}}\langle x\rangle^{-t+d/4}$ is bounded from $H^{t'+d/4}$ to $H^{t'-m-m'}$ by (i). It remains to show that the operator $\langle x\rangle^{t+s-d/2}T_pq(x,D)\langle x\rangle^{-t+d/4}$ is bounded from $H^{t'+d/4}$ to $H^{t'-m-m'}$.

We have

$$\langle x\rangle^{t+s-d/2}T_pq(x,D)\langle x\rangle^{-t+d/4}=\langle x\rangle^{t+s-d/2}T_p\langle x\rangle^{-t+d/4}\langle x\rangle^{t-d/4}q(x,D)\langle x\rangle^{-t+d/4},$$

$\langle x\rangle^{t-d/4}q(x,D)\langle x\rangle^{-t+d/4} = q_1(x,D)$ with $q_1(x,\xi)$ in $S(0,m')$ by Proposition 3.2. $\langle x\rangle^{t-d/4}q(x,D)\langle x\rangle^{-t+d/4}$ is bounded from $H^{t'+d/4}$ to $H^{t'+d/4-m'}$ since $q_1(x,\xi)$ is in $S(0,m')$, and $\langle x\rangle^{t+s-d/2}T_p\langle x\rangle^{-t+d/4}$ is bounded from $H^{t'+d/4-m'}$ to $H^{t'-m-m'}$ by (i). Thus, $\langle x\rangle^{t+s-d/2}T_pq(x,D)\langle x\rangle^{-t+d/4}$ is bounded from $H^{t'+d/4}$ to $H^{t'-m-m'}$.

The proof of (iv) is similar to the proof of (iii). ◻

The following proposition will be used to bound the terms coming from the nonlinearity in the energy estimates of the third step of the proof.

PROPOSITION 4.12. *Let $s>d/2$ such that $s-d/2$ is not an integer. Let $m\geq 0$ and $k\geq 0$ such that $k+m=2s-d/2$. Let $p(x,\xi)$ be in $S(m,k)$ and q in $\widetilde{\Sigma}_s^0$. Let u be in H_0^s. Then, there exist N symbols p_j, $1\leq j\leq N$, in $S(m,k)$ with support included in the support of $p(x,\xi)$ such that*

$$\|p(x,D)T_qu\|_{L^2}\leq C\left(\sum_{1\leq j\leq N}\|p_j(x,D)u\|_{L^2}+1\right). \tag{4.85}$$

Moreover, when $m=0$ or when $q\equiv 0$ in a neighborhood of the support of p, we can take $p_j(x,\xi)=0$ for all $1\leq j\leq N$.

Proof. By Proposition 3.2, $p(x,D)=p_1(x,D)\langle x\rangle^k$, where $p_1(x,\xi)$ is in $S(m,0)$. If $m=0$, $p_1(x,D)$ is bounded on L^2. It remains to show that $\langle x\rangle^{2s-d/2}T_qu$ belongs to L^2. As u is in H_0^s, $\langle x\rangle^{s-d/4}u$ is in $H^{d/4}$ by Proposition 4.3. Moreover, $\langle x\rangle^{2s-d/2}T_q\langle x\rangle^{d/4-s}$ is bounded from $H^{d/4}$ to L^2 by (i) of Proposition 4.11. This concludes the case $m=0$.

If $0<m\leq 2s-d/2$, we choose a real number ζ such that $\max(0,k-s)\leq\zeta\leq k$ and $s-\zeta-d/2$ is not an integer. Then

$$p(x,D)T_qu=p_1(x,D)\langle x\rangle^kT_q\langle x\rangle^{-k+\zeta}\langle x\rangle^{k-\zeta}u. \tag{4.86}$$

As $0\leq k-\zeta\leq s$ and u is in H_0^s, $\langle x\rangle^{k-\zeta}u$ is in $H^{s-k+\zeta}$ by Proposition 4.3. Let $0\leq j<s-d/2$. As $s-\zeta-d/2$ is not an integer, we consider the cases $s-\zeta-d/2-j<0$ and $s-\zeta-d/2-j>0$. If $s-\zeta-d/2-j<0$, $\langle x\rangle^kT_{q_{-j}}\langle x\rangle^{-k+\zeta}$ is bounded from H^t to $H^{t+s-\zeta-d/2}$ for all t by Proposition 4.8. Thus, $\langle x\rangle^kT_{q_{-j}}\langle x\rangle^{-k+\zeta}$ is bounded from $H^{s-k+\zeta}$ to H^m. As $\langle x\rangle^{k-\zeta}u$ is in $H^{s-k+\zeta}$, and $p_1(x,D)$ is bounded from H^m to L^2, we get

$$\|p(x,D)T_{q_{-j}}u\|_{L^2}\leq C,\quad s-\zeta-d/2<j. \tag{4.87}$$

If $s-\zeta-d/2-j>0$, then Proposition 4.8 implies

$$\langle x\rangle^kT_{q_j}\langle x\rangle^{-k+\zeta}=\sum_{|\alpha|\leq s-\zeta-d/2-j}\frac{1}{i^\alpha\alpha!}T_{\partial_x^\alpha(\langle x\rangle^{-k+\zeta})\langle x\rangle^k\partial_\xi^\alpha q_{-j}}+R_1^j, \tag{4.88}$$

where R_1^j is bounded from $H^{s-k+\zeta}$ to H^m. As $\langle x\rangle^{k-\zeta}u$ is in $H^{s-k+\zeta}$, and $p_1(x,D)$ is bounded from H^m to L^2, we get

$$\|p_1(x,D)\langle x\rangle^{k-\zeta}R_1^j u\|_{L^2} \leq C.$$

Moreover, as p_1 is in S^m,

$$\forall t \text{ in } \mathbb{R},\ p_1(x,D) - T_{p_1} \text{ is bounded from } H^t \text{ to } H^{+\infty} \tag{4.89}$$

(see, for example, [10]). So, it suffices to look at

$$T_{p_1}T_{\partial_x^\alpha(\langle x\rangle^{-k+\zeta})\langle x\rangle^k\partial_\xi^\alpha q_{-j}}.$$

As p_1 is $C^{s-\zeta-d/2-j}$ in x and $\partial_x^\alpha(\langle x\rangle^{-k+\zeta})\langle x\rangle^k\partial_\xi^\alpha q_{-j}$ is $C^{s-\zeta-d/2-j}$ in x, the symbolic calculus for the paradifferential operators (see, for example, [10]) yields

$$T_{p_1}T_{\partial_x^\alpha(\langle x\rangle^{-k+\zeta})\langle x\rangle^k\partial_\xi^\alpha q_{-j}} = \sum_{|\beta|<s-\zeta-d/2-j}\frac{1}{i^\beta\beta!}T_{\partial_\xi^\beta p_1\partial_x^\beta\left(\partial_x^\alpha(\langle x\rangle^{-k+\zeta})\langle x\rangle^k\partial_\xi^\alpha q_{-j}\right)} + R_2^j, \tag{4.90}$$

where R_2^j is $s-\zeta-d/2-m$-regularizing. When $q\equiv 0$ in a neighborhood of the support of p, (4.86), (4.87), (4.88), and (4.90) imply (4.85) with $p_j(x,\xi)=0$ for all $1\leq j\leq N$.

It remains to prove the general case. Let c_γ, $\gamma\in\mathbb{R}^d$, be defined by

$$c_0 = 1,\quad c_\gamma = -\sum_{0<\omega\leq\gamma}\frac{1}{i^{|\omega|}\omega!}c_{\gamma-\omega}.$$

The equalities

$$\sum_{0\leq\omega\leq\gamma}\frac{1}{i^{|\omega|}\omega!}c_{\gamma-\omega} = \delta_{\gamma 0}$$

and the symbolic calculus for the paradifferential operators imply that

$$\begin{aligned}&\sum_{|\gamma|<s-\zeta-d/2-j-|\beta|} c_\gamma T_{\partial_\xi^\gamma\partial_x^\beta\left(\partial_x^\alpha(\langle x\rangle^{-k+\zeta})\langle x\rangle^k\partial_\xi^\alpha q_{-j}\right)}T_{\partial_x^\gamma\partial_\xi^\beta p_1}\\ &\quad = T_{\partial_\xi^\beta p_1\partial_x^\beta\left(\partial_x^\alpha(\langle x\rangle^{-k+\zeta})\langle x\rangle^k\partial_\xi^\alpha q_{-j}\right)} + R_3^j,\end{aligned} \tag{4.91}$$

where R_3^j is $s-\zeta-d/2-m$-regularizing. By using the analogue of (4.89) for $\partial_x^\gamma\partial_\xi^\beta p_1$, (4.91) becomes

$$\begin{aligned}&\sum_{|\gamma|<s-\zeta-d/2-j-|\beta|} c_\gamma T_{\partial_\xi^\gamma\partial_x^\beta\left(\partial_x^\alpha(\langle x\rangle^{-k+\zeta})\langle x\rangle^k\partial_\xi^\alpha q_{-j}\right)}\partial_x^\gamma\partial_\xi^\beta p_1(x,D)\\ &\quad = T_{\partial_\xi^\beta p_1\partial_x^\beta\left(\partial_x^\alpha(\langle x\rangle^{-k+\zeta})\langle x\rangle^k\partial_\xi^\alpha q_{-j}\right)} + R_4^j,\end{aligned} \tag{4.92}$$

where R_4^j is $s-\zeta-d/2-m$-regularizing. As R_2^j and R_4^j are $s-\zeta-d/2-m$-regularizing, we have

$$\|R_2^j\langle x\rangle^{k-\zeta}u\|_{L^2} + \|R_4^j\langle x\rangle^{k-\zeta}u\|_{L^2} \leq C, \tag{4.93}$$

since $\langle x\rangle^{k-\zeta}u$ is in $H^{s-k+\zeta}$. Finally, (4.86), (4.87), (4.88), (4.90), (4.92), and (4.93) imply

$$\begin{aligned}\|p(x,D)T_q u\|_{L^2} \le C\Bigg(&\sum_{\max(j+|\alpha|,j+|\beta|+|\gamma|)<s-\zeta-d/2}\\ &\times \left\|T_{\partial_\xi^\gamma\partial_x^\beta(\partial_x^\alpha(\langle x\rangle^{-k+\zeta})\langle x\rangle^k\partial_\xi^\alpha q_{-j})}\partial_x^\gamma\partial_\xi^\beta p_1(x,D)\langle x\rangle^{k-\zeta}u\right\|_{L^2}+1\Bigg).\end{aligned} \tag{4.94}$$

As $j+|\beta|+ < s-\zeta-d/2$, $\partial_\xi^\gamma\partial_x^\beta(\partial_x^\alpha(\langle x\rangle^{-k+\zeta})\langle x\rangle^k\partial_\xi^\alpha q_{-j})$ is in $\Sigma^{-j-|\gamma|}_{s-\zeta-d/2-j-|\beta|}$ and the paradifferential operator $T_{\partial_\xi^\gamma\partial_x^\beta(\partial_x^\alpha(\langle x\rangle^{-k+\zeta})\langle x\rangle^k\partial_\xi^\alpha q_{-j})}$ is bounded on L^2. Inequality (4.94) becomes

$$\begin{aligned}&\|p(x,D)T_q u\|_{L^2}\\ &\le C\left(\sum_{\max(j+|\alpha|,j+|\beta|+|\gamma|)<s-\zeta-d/2}\|\partial_x^\gamma\partial_\xi^\beta p_1(x,D)\langle x\rangle^{k-\zeta}u\|_{L^2}+1\right).\end{aligned} \tag{4.95}$$

Proposition 3.2 implies that $\partial_x^\gamma\partial_\xi^\beta p_1(x,D)\langle x\rangle^{k-\zeta} = p_{j,\alpha,\beta,\gamma}(x,D)$ with $p_{j,\alpha,\beta,\gamma}(x,\xi)$ in $S(m-|\beta|,k-|\gamma|-\zeta)$. As $\zeta\ge 0$, $p_{j,\alpha,\beta,\gamma}(x,\xi)$ is in $S(m,k)$, which together with (4.95) implies (4.85). ∎

5. Proof of the main theorems. We extend the results of [6] to the following equation:

$$\begin{cases} i\dfrac{\partial u}{\partial t}+\dfrac{1}{2}\displaystyle\sum_{j,l=1}^d \partial_{x_j}a^{jl}(x)\partial_{x_l}u = f(u,\bar u), \quad 0<t<T, \quad x\in\mathbb{R}^d,\\ u|_{t=0}=u_0,\end{cases} \tag{5.1}$$

where the coefficients satisfy (2.1), (2.2), and (2.4) and where f is in $C^\infty(\mathbb{C}^2)$ and vanishes at $(0,0)$. Then

$$-\frac{1}{2}\sum_{j,l=1}^d \partial_{x_j}a^{jl}(x)\partial_{x_l}u = Au = a_2(x,D)u + a_1(x,D)u, \tag{5.2}$$

where

$$a_2(x,\xi) = 1/2\sum_{j,l=1}^d a^{jl}(x)\xi_j\xi_l \in S(2,0) \tag{5.3}$$

and

$$a_1(x,\xi) = -i\sum_{j,l=1}^d \partial_{x_j}a^{jl}(x)\xi_l \in S(1,-1). \tag{5.4}$$

By eventually shrinking $\mathcal{E}$ and the size of the support of θ, we may assume in the following the existence of a real function $\theta_1(x)$ in $S(0,0)$ such that θ_1 is equal to 1 on the support of θ and $\theta_1 u$ is in $L^\infty(0,T,H_0^s)$.

We now start with the first step of the proof. We define

$$\lambda_1 = \partial_u f(0,0), \quad \lambda_2 = \partial_{\bar u}f(0,0), \quad g(u,\bar u) = f(u,\bar u)-\lambda_1 u-\lambda_2\bar u, \tag{5.5}$$

g vanishes to the second order at 0, and θu is a solution of

$$i\frac{\partial \theta u}{\partial t} - A\theta u = \lambda_1\theta u + \lambda_2\overline{\theta u} + \theta g\big(\theta_1 u, \overline{\theta_1 u}\big) + [\theta, A]u, \quad 0 < t < T, \quad x \in \mathbb{R}^d. \tag{5.6}$$

As $\theta_1 u$ is in $L^\infty(0,T,H^s_0)$, and as g vanishes to the second order at 0, (4.46) and Lemma 4.7 imply that θu is a solution of

$$\begin{aligned} i\frac{\partial \theta u}{\partial t} - A\theta u = {}& \lambda_1\theta u + \lambda_2\overline{\theta u} + \theta T_{\frac{\partial g}{\partial u}\left(\theta_1 u, \overline{\theta_1 u}\right)}\theta_1 u + \theta T_{\frac{\partial g}{\partial \bar{u}}\left(\theta_1 u, \overline{\theta_1 u}\right)}\overline{\theta_1 u} \\ & + \theta r\big(\theta_1 u, \overline{\theta_1 u}\big) + [\theta, A]u, \quad 0 < t < T, \quad x \in \mathbb{R}^d, \end{aligned} \tag{5.7}$$

where $r(\theta_1 u, \overline{\theta_1 u})$ is in $L^\infty(0,T,H^{2s-d/2}_0)$. As $\theta_1 u$ is in $L^\infty(0,T,H^s_0)$ and as g vanishes to the second order at 0, we have seen when proving Lemma 4.7 that $\partial_u g(\theta_1 u, \overline{\theta_1 u})$ and $\partial_{\bar{u}} g(\theta_1 u, \overline{\theta_1 u})$ are in $L^\infty(0,T,H^s_0)$. Therefore, $\partial_u g(\theta_1 u, \overline{\theta_1 u})$ and $\partial_{\bar{u}} g(\theta_1 u, \overline{\theta_1 u})$ are in $L^\infty(0,T,\widetilde{\Sigma}^0_s)$. Items (iii) and (iv) of Proposition 4.11 yield

$$\begin{aligned} &[\theta, T_{\frac{\partial g}{\partial u}(\theta_1 u, \overline{\theta_1 u})}] = T_{l_\theta} + R, \\ &[\theta, T_{\frac{\partial g}{\partial \bar{u}}(\theta_1 u, \overline{\theta_1 u})}] = T_{\widetilde{l_\theta}} + \widetilde{R}, \end{aligned} \tag{5.8}$$

where l_θ and $\widetilde{l_\theta}$ are in $L^\infty(0,T,\widetilde{\Sigma}^0_s)$ with support included in the support of $\nabla\theta$, and where R and $\widetilde{R}$ are bounded from $L^\infty(0,T,H^t_{t'})$ to $L^\infty(0,T,H^{t+s-d/2}_{t'})$. We define $r_1(u,\bar{u}) = R\theta_1 u + \widetilde{R}\overline{\theta_1 u} + \theta r(\theta_1 u, \overline{\theta_1 u})$. Then, (5.7) becomes

$$\begin{aligned} i\frac{\partial \theta u}{\partial t} - A\theta u = {}& \lambda_1\theta u + \lambda_2\overline{\theta u} + T_{\frac{\partial g}{\partial u}(\theta_1 u, \overline{\theta_1 u})}\theta u + T_{\frac{\partial g}{\partial \bar{u}}(\theta_1 u, \overline{\theta_1 u})}\overline{\theta u} \\ & + r_1(u,\bar{u}) + T_{l_\theta}\theta_1 u + T_{\widetilde{l_\theta}}\overline{\theta_1 u} + [\theta, A]u, \quad 0 < t < T, \quad x \in \mathbb{R}^d, \end{aligned} \tag{5.9}$$

where $r_1(u,\bar{u})$ is in $L^\infty(0,T,H^{2s-d/2}_0)$.

We want to adapt the strategy of [6] to (5.9). However, we do not know how to deal with the terms evaluated at $\overline{\theta u}$. In fact, if $Jz = \bar{z}$ for any complex number z and if $b(x,\xi)$ is a symbol, we have

$$Jb(x,D) = b^J(x,D)J, \text{ where } b^J(x,\xi) = \bar{b}(x,-\xi). \tag{5.10}$$

So, microlocal information on θu at (x_0,ξ_0) implies for $\overline{\theta u}$ microlocal information at $(x_0,-\xi_0)$, but not at (x_0,ξ_0). Therefore, we now proceed with the second step of the proof. We look for an operator Q such that $v = \theta u - Q\overline{\theta u}$ satisfies an equation of type (5.9) without term evaluated at $\bar{v}$. This is the aim of the following proposition.

Proposition 5.1. *Let u be the solution of* (5.1) *and let s be in I_d defined by* (2.10). *Assume there exist real functions $\theta(x)$ and $\theta_1(x)$ in $S(0,0)$ such that θ_1 is equal to 1 on the support of θ and $\theta_1 u$ is in $L^\infty(0,T,H^s_0)$. Then, there exists an operator Q with $Q = q_1(x,D) + T_{q_2}$ such that $q_1(x,\xi)$ is in $S(-2,0)$ and $q_2(t,x,\xi)$ is in $L^\infty(0,T,\widetilde{\Sigma}^{-2}_s)$. Moreover, $v = \theta u - Q\overline{\theta u}$ is in $L^\infty(0,T,H^s_0)$ and satisfies*

$$\begin{cases} i\dfrac{\partial v}{\partial t} - Av = (l_1(x,D) + T_{l_2})v + r_2(u,\bar{u}) \\ \quad +(l^1_\theta(x,D) + T_{l^2_\theta})\theta_1 u + (\widetilde{l^1_\theta}(x,D) + T_{\widetilde{l^2_\theta}})\overline{\theta_1 u}, \quad 0 < t < T, \quad x \in \mathbb{R}^d, \\ v|_{t=0} = \theta u_0 - (q_1(x,D) + T_{q_2(0,.)})\overline{\theta u_0} = v_0, \end{cases} \tag{5.11}$$

where $r_2(u,\bar{u})$ is in $L^\infty(0,T,H_0^{2s-d/2})$, $l_1(x,\xi)$ is in $S(0,0)$, $l^1_\theta(x,\xi)$ and $\widetilde{l^1_\theta}(x,\xi)$ are in $S(1,-1)$, $l_2(t,x,\xi)$, $l^2_\theta(t,x,\xi)$, and $\widetilde{l^2_\theta}(t,x,\xi)$ are in $L^\infty(0,T,\widetilde{\Sigma}^0_s)$, and where the support of l^1_θ, $\widetilde{l^1_\theta}$, l^2_θ and of $\widetilde{l^2_\theta}$ are included in the support of $\nabla\theta$. Finally, when $\lambda_2=\partial_{\bar{u}}f(0,0)=0$, we may take $q_1(x,\xi)\equiv 0$.

The proof of Proposition 5.1 is a consequence of the following lemmas.

LEMMA 5.2. *Let u be the solution of (5.1) and let $s>d/2$ such that $s-d/2$ is not an integer. Assume there exist real functions $\theta(x)$ and $\theta_1(x)$ in $S(0,0)$ such that θ_1 is equal to 1 on the support of θ and $\theta_1 u$ is in $L^\infty(0,T,H_0^s)$. Let an operator Q with $Q=q_1(x,D)+T_{q_2}$ be such that $q_1(x,\xi)$ is in $S(-2,0)$, $q_2(t,x,\xi)$ is in $L^\infty(0,T,\widetilde{\Sigma}^{-2}_s)$, and $\partial_t q_2(t,x,\xi)$ is in $L^\infty(0,T,\widetilde{\Sigma}^0_s)$. Then, $v=\theta u-Q\overline{\theta u}$ is in $L^\infty(0,T,H_0^s)$ and satisfies*

$$(5.12)\qquad \begin{cases} i\dfrac{\partial v}{\partial t}-Av=(l_1(x,D)+T_{l_2})v+(\widetilde{l_1}(x,D)+T_{\widetilde{l_2}})\bar{v}+r_2(u,\bar{u})\\ \quad +(l^1_\theta(x,D)+T_{l^2_\theta})\theta_1 u+(\widetilde{l^1_\theta}(x,D)+T_{\widetilde{l^2_\theta}})\overline{\theta_1 u},\quad 0<t<T,\quad x\in\mathbb{R}^d,\\ v|_{t=0}=\theta u_0-(q_1(x,D)+T_{q_2(0,.)})\overline{\theta u_0}=v_0, \end{cases}$$

where $r_2(u,\bar{u})$ is in $L^\infty(0,T,H_0^{2s-d/2})$, $l_1(x,\xi)$ and $\widetilde{l_1}(x,\xi)$ are in $S(0,0)$, $l^1_\theta(x,\xi)$ and $\widetilde{l^1_\theta}(x,\xi)$ are in $S(1,-1)$, $l_2(t,x,\xi)$, $\widetilde{l^2}(t,x,\xi)$, $l^2_\theta(t,x,\xi)$, and $\widetilde{l^2_\theta}(t,x,\xi)$ are in $L^\infty(0,T,\widetilde{\Sigma}^0_s)$, and where the support of l^1_θ, $\widetilde{l^1_\theta}$, l^2_θ and of $\widetilde{l^2_\theta}$ are included in the support of $\nabla\theta$.

LEMMA 5.3. *There exists $q_1(x,\xi)$ in $S(-2,0)$ such that v satisfies (5.12) with $\widetilde{l_1}(x,\xi)=0$. Moreover, when $\lambda_2=\partial_{\bar{u}}f(0,0)=0$, we may take $q_1(x,\xi)\equiv 0$.*

LEMMA 5.4. *We make the assumptions of Lemma 5.2 and we assume furthermore that s is in I_d defined by (2.10). With the choice of $q_1(x,\xi)$ made in Lemma 5.3, there exists $q_2(t,x,\xi)$ in $L^\infty(0,T,\widetilde{\Sigma}^{-2}_s)$ with $\partial_t q_2(t,x,\xi)$ in $L^\infty(0,T,\widetilde{\Sigma}^0_s)$ such that v satisfies (5.11).*

Proof of Lemma 5.2. φ and ψ are real functions and may be chosen even in ξ. So, J commutes with S_p and Δ_p. Therefore, (5.10) yields $JT_p=T_{p^J}J$.

As θu is in $L^\infty(0,T,H_0^s)$ and as $v=\theta u-Q\overline{\theta u}$, v is in $L^\infty(0,T,H_0^s)$ by Lemma 4.4 and (i) of Proposition 4.11. As θu satisfies (5.9), v is a solution of

$$(5.13)\qquad \begin{aligned} i\frac{\partial v}{\partial t}-Av&=\lambda_1\theta u+\lambda_2\overline{\theta u}+T_{\frac{\partial g}{\partial u}(\theta_1 u,\overline{\theta_1 u})}\theta u+T_{\frac{\partial g}{\partial \bar{u}}(\theta_1 u,\overline{\theta_1 u})}\overline{\theta u}\\ &\quad +r_1(u,\bar{u})+T_{l_\theta}\theta_1 u+T_{\widetilde{l_\theta}}\overline{\theta_1 u}+[\theta,A]\theta_1 u\\ &\quad -iT_{\partial_t q_2}\overline{\theta u}-Qi\partial_t\overline{\theta u}-AQ\overline{\theta u},\quad 0<t<T,\quad x\in\mathbb{R}^d. \end{aligned}$$

Taking the conjugate of (5.9), we have

$$(5.14)\qquad \begin{aligned} i\frac{\partial\overline{\theta u}}{\partial t}-A\overline{\theta u}&=\overline{\lambda_1\theta u}+\overline{\lambda_2}\theta u+T_{\overline{\frac{\partial g}{\partial u}(\theta_1 u,\overline{\theta_1 u})}}\overline{\theta u}+T_{\overline{\frac{\partial g}{\partial \bar{u}}(\theta_1 u,\overline{\theta_1 u})}}\theta u\\ &\quad +\overline{r_1(u,\bar{u})}+T_{l^J_\theta}\overline{\theta_1 u}\\ &\quad +T_{\widetilde{l_\theta}^J}\theta_1 u+[\theta,A]\overline{\theta_1 u},\quad 0<t<T,\quad x\in\mathbb{R}^d, \end{aligned}$$

which together with (5.13) yields

$$
\begin{aligned}
i\frac{\partial v}{\partial t} - Av = {} & (\lambda_1 + T_{\frac{\partial g}{\partial u}(\theta_1 u,\overline{\theta_1 u})} + Q(\overline{\lambda_2} + T_{\overline{\frac{\partial g}{\partial \bar{u}}(\theta_1 u,\overline{\theta_1 u})}}))\theta u \\
& + \Big(\lambda_2 + T_{\frac{\partial g}{\partial \bar{u}}(\theta_1 u,\overline{\theta_1 u})} - iT_{\partial_t q_2} - QA + Q\big(\overline{\lambda_1} + T_{\overline{\frac{\partial g}{\partial u}(\theta_1 u,\overline{\theta_1 u})}}\big) \\
& \quad - AQ\Big)\overline{\theta u} + r_1(u,\bar{u}) + Q\overline{r_1(\theta_1 u,\overline{\theta_1 u})} + \big(T_{l_\theta} + QT_{\widetilde{l_\theta}^J}\big)\theta_1 u \\
& + (T_{\widetilde{l_\theta}} + QT_{l_\theta^J})\overline{\theta_1 u} + [\theta, A]\theta_1 u \\
& + Q[\theta, A]\overline{\theta_1 u}, \quad 0 < t < T, \quad x \in \mathbb{R}^d.
\end{aligned} \tag{5.15}
$$

Proposition 3.2 implies that

$$[\theta, A] = l_\theta^1(x, D) \quad \text{and} \quad q_1(x, D)[\theta, A] = \widetilde{l_\theta^1}(x, D),$$

where $l_\theta^1(x,\xi)$ is in $S(1,-1)$ and $\widetilde{l_\theta^1}(x,\xi)$ is in $S(-1,-1)$, and where $l_\theta^1(x,\xi)$ and $\widetilde{l_\theta^1}(x,\xi)$ have their support included in the support of $\nabla\theta$. Items (iii) and (iv) of Proposition 4.11 imply that

$$
\begin{aligned}
& T_{l_\theta} + QT_{\widetilde{l_\theta}^J} = T_{l_\theta^2} + R, \\
& T_{\widetilde{l_\theta}} + QT_{l_\theta^J} + T_{q_2}[\theta, A] = T_{\widetilde{l_\theta^2}} + \widetilde{R},
\end{aligned} \tag{5.16}
$$

where l_θ^2 and $\widetilde{l_\theta^2}$ are in $L^\infty(0, T, \widetilde{\Sigma}_s^0)$ with support included in the support of $\nabla\theta$ and where R and $\widetilde{R}$ are bounded from $L^\infty(0, T, H_{t'}^t)$ to $L^\infty(0, T, H_{t'}^{t+s-d/2})$. We define

$$r_2(u,\bar{u}) = R\theta_1 u + \widetilde{R}\overline{\theta_1 u} + r_1\big(\theta_1 u, \overline{\theta_1 u}\big) + Q\overline{r_1\big(\theta_1 u, \overline{\theta_1 u}\big)}.$$

Equation (5.15) becomes

$$
\begin{aligned}
i\frac{\partial v}{\partial t} - Av = {} & \Big(\lambda_1 + T_{\frac{\partial g}{\partial u}(\theta_1 u,\overline{\theta_1 u})} + Q(\overline{\lambda_2} + T_{\overline{\frac{\partial g}{\partial \bar{u}}(\theta_1 u,\overline{\theta_1 u})}})\Big)\theta u \\
& + \Big(\lambda_2 + T_{\frac{\partial g}{\partial \bar{u}}(\theta_1 u,\overline{\theta_1 u})} - iT_{\partial_t q_2} - QA + Q\big(\overline{\lambda_1} + T_{\overline{\frac{\partial g}{\partial u}(\theta_1 u,\overline{\theta_1 u})}}\big) \\
& \quad - AQ\Big)\overline{\theta u} + r_2(u,\bar{u}) + \big(l_\theta^1(x, D) + T_{l_\theta^2}\big)\theta_1 u \\
& + \Big(\widetilde{l_\theta^1}(x, D) + T_{\widetilde{l_\theta^2}}\Big)\overline{\theta_1 u}, \quad 0 < t < T, \quad x \in \mathbb{R}^d,
\end{aligned} \tag{5.17}
$$

where $r_2(u,\bar{u})$ is in $L^\infty(0, T, H_0^{2s-d/2})$.

Taking the conjugate of $v = \theta u - Q\overline{\theta u}$, we get $\bar{v} = \overline{\theta u} - Q^J\theta u$. So

$$
\begin{aligned}
& (1 - QQ^J)\theta u = v + Q\bar{v}, \\
& (1 - Q^J Q)\overline{\theta u} = \bar{v} + Q^J v.
\end{aligned} \tag{5.18}
$$

By Propositions 3.2 and 4.11, there exist two operators P and $\widetilde{P}$, with $P = p_1(x, D) + T_{p_2}$ and $\widetilde{P} = \widetilde{p_1}(x, D) + T_{\widetilde{p_2}}$, such that $p_1(x,\xi)$ is in $S(0,0)$, $p_2(t,x,\xi)$ is in $L^\infty(0, T, \widetilde{\Sigma}_s^0)$, $\widetilde{p_1}(x,\xi)$ is in $S(0,0)$, and $\widetilde{p_2}(t,x,\xi)$ is in $L^\infty(0, T, \widetilde{\Sigma}_s^0)$ with

$$
\begin{aligned}
& P(1 - QQ^J) = 1 + R_1, \\
& \widetilde{P}(1 - Q^J Q) = 1 + R_2,
\end{aligned} \tag{5.19}
$$

where R_1 and R_2 are bounded from $L^\infty(0,T,H^t_{t'})$ to $L^\infty(0,T,H^{t+s-d/2}_{t'})$. Then, (5.17) becomes

$$i\frac{\partial v}{\partial t} - Av = L_3 v + \widetilde{L_3}\bar{v} + r_3(u,\bar{u}) + \big(l^1_\theta(x,D) + T_{l^2_\theta}\big)\theta_1 u + \big(\widetilde{l^1_\theta}(x,D) + T_{\widetilde{l^2_\theta}}\big)\overline{\theta_1 u}, \quad 0 < t < T, \quad x \in \mathbb{R}^d, \tag{5.20}$$

where

(5.21)

$$\begin{aligned} L_3 &= \Big(\lambda_1 + T_{\frac{\partial g}{\partial u}(\theta_1 u,\overline{\theta_1 u})} + Q(\overline{\lambda_2} + T_{\overline{\frac{\partial g}{\partial \bar{u}}(\theta_1 u,\overline{\theta_1 u})}})\Big)P \\ &\quad + \Big(\lambda_2 + T_{\frac{\partial g}{\partial \bar{u}}(\theta_1 u,\overline{\theta_1 u})} - iT_{\partial_t q_2} + QA + Q\Big(\overline{\lambda_1} + T_{\overline{\frac{\partial g}{\partial u}(\theta_1 u,\overline{\theta_1 u})}}\Big) + AQ\Big)\widetilde{P}Q^J, \\ \widetilde{L_3} &= \Big(\lambda_2 + T_{\frac{\partial g}{\partial \bar{u}}(\theta_1 u,\overline{\theta_1 u})} - iT_{\partial_t q_2} + QA + Q(\overline{\lambda_1} + T_{\overline{\frac{\partial g}{\partial u}(\theta_1 u,\overline{\theta_1 u})}}) + AQ\Big)\widetilde{P} \\ &\quad + \Big(\lambda_1 + T_{\frac{\partial g}{\partial u}(\theta_1 u,\overline{\theta_1 u})} + Q(\overline{\lambda_2} + T_{\overline{\frac{\partial g}{\partial \bar{u}}(\theta_1 u,\overline{\theta_1 u})}})\Big)PQ, \\ r_3(u,\bar{u}) &= r_2(u,\bar{u}) - \Big(\lambda_1 + T_{\frac{\partial g}{\partial u}(\theta_1 u,\overline{\theta_1 u})} + Q(\overline{\lambda_2} + T_{\overline{\frac{\partial g}{\partial \bar{u}}(\theta_1 u,\overline{\theta_1 u})}})\Big)R_1\theta u \\ &\quad - \Big(\lambda_2 + T_{\frac{\partial g}{\partial \bar{u}}(\theta_1 u,\overline{\theta_1 u})} - iT_{\partial_t q_2} + QA + Q(\overline{\lambda_1} + T_{\overline{\frac{\partial g}{\partial u}(\theta_1 u,\overline{\theta_1 u})}}) + AQ\Big)R_2\overline{\theta u}. \end{aligned}$$

By assumption, $\partial_t q_2(t,x,\xi)$ is in $L^\infty(0,T,\widetilde{\Sigma}^0_s)$. $r_3(u,\bar{u})$ is then in $L^\infty(0,T,H^{2s-d/2}_0)$ by Lemma 4.4 and by (i) of Proposition 4.11. Moreover, by (ii), (iii), and (iv) of Proposition 4.11, we have

$$\begin{aligned} L_3 &= l_1(x,D) + T_{l_2} + R_3, \\ \widetilde{L_3} &= \widetilde{l_1}(x,D) + T_{\widetilde{l_2}} + R_4, \end{aligned} \tag{5.22}$$

where $l_1(x,\xi)$ and $\widetilde{l_1}(x,\xi)$ are in $S(0,0)$, $l_2(t,x,\xi)$ and $\widetilde{l_2}(t,x,\xi)$ are in $L^\infty(0,T,\widetilde{\Sigma}^0_s)$, and R_3 and R_4 are bounded from $L^\infty(0,T,H^t_{t'})$ to $L^\infty(0,T,H^{t+s-d/2}_{t'})$. Then, (5.20) becomes

(5.23)

$$\begin{aligned} i\frac{\partial v}{\partial t} - Av &= l_1(x,D)v + T_{l_2}v + \widetilde{l_1}(x,D)\bar{v} + T_{\widetilde{l_2}}\bar{v} + r_4(u,\bar{u}) \\ &\quad + \big(l^1_\theta(x,D) + T_{l^2_\theta}\big)\theta_1 u + \big(\widetilde{l^1_\theta}(x,D) + T_{\widetilde{l^2_\theta}}\big)\overline{\theta_1 u}, \quad 0 < t < T, \quad x \in \mathbb{R}^d, \end{aligned}$$

where $r_4(u,\bar{u}) = r_3(u,\bar{u}) + R_3\theta u + R_4\overline{\theta u}$ is in $L^\infty(0,T,H^{2s-d/2}_0)$. This ends the proof of (5.12). ∎

Proof of Lemma 5.3. To obtain $\widetilde{l_1}(x,\xi) = 0$, it suffices by (5.19), (5.21), and (5.22) to find $p_1(x,\xi)$ and $\widetilde{p^1}(x,\xi)$ in $S(0,0)$ and $q_1(x,\xi)$ in $S(-2,0)$ satisfying

$$\begin{aligned} &p_1(x,D)\big(1 - q_1(x,D)q_1^J(x,D)\big) = 1 + r_1(x,D), \\ &\widetilde{p_1}(x,D)\big(1 - q_1^J(x,D)q_1(x,D)\big) = 1 + r_2(x,D), \\ &\big(\lambda_2 + q_1(x,D)A + \overline{\lambda_1}q_1(x,D) + Aq_1(x,D)\big)\widetilde{p_1}(x,D) \\ &\quad + \big(\lambda_1 + \overline{\lambda_2}q_1(x,D)\big)p_1(x,D)q_1(x,D) = r_3(x,D), \end{aligned} \tag{5.24}$$

where $r_1(x,\xi)$, $r_2(x,\xi)$, and $r_3(x,\xi)$ belong to $S(-(s-d/2),-(s-d/2))$. We look for $p_1(x,\xi)$, $\widetilde{p^1}(x,\xi)$, and $q_1(x,\xi)$ with $p_1(x,\xi)=\sum_{j<s-d/2}p^1_{-j}(x,\xi)$, $\widetilde{p_1}(x,\xi)=\sum_{j<s-d/2}\widetilde{p^1_{-j}}(x,\xi)$, and $q_1(x,\xi)=\sum_{j<s-d/2}q^1_{-j}(x,\xi)$, where $p^1_j(x,\xi)$ and $\widetilde{p^1_j}(x,\xi)$ are in $S(-j,-j)$ and $q^1_{-j}(x,\xi)$ is in $S(-2-j,-j)$. By Proposition 3.2, it suffices to look for $p^1_j(x,\xi)$, $\widetilde{p^1_j}(x,\xi)$, and $q^1_{-j}(x,\xi)$, $0\le j<s-d/2$, satisfying

$$p^1_{-j}-\sum_{|\alpha|+|\beta|+k+l+m=j}\frac{1}{i^{|\alpha|+|\beta|}\alpha!\beta!}\partial^\alpha_\xi p^1_{-k}\partial^\alpha_x\left(\partial^\beta_\xi q^1_{-l}\partial^\beta_x q^{1J}_{-m}\right)=\delta_{j0}, \tag{5.25}$$

$$\widetilde{p^1_{-j}}-\sum_{|\alpha|+|\beta|+k+l+m=j}\frac{1}{i^{|\alpha|+|\beta|}\alpha!\beta!}\partial^\alpha_\xi \widetilde{p^1_{-k}}\partial^\alpha_x\left(\partial^\beta_\xi q^{1J}_{-l}\partial^\beta_x q^{1}_{-m}\right)=\delta_{j0}, \tag{5.26}$$

and

$$\begin{aligned}
&\lambda_2\widetilde{p^1_{-j}}-\sum_{|\alpha|+|\beta|+k+l+m=j}\frac{1}{i^{|\alpha|+|\beta|}\alpha!\beta!}\Big(\partial^\alpha_\xi q^1_{-k}\partial^\alpha_x\left(\partial^\beta_\xi a_{2-l}\partial^\beta_x\widetilde{p^1_{-m}}\right)\\
&\qquad+\partial^\alpha_\xi a_{2-k}\partial^\alpha_x\left(\partial^\beta_\xi q^1_{-l}\partial^\beta_x\widetilde{p^1_{-m}}\right)\Big)\\
&\qquad+\sum_{|\alpha|+k+l=j}\frac{1}{i^{|\alpha|}\alpha!}\left(\overline{\lambda_1}\partial^\alpha_\xi q^1_{-k}\partial^\alpha_x\widetilde{p^1_{-l}}+\lambda_1\partial^\alpha_\xi p^1_{-k}\partial^\alpha_x q^1_{-l}\right)\\
&\qquad+\sum_{|\alpha|+|\beta|+k+l+m=j}\frac{1}{i^{|\alpha|+|\beta|}\alpha!\beta!}\overline{\lambda_2}\partial^\alpha_\xi q^1_{-k}\partial^\alpha_x\left(\partial^\beta_\xi p^1_{-l}\partial^\beta_x q^1_{-m}\right)=0.
\end{aligned}\tag{5.27}$$

For $j=0$, (5.25), (5.26), and (5.27) imply

$$\begin{array}{ll}
\text{(i)} & p^1_0\left(1-q^1_0q^{1J}_0\right)=1,\\
\text{(ii)} & \widetilde{p^1_0}\left(1-q^{1J}_0q^1_0\right)=1,\\
\text{(iii)} & \left(\lambda_2+q^1_0a_2+\overline{\lambda_1}q^1_0+a_2q^1_0\right)\widetilde{p^1_0}+\left(\lambda_1+\overline{\lambda_2}q^1_0\right)p^1_0q^1_0=0.
\end{array}\tag{5.28}$$

(i) and (ii) imply $p^1_0=\widetilde{p^1_0}\neq 0$. Thus, (iii) yields

$$\overline{\lambda_2}\left(q^1_0\right)^2+2(a_2+\mathrm{Re}(\lambda_1))q^1_0+\lambda_2=0. \tag{5.29}$$

We set

$$q^1_0(x,\xi)=\frac{-\mathrm{Re}(\lambda_1)-a_2(x,\xi)+\left((\mathrm{Re}(\lambda_1)+a_2(x,\xi))^2-|\lambda_2|^2\right)^{1/2}}{\overline{\lambda_2}}. \tag{5.30}$$

Thus, $q^1_0(x,\xi)$ is in $S(-2,0)$ and satisfies (5.29). Then, $p^1_0(x,\xi)$ is in $S(0,0)$ by (i) and $\widetilde{p^1_0(x,\xi)}$ is in $S(0,0)$ by (ii).

Let $1\le j<s-d/2$. Assume we have found q^1_{-l}, p^1_{-l}, and $\widetilde{p^1_{-l}}$, $0\le l\le j-1$, satisfying (5.25), (5.26), and (5.27) for $0\le l\le j-1$ and such that q^1_{-l} belongs to $S(-2-l,-l)$, and p^1_{-l} and $\widetilde{p^1_{-l}}$ are in $S(-l,-l)$. Equations (5.25), (5.26), and (5.27)

for j yield

$$
\begin{aligned}
&\text{(i)}_{\mathrm{j}} \quad \big(1 - q_0^1 q_0^{1J}\big) p_{-j}^1 - p_0^1 q_0^1 q_{-j}^{1J} - p_0^1 q_{-j}^1 q_0^{1J} = \alpha_j^1,\\
&\text{(ii)}_{\mathrm{j}} \quad \big(1 - q_0^1 q_0^{1J}\big) \widetilde{p_{-j}^1} - \widetilde{p_0^1} q_0^1 q_{-j}^{1J} - \widetilde{p_0^1} q_{-j}^1 q_0^{1J} = \alpha_j^2,\\
&\text{(iii)}_{\mathrm{j}} \quad \big(\lambda_2 + 2q_0^1 a_2 + \overline{\lambda_1} q_0^1\big)\widetilde{p_{-j}^1} + \big(\lambda_1 + \overline{\lambda_2} q_0^1\big) q_0^1 p_{-j}^1\\
&\qquad\qquad + s\big(2a_2 + 2\mathrm{Re}(\lambda_1) + 2\overline{\lambda_2} q_0^1\big) p_0^1 q_{-j}^1 = \alpha_j^3,
\end{aligned}
\tag{5.31}
$$

where α_j^1, α_j^2, and α_j^3 are polynomials of the derivatives of q_{-l}^1, p_{-l}^1, and $\widetilde{p_{-l}^1}$, $0 \le l \le j-1$, and are in $S(-j,-j)$ since q_{-l}^1 is in $S(-2-l,-l)$, and p_{-l}^1 and $\widetilde{p_{-l}^1}$ are in $S(-l,-l)$. Since $p_0^1 = \widetilde{p_0^1}$, (i)$_{\mathrm{j}}$ and (ii)$_{\mathrm{j}}$ yield

$$\big(1 - q_0^1 q_0^{1J}\big)\big(p_{-j}^1 - \widetilde{p_{-j}^1}\big) = \alpha_j^1 - \alpha_j^2.$$

So, $p_{-j}^1 - \widetilde{p_{-j}^1} = \alpha_j^4$, where α_j^4 is a polynomial of the derivatives of q_{-l}^1, p_{-l}^1, and $\widetilde{p_{-l}^1}$, $0 \le l \le j-1$, and is in $S(-j,-j)$. Using (5.29), we get

$$
\begin{aligned}
&\big(\lambda_2 + 2q_0^1 a_2 + \overline{\lambda_1} q_0^1\big)\widetilde{p_{-j}^1} + \big(\lambda_1 + \overline{\lambda_2} q_0^1\big) q_0^1 p_{-j}^1 = \big(\lambda_2 + 2q_0^1 a_2 + \overline{\lambda_1} q_0^1\big)\big(\widetilde{p_{-j}^1} - p_{-j}^1\big)\\
&\quad = -\big(\lambda_2 + 2q_0^1 a_2 + \overline{\lambda_1} q_0^1\big)\alpha_j^4 = \alpha_j^5,
\end{aligned}
\tag{5.32}
$$

where α_j^5 is a polynomial in the derivatives of q_{-l}^1, p_{-l}^1, and $\widetilde{p_{-l}^1}$, $0 \le l \le j-1$, and is in $S(-j,-j)$. (iii)$_{\mathrm{j}}$ and (5.32) imply that

$$\big(2a_2 + 2\mathrm{Re}(\lambda_1) + 2\overline{\lambda_2} q_0^1\big) p_0^1 q_{-j}^1 = \alpha_j^3 - \alpha_j^5. \tag{5.33}$$

Equation (5.33) gives q_{-j}^1 in $S(-2-j,-j)$ as a function of the derivatives of q_{-l}^1, p_{-l}^1, and $\widetilde{p_{-l}^1}$, $0 \le l \le j-1$. Then, (i)$_{\mathrm{j}}$ gives p_{-j}^1 in $S(-j,-j)$ as a function of q_{-j}^1 and the derivatives of q_{-l}^1, p_{-l}^1, and $\widetilde{p_{-l}^1}$, $0 \le l \le j-1$, and (ii)$_{\mathrm{j}}$ gives $\widetilde{p_{-j}^1}$ in $S(-j,-j)$ as a function of q_{-j}^1 and the derivatives of q_{-l}^1, p_{-l}^1, and $\widetilde{p_{-l}^1}$, $0 \le l \le j-1$. So, we obtain $p_1(x,\xi)$ and $\widetilde{p^1}(x,\xi)$ in $S(0,0)$ and $q_1(x,\xi)$ in $S(-2,0)$ satisfying (5.24) by iteration.

Finally, when $\lambda_2 = \partial_{\bar u} f(0,0) = 0$, $q_1(x,\xi) = 0$ and $p_1(x,\xi) = \widetilde{p_1}(x,\xi) = 1$ is a solution of (5.24). So, we may take $q_1(x,\xi) \equiv 0$. ∎

Proof of Lemma 5.4. Assume that $p_1(x,\xi)$, $\widetilde{p^1}(x,\xi)$, and $q_1(x,\xi)$ are given as above. In particular, $p_1(x,\xi)$, $\widetilde{p^1}(x,\xi)$, and $q_1(x,\xi)$ satisfy (5.24). Moreover, we see from (5.30) that $q_0^1(x,\xi)$ has an asymptotic expansion of the form

$$q_0^1(x,\xi) = \sum_{l<s-d/2} q_{0,-2-l}^{1h}(x,\xi) + r(x,\xi), \quad r(x,\xi) \in S(-(s-d/2)-2, 0),$$

where $q_{0,-2-l}^{1h}(x,\xi)$ is in $S(-2-l,0)$ and is homogeneous of degree $-2-l$ in ξ outside a neighborhood of $\xi = 0$. Then, (i) and (ii) of (5.28) imply that

$$
\begin{aligned}
p_0^1(x,\xi) &= \sum_{l<s-d/2} p_{0,-l}^{1h}(x,\xi) + r(x,\xi), \quad r(x,\xi) \in S(-(s-d/2),0),\\
\widetilde{p_0^1}(x,\xi) &= \sum_{l<s-d/2} \widetilde{p_{0,-l}^{1h}}(x,\xi) + r(x,\xi), \quad r(x,\xi) \in S(-(s-d/2),0),
\end{aligned}
$$

where $p^{1h}_{0,-l}(x,\xi)$ and $\widetilde{p^{1h}_{0,-l}}(x,\xi)$ are in $S(-l,0)$ and are homogeneous of degree $-l$ in ξ outside a neighborhood of $\xi=0$.

Similarly, we show by iteration on $0\le j<s-d/2$ that

$$\begin{aligned}
p^1_{-j}(x,\xi) &= \sum_{l<s-d/2} p^{1h}_{j,-l}(x,\xi)+r(x,\xi), \quad r(x,\xi)\in S(-(s-d/2),0),\\
\widetilde{p^1_{-j}}(x,\xi) &= \sum_{l<s-d/2} \widetilde{p^{1h}_{j,-l}}(x,\xi)+r(x,\xi), \quad r(x,\xi)\in S(-(s-d/2),0),\\
q^1_{-j}(x,\xi) &= \sum_{l<s-d/2} q^{1h}_{j,-2-l}(x,\xi)+r(x,\xi), \quad r(x,\xi)\in S(-(s-d/2)-2,0),
\end{aligned}$$

where $p^{1h}_{j,-l}(x,\xi)$ and $\widetilde{p^{1h}_{j,-l}}(x,\xi)$ are in $S(-l,0)$ and are homogeneous of degree $-l$ in ξ outside a neighborhood of $\xi=0$, and $q^{1h}_{j,-2-l}(x,\xi)$ is in $S(-2-l,0)$ and is homogeneous of degree $-2-l$ in ξ outside a neighborhood of $\xi=0$. Grouping the various expansions, we obtain that $p_1(x,\xi)$, $\widetilde{p_1}(x,\xi)$, and $q_1(x,\xi)$ have asymptotic expansions of the form

$$\begin{aligned}
p_1(x,\xi) &= \sum_{j<s-d/2} p^{1h}_{-j}(x,\xi)+r_1(x,\xi), \quad r_1(x,\xi)\in S(-(s-d/2),0),\\
\widetilde{p_1}(x,\xi) &= \sum_{j<s-d/2} \widetilde{p^{1h}_{-j}}(x,\xi)+r_2(x,\xi), \quad r_2(x,\xi)\in S(-(s-d/2),0),\\
q_1(x,\xi) &= \sum_{j<s-d/2} q^{1h}_{-2-j}(x,\xi)+r_3(x,\xi), \quad r_3(x,\xi)\in S(-(s-d/2)-2,0),
\end{aligned} \tag{5.34}$$

where $p^{1h}_{-j}(x,\xi)$ and $\widetilde{p^{1h}_{-j}}(x,\xi)$ are in $S(-j,0)$ and are homogeneous of degree $-j$ in ξ outside a neighborhood of $\xi=0$ and $q^{1h}_{-2-j}(x,\xi)$ is in $S(-2-j,0)$ and is homogeneous of degree $-2-j$ in ξ outside a neighborhood of $\xi=0$. In particular, we have

$$p^{1h}_0(x,\xi)=1, \quad \widetilde{p^{1h}_0}(x,\xi)=1, \quad q^{1h}_{-2}(x,\xi)=-\frac{\lambda_2}{2a_2}. \tag{5.35}$$

We look for $p_2(t,x,\xi)$ and $\widetilde{p_2}(t,x,\xi)$ in $L^\infty(0,T,\widetilde{\Sigma}^0_s)$ and $q_2(t,x,\xi)$ in $L^\infty(0,T,\widetilde{\Sigma}^{-2}_s)$ such that Q, P, and $\widetilde{P}$ satisfy (5.19) and $\widetilde{l_2}(t,x,\xi)=0$, and such that $\partial_t q_2(t,x,\xi)$ is in $L^\infty(0,T,\widetilde{\Sigma}^0_s)$. It suffices to find q^2_{-2-j}, p^2_{-j}, and $\widetilde{p^2_{-j}}$, $0\le j<s-d/2$, such that $\partial^k_t q^2_{-2-j}(t,x,\xi)$ is in $L^\infty(0,T,\Sigma^{-2-j}_{s,-j-2k})$, and $\partial^k_t p^2_{-j}(t,x,\xi)$ and $\partial^k_t\widetilde{p^2_{-j}}(t,x,\xi)$ are in $L^\infty(0,T,\Sigma^{-j}_{s,-j-2k+4})$ for $2k<s+2-d/2-j$. Moreover, by (ii), (iii), and (iv) of Proposition 4.11, q^2_{-2-j}, p^2_{-j}, and $\widetilde{p^2_{-j}}$ must satisfy for $0\le j<s-d/2$ the following equalities:

$$p^2_{-j}=\alpha^1_j\big(\partial^\alpha_x\partial^\beta_\xi p^{1h}_{-l},\partial^\alpha_x\partial^\beta_\xi p^2_{-l},\partial^\alpha_x\partial^\beta_\xi q^{1h}_{-2-l},\partial^\alpha_x\partial^\beta_\xi q^2_{-2-l}, \quad 0\le l\le j-4\big), \tag{5.36}$$

$$\widetilde{p^2_{-j}}=\alpha^2_j\big(\partial^\alpha_x\partial^\beta_\xi \widetilde{p^{1h}_{-l}},\partial^\alpha_x\partial^\beta_\xi \widetilde{p^2_{-l}},\partial^\alpha_x\partial^\beta_\xi q^{1h}_{-2-l},\partial^\alpha_x\partial^\beta_\xi q^2_{-2-l}, \quad 0\le l\le j-4\big), \tag{5.37}$$

and

(5.38)

$$\begin{aligned}
2a_2\big(1+\widetilde{p_0^2}\big)q^2_{-2-j} &= \alpha^3_j\left(\partial^\alpha_x\partial^\beta_\xi a_{2-l}, \partial^\alpha_x\partial^\beta_\xi \widetilde{p^{1h}_{-l}}, \partial^\alpha_x\partial^\beta_\xi \widetilde{p^2_{-l}}, \partial^\alpha_x\partial^\beta_\xi q^{1h}_{-2-l}, \quad 0\le l\le j\right)\\
&\quad+\alpha^4_j\left(\partial^\alpha_x\partial^\beta_\xi a_{2-l}, \partial^\alpha_x\partial^\beta_\xi \widetilde{p^{1h}_{-l}}, \partial^\alpha_x\partial^\beta_\xi \widetilde{p^2_{-l}}, \partial^\alpha_x\partial^\beta_\xi q^{1h}_{-2-l}, \partial^\alpha_x\partial^\beta_\xi q^2_{-2-l}, \quad 0\le l\le j-1\right)\\
&\quad+\alpha^5_j\left(\partial^\alpha_x\partial^\beta_\xi \widetilde{p^{1h}_{-l}}, \partial^\alpha_x\partial^\beta_\xi \widetilde{p^2_{-l}}, \partial^\alpha_x\partial^\beta_\xi \partial_t q^2_{-2-l}, \quad 0\le l\le j-2\right)\\
&\quad+\alpha^6_j\left(\partial^\alpha_x\partial^\beta_\xi p^{1h}_{-l}, \partial^\alpha_x\partial^\beta_\xi p^2_{-l}, \partial^\alpha_x\partial^\beta_\xi \widetilde{p^{1h}_{-l}}, \partial^\alpha_x\partial^\beta_\xi \widetilde{p^2_{-l}}, \partial^\alpha_x\partial^\beta_\xi q^{1h}_{-2-l}, \partial^\alpha_x\partial^\beta_\xi q^2_{-2-l}, \ 0\le l\le j-2\right)\\
&\quad+\alpha^7_j\left(\partial^\alpha_x\partial^\beta_\xi \widetilde{p^{1h}_{-l}}, \partial^\alpha_x\partial^\beta_\xi \widetilde{p^2_{-l}}, \partial^\alpha_x\partial^\beta_\xi q^{1h}_{-2-l}, \partial^\alpha_x\partial^\beta_\xi q^2_{-2-l}, \quad 0\le l\le j-4, \quad \partial^\alpha_x\frac{\partial g}{\partial u}, \partial^\alpha_x\frac{\partial g}{\partial \bar u}\right),
\end{aligned}$$

where the α^l_j, $1\le l\le 7$, are given polynomials with complex coefficients.

Remark. In (5.36), (5.37), and (5.38), we do not write the terms containing $r_1(x,\xi)$, $r_2(x,\xi)$, and $r_3(x,\xi)$ defined in (5.34). In fact, when $r(x,\xi)$ is in $S(-(s-d/2),0)$ and $q(x,\xi)$ is in $\widetilde{\Sigma}^0_s$, $q(x,D)r(x,D)$ and $r(x,D)q(x,D)$ are continuous from $H^t_{t'}$ to $H^{t+s-d/2}_{t'}$ by Lemma 4.4 and (i) of Proposition 4.11. So, the terms containing $r_1(x,\xi)$, $r_2(x,\xi)$, and $r_3(x,\xi)$ can be incorporated in the remainder $r_2(u,\bar u)$ of (5.12).

For $j=0$, (5.36), (5.37), and (5.38) imply

$$p^2_0=0,\quad \widetilde{p^2_0}=0\quad\text{ and }\quad 2a_2q^2_{-2}=\frac{\partial g}{\partial\bar u}(\theta_1u,\overline{\theta_1u}). \tag{5.39}$$

$\partial^k_t p^2_0(t,x,\xi)$ and $\partial^k_t\widetilde{p^2_0}(t,x,\xi)$ are in $L^\infty(0,T,\Sigma^0_{s,-2k+4})$ for $2k<s+2-d/2$. By restricting the support of θ_1, we may assume there exists $\theta_2(x)$ in $S(0,0)$ equal to 1 on the support of θ_1 such that θ_2u is in $L^\infty(0,T,H^s_0)$. Therefore, as s is in I_d and $\partial_{\bar u}g(0,0)=0$, Corollary 4.6 implies that $\partial^k_t\partial_{\bar u}g(\theta_1u,\overline{\theta_1u})$ is in $L^\infty(0,T,H^s_{-2k})$ for $2k<s-d/2+2$. As $a_2(x,\xi)$ is in $S(2,0)$, homogeneous of degree 2 in ξ, and elliptic, $\partial^k_t q^2_{-2}(t,x,\xi)$ is in $L^\infty(0,T,\Sigma^{-2}_{s,-2k})$ for $2k<s-d/2+2$ by (5.39), which ends the construction of $p^2_0(t,x,\xi)$, $\widetilde{p^2_0}(t,x,\xi)$, and q^2_{-2}.

Let $1\le j<s-d/2$. Assume q^2_{-2-l}, p^2_{-l}, and $\widetilde{p^2_{-l}}$ are given for $0\le l\le j-1$ such that they satisfy (5.36), (5.37), and (5.38) for $0\le l\le j-1$, and such that $\partial^k_t q^2_{-2-l}(t,x,\xi)$ is in $L^\infty(0,T,\Sigma^{-2-l}_{s,-l-2k})$, and $\partial^k_t p^2_{-l}(t,x,\xi)$ and $\partial^k_t p^2_{-l}(t,x,\xi)$ are in $L^\infty(0,T,\Sigma^{-l}_{s,-l-2k+4})$ for $2k<s+2-d/2-l$. Equations (5.36) and (5.37) for j give p^2_{-j} and p^2_{-j} and (5.38) gives q^2_{-2-j}. It remains to show that $\partial^k_t q^2_{-2-j}(t,x,\xi)$ is in $L^\infty(0,T,\Sigma^{-2-j}_{s,-j-2k})$, and that $\partial^k_t p^2_{-j}(t,x,\xi)$ and $\partial^k_t p^2_{-j}(t,x,\xi)$ are in $L^\infty(0,T,\Sigma^{-j}_{s,-j-2k+4})$ for $2k<s+2-d/2-j$. By (5.36), (5.37), and (5.38), it suffices to show that $\partial^k_t\alpha^1_j(t,x,\xi)$ and $\partial^k_t\alpha^2_j(t,x,\xi)$ are in $L^\infty(0,T,\Sigma^{-j}_{s,-j-2k+4})$ and that $\partial^k_t\alpha^l_j(t,x,\xi)$, $3\le l\le 7$, is in $L^\infty(0,T,\Sigma^{-j}_{s,-j-2k})$ for $2k<s+2-d/2-j$. Furthermore, $\alpha^l_j(t,x,\xi)$, $1\le l\le 7$, is the sum of the following terms:

1. $\partial^\alpha_\xi p_{-k}\partial^\alpha_x(\partial^\beta_\xi q_{-2-l}\partial^\beta_x(q_{-2-m}))$ and $\partial^\alpha_\xi q_{-2-k}\partial^\alpha_x(\overline{\frac{\partial g}{\partial\bar u}}(\theta_1u,\overline{\theta_1u})\partial^\beta_\xi(p_{-l})\partial^\beta_x q_{-2-m})$, where $|\alpha|+|\beta|+k+l+m=j-4$,
2. $\partial^\alpha_\xi\partial_t q^2_{-2-k}\partial^\alpha_x\widetilde{p_{-l}}$, $\partial^\alpha_\xi q_{-2-k}\partial^\alpha_x p_{-l}$, and $\partial^\alpha_\xi q_{-2-k}\partial^\alpha_x(\overline{\frac{\partial g}{\partial u}}(\theta_1u,\overline{\theta_1u})p_{-m})$, where $|\alpha|+k+l=j-2$,
3. $\partial^\alpha_\xi q_{-2-k}\partial^\alpha_x(\partial^\beta_\xi a_{2-l}\partial^\beta_x p_{-m})$, where $|\alpha|+|\beta|+k+l+m=j$,
4. p^2_{-j} and $\frac{\partial g}{\partial\bar u}\widetilde{p_{-j}}$,

where p_{-l} is either p^{1h}_{-l}, p^2_{-l}, $\widetilde{p^{1h}_{-l}}$, or $\widetilde{p^2_{-l}}$, and q_{-2-l} is either q^{1h}_{-2-l} or q^2_{-2-l}. It suffices to consider the products of two terms, and using Leibniz rule we have to prove that

$$\begin{aligned}&S(m',0)\Sigma^{m''}_{s,-l''-2k''} \subset \Sigma^{-j}_{s,-l''-2k''},\\ &\qquad \text{where } m'+m''=-j,\ k''\leq k,\ l''\leq j,\\ &\text{and } \Sigma^{m'}_{s,-l'-2k'}\Sigma^{m''}_{s,-l''-2k''} \subset \Sigma^{-j}_{s,-l'-l''-2k},\\ &\qquad \text{where } m'+m''=-j,\ k'+k''=k,\ l'+l''=j-2,\ \text{ or } j-4.\end{aligned} \tag{5.40}$$

By (c) of Lemma 4.10,

$$S(m',0)\Sigma^{m''}_{s,-l''-2k''} \subset \Sigma^{m'+m''}_{s,-l''-2k''} = \Sigma^{-j}_{s,-l''-2k''}. \tag{5.41}$$

We set $s'=-l'-2k'$ and $s''=-l''-2k''$. Then, since $l'+l''\leq j-2$ we have

$$s+s'+s''=s-l'-l''-2k\geq s-j-2k+2>0$$

and

$$s+\max(s',s'')=s+\max(-l'-2k',-l''-2k'')\geq s-l'-l''-2k\geq s-j-2k+2>d/2.$$

Item (b) of Lemma 4.10 implies that

$$\Sigma^{m'}_{s,-l'-2k'}\Sigma^{m''}_{s,-l''-2k''} \subset \Sigma^{m'+m''}_{s,-l'-l''-2k'-2k''} = \Sigma^{-j}_{s,-l'-l''-2k}. \tag{5.42}$$

Equations (5.41) and (5.42) yield (5.40). So, $\partial_t^k q^2_{-2-j}(t,x,\xi)$ is in $L^\infty(0,T,\Sigma^{-2-j}_{s,-j-2k})$, and $\partial_t^k p^2_{-j}(t,x,\xi)$ and $\partial_t^k p^2_{-j}(t,x,\xi)$ are in $L^\infty(0,T,\Sigma^{-j}_{s,-j-2k+4})$ for $2k<s+2-d/2-j$.

Finally, we obtain $q_2(t,x,\xi)$ in $L^\infty(0,T,\widetilde{\Sigma}^{-2}_s)$ such that $\partial_t q_2(t,x,\xi)$ is in $L^\infty(0,T,\widetilde{\Sigma}^0_s)$ and $\widetilde{l_2}(t,x,\xi)=0$ by iteration. Moreover, $\widetilde{l_1}(x,\xi)=0$ by Lemma 5.3. Since v satisfies (5.12) with $\widetilde{l_1}(x,\xi)=\widetilde{l_2}(t,x,\xi)=0$, v satisfies (5.11). □

We finally proceed with the third and last step of the proof. Adapting the strategy of [6] to (5.11), we prove a microlocal smoothing effect result for v which in turn yields a result of microlocal smoothing effect for u.

Let $b(x,\xi)$ be in $S(m,k)$. Proposition 3.2 yields

$$\begin{aligned} i[A,b(x,D)] &= i[a(x,D),b(x,D)]+i[a_1(x,D),b(x,D)]\\ &= \{a,b\}(x,D)+e(x,D)+R_L,\end{aligned} \tag{5.43}$$

where $L=[\max(m+2,k)]+1$. R_L is a bounded operator on L^2 and $e(x,\xi)$ in $S(m,k-2)$ is defined by

$$\begin{aligned} e(x,\xi) &= \sum_{2\leq|\alpha|<L}\frac{1}{i^{|\alpha|-1}\alpha!}\big(\partial_\xi^\alpha a\partial_x^\alpha b-\partial_x^\alpha a\partial_\xi^\alpha b\big)\\ &\quad+\sum_{1\leq|\alpha|<L}\frac{1}{i^{|\alpha|-1}\alpha!}\big(\partial_\xi^\alpha a_1\partial_x^\alpha b-\partial_x^\alpha a_1\partial_\xi^\alpha b\big).\end{aligned} \tag{5.44}$$

Let v a solution of (5.11) and B an operator. Then,

(5.45)

$$\begin{aligned}
&\partial_t\langle Bv,v\rangle = \langle B\partial_t v,v\rangle + \langle Bv,\partial_t v\rangle + \left\langle \frac{\partial B}{\partial t}v,v\right\rangle \\
&= \langle -iB\big(Av + l_1(x,D)v + T_{l_2}v + r_2(u,\bar u) + \big(l^1_\theta(x,D) + T_{l^2_\theta}\big)\theta_1 u \\
&\quad + \big(\widetilde{l^1_\theta}(x,D) + T_{\widetilde{l^2_\theta}}\big)\overline{\theta_1 u}\big),v\rangle + \langle Bv, -i\big(Av + l_1(x,D)v + T_{l_2}v \\
&\quad + r_2(u,\bar u) + \big(l^1_\theta(x,D) + T_{l^2_\theta}\big)\theta_1 u + \big(\widetilde{l^1_\theta}(x,D) + T_{\widetilde{l^2_\theta}}\big)\overline{\theta_1 u}\big)\rangle + \left\langle \frac{\partial B}{\partial t}v,v\right\rangle \\
&= -i\langle [B,A]v,v\rangle - i\langle B\big(l_1(x,D)v + T_{l_2}v + r_2(u,\bar u) \\
&\quad + \big(l^1_\theta(x,D) + T_{l^2_\theta}\big)\theta_1 u + \big(\widetilde{l^1_\theta}(x,D) + T_{\widetilde{l^2_\theta}}\big)\overline{\theta_1 u}\big),v\rangle + i\langle Bv, l_1(x,D)v + T_{l_2}v \\
&\quad + r_2(u,\bar u) + \big(l^1_\theta(x,D) + T_{l^2_\theta}\big)\theta_1 u + \big(\widetilde{l^1_\theta}(x,D) + T_{\widetilde{l^2_\theta}}\big)\overline{\theta_1 u}\rangle + \left\langle \frac{\partial B}{\partial t}v,v\right\rangle.
\end{aligned}$$

Integrating (5.45) for $B = b(x,D)$ between 0 and t and taking the real part,

$$\begin{aligned}
(5.46)\quad &\mathrm{Re}\langle b(x,D)v(t),v(t)\rangle + \int_0^t \mathrm{Re}\langle c(x,D)v(\tau),v(\tau)\rangle d\tau \\
&\le \mathrm{Re}\langle b(x,D)v_0,v_0\rangle + \int_0^t \mathrm{Re}\langle e(x,D)v(\tau),v(\tau)\rangle d\tau \\
&\quad + \int_0^t (|\langle b(x,D)(l_1(x,D)v + T_{l_2}v + r_2(u,\bar u) + \big(l^1_\theta(x,D) + T_{l^2_\theta}\big)\theta_1 u \\
&\quad + \big(\widetilde{l^1_\theta}(x,D) + T_{\widetilde{l^2_\theta}}\big)\overline{\theta_1 u}),v\rangle| + |\langle b(x,D)v, l_1(x,D)v + T_{l_2}v + r_2(u,\bar u) \\
&\quad + (l^1_\theta(x,D) + T_{l^2_\theta})\theta_1 u + (\widetilde{l^1_\theta}(x,D) + T_{\widetilde{l^2_\theta}})\overline{\theta_1 u}\rangle|)d\tau + C\int_0^t \|v\|^2_{L^2}d\tau,
\end{aligned}$$

where $c = -\{a_2,b\}$. Let $\mu > 0$. Integrating (4.29) for $B = t^\mu b(x,D)$ between 0 and t and taking the real part,

(5.47)

$$\begin{aligned}
&t^\mu \mathrm{Re}\langle b(x,D)v(t),v(t)\rangle + \int_0^t \tau^\mu \mathrm{Re}\langle c(x,D)v(\tau),v(\tau)\rangle d\tau \\
&\le \int_0^t \tau^\mu \mathrm{Re}\langle e(x,D)v(\tau),v(\tau)\rangle d\tau + \int_0^t \mu\tau^{\mu-1}\langle b(x,D)v(\tau),v(\tau)\rangle d\tau \\
&\quad + \int_0^t \tau^\mu(|\langle b(x,D)\big(l_1(x,D)v + T_{l_2}v + r_2(u,\bar u) + \big(l^1_\theta(x,D) + T_{l^2_\theta}\big)\theta_1 u \\
&\quad + \big(\widetilde{l^1_\theta}(x,D) + T_{\widetilde{l^2_\theta}}\big)\overline{\theta_1 u}\big),v\rangle| + |\langle b(x,D)v, l_1(x,D)v + T_{l_2}v + r_2(u,\bar u) \\
&\quad + \big(l^1_\theta(x,D) + T_{l^2_\theta}\big)\theta_1 u + \big(\widetilde{l^1_\theta}(x,D) + T_{\widetilde{l^2_\theta}}\big)\overline{\theta_1 u}\rangle|)d\tau + C\int_0^t \|v\|^2_{L^2}d\tau.
\end{aligned}$$

Theorem 5.5 shows that if v_0 decreases along the backward bicharacteristic through (x_0,ξ_0), then $v(t,.)$ decreases at the same speed for all $t > 0$ along this curve. Theorem 5.6 uses the decrease established in Theorem 5.5 to prove a result of regularity for v at (x_0,ξ_0) by iteration.

Theorem 5.5. *Let (x_0,ξ_0) not be trapped backwards. Let there exist two conic neighborhoods $\mathcal{E}^0 \subset \mathcal{E}$ of $\{\varphi(s,x_0,\xi_0), s \le 0\}$. Let $T > 0$, s in I_d, and u in $C(0,T,L^2(\mathbb{R}^d))$ be a solution of* (5.1). *Suppose there is $\theta(x)$ in $S(0,0)$ equal to 1 on $\mathcal{E}$ such that θu is in $L^\infty(0,T,H_0^s)$. Let v in $L^\infty(0,T,H_0^s)$ be a solution of* (5.11). *Moreover, suppose there exists $s_0(x,\xi)$ in $S(0,2s-d/2)$ with support in $\mathcal{E}$ such that $(s_0^2,-\{a_2,s_0^2\})$ satisfies* (3.1) *and* (3.2) *and $\langle x\rangle^{2(2s-d/2)} \le s_0^2(x,\xi)$ on the set $\mathcal{E}^0 \cap \{|\xi| \ge 1\}$. Suppose v_0 satisfies*

$$\langle s_0(x,D)v_0, s_0(x,D)v_0\rangle < +\infty. \tag{5.48}$$

Then, there exists a neighborhood $\mathcal{D}^0 \subset \mathcal{E}^0$ containing the backward bicharacteristic $\{\varphi(s,x_0,\xi_0), s \le 0\}$ and a pair of symbols (b^0,c^0),

$$\begin{aligned} &0 \le b^0(x,\xi) \in S(0,2(2s-d/2)),\\ &0 \le c^0(x,\xi) = -\{a_2,b^0\}(x,\xi) \in S(1,2(2s-d/2)-1), \end{aligned} \tag{5.49}$$

such that

$$\begin{aligned} &\big\{(x,\xi)\,/\,c^0(x,\xi) \ge \langle x\rangle^{2(2s-d/2)-1}\langle\xi\rangle\big\}\\ &\qquad\qquad \cap \big\{(x,\xi)\,/\,b^0(x,\xi) \ge \langle x\rangle^{2(2s-d/2)}\big\} = \mathcal{D}^0, \end{aligned} \tag{5.50}$$

and for all $T > 0$,

$$\sup_{0\le t\le T} \|\sqrt{b^0}(x,D)v(t)\|_{L^2}^2 + \int_0^T \|\sqrt{c^0}(x,D)v(t)\|_{L^2}^2 dt < +\infty. \tag{5.51}$$

Theorem 5.6. *Let (x_0,ξ_0) not be trapped backwards. Let there exist two conic neighborhoods $\mathcal{E}^0 \subset \mathcal{E}$ of $\{\varphi(s,x_0,\xi_0), s \le 0\}$. Let $T > 0$, s in I_d, and u in $C(0,T,L^2(\mathbb{R}^d))$ be a solution of* (5.1). *Suppose there is $\theta(x)$ in $S(0,0)$ equal to 1 on $\mathcal{E}$ such that θu is in $L^\infty(0,T,H_0^s)$. Let v in $L^\infty(0,T,H_0^s)$ be a solution of* (5.11). *Moreover, suppose there exists $s_0(x,\xi)$ in $S(0,2s-d/2)$ with support in $\mathcal{E}$ such that $(s_0^2,-\{a_2,s_0^2\})$ satisfies* (3.1) *and* (3.2) *and $\langle x\rangle^{2(2s-d/2)} \le s_0^2(x,\xi)$ on the set $\mathcal{E}^0 \cap \{|\xi| \ge 1\}$. Suppose v_0 satisfies* (5.48). *We define $\kappa = 2(2s-d/2) - [2(2s-d/2)]$ and there are nested neighborhoods $\mathcal{D}^l \subset \mathcal{D}^{l-1} \subset \cdots \subset \mathcal{D}^0 \subset \mathcal{E}^0$ containing the backward bicharacteristic $\{\varphi(s,x_0,\xi_0), s \le 0\}$ and symbol pairs (b_l,c_l),*

$$\begin{aligned} &0 \le b_l(x,\xi) \in S(l+\kappa, 2(2s-d/2)-l-\kappa),\\ &0 \le c_l(x,\xi) = -\{a_2,b_l\}(x,\xi) \in S(l+\kappa+1, 2(2s-d/2)-l-\kappa-1), \end{aligned} \tag{5.52}$$

with $0 \le l \le [2(2s-d/2)]$ such that for $l > 0$,

$$\begin{aligned} \operatorname{supp}(b_l(x,\xi)) \subset &\big\{(x,\xi)/c_{l-1}(x,\xi) \ge \langle x\rangle^{2(2s-d/2)-l-\kappa}\langle\xi\rangle^{l+\kappa}\big\}\\ &\cap \big\{(x,\xi)/b_{l-1}(x,\xi) \ge \langle x\rangle^{2(2s-d/2)-l-\kappa+1}\langle\xi\rangle^{l+\kappa-1}\big\} = \mathcal{D}^{l-1}, \end{aligned} \tag{5.53}$$

and for all $T > 0$,

$$\sup_{0\le t\le T} t^{\mu_l}\|\sqrt{b_l}(x,D)v(t)\|_{L^2}^2 + \int_0^T t^{\mu_l}\|\sqrt{c_l}(x,D)v(t)\|_{L^2}^2 dt < +\infty. \tag{5.54}$$

The exponents can be chosen $\mu_l = l + \kappa(1+\delta)$ for any $\delta > 0$. When $l = [2(2s-d/2)]$ there is an exception to (5.53) *that for any choice of $\nu < -1$, we may take $c_{[2(2s-d/2)]}(x,\xi)$ such that*

$$\mathcal{D}^{[2(2s-d/2)]} = \big\{(x,\xi)/c_{[2(2s-d/2)]}(x,\xi) \ge \langle x\rangle^{\nu}\langle\xi\rangle^{2(2s-d/2)}\big\} \ne \emptyset. \tag{5.55}$$

Remarks.

1. Suppose u_0 is in $H_0^s(\mathbb{R}^d)$. Then, the solution u of (5.1) is in $L^\infty(0,T,H_0^s)$ by Corollary 4.5. In this case, we can choose $\theta(x)=1$ in Theorem 5.6.
2. Let $b(x,\xi)$ be one of the symbols in the statement of Theorem 5.5 or 5.6. We will construct those symbols using Proposition 3.1. Therefore, we can choose b and c such that $\sqrt{b}$ is in $S(m/2,k/2)$ and $\sqrt{c}$ is in $S((m+1)/2,(k-1)/2)$ for $b\in S(m,k)$ and $c\in S(m+1,k-1)$.

Proof of Theorem 5.5. Let L be an integer defined by $L=[2s-d/2]+1$ and $0<\eta<1/2$ defined by $\eta=(2s-d/2)/(2L)$. Starting with $(s_0^2,-\{a_2,s_0^2\})$ and $\mathcal{E}^0$ and using Proposition 3.1, we construct a sequence of conic neighborhoods $\mathcal{E}^l\subset\mathcal{E}^{l-1}\subset\cdots\subset\mathcal{E}^0$ containing the backward bicharacteristic $\{\varphi(s,x_0,\xi_0), s\leq 0\}$ for $1\leq l\leq 2L$ and pairs of symbols (b_l^0,c_l^0) satisfying (3.1) and (3.2) with $b_l^0(x,\xi)$ in $S(0,2l\eta)$, $c_l^0(x,\xi)$ in $S(1,2l\eta-1)$, and

$$\operatorname{supp}(b_l^0(x,\xi))\subset\{(x,\xi)\,/\,b_{l-1}^0(x,\xi)\geq\langle x\rangle^{2(l-1)\eta}\}=\mathcal{E}^{l-1}. \tag{5.56}$$

Equation (5.46) implies

$$\begin{aligned}\operatorname{Re}\,\langle\ & b_l^0(x,D)v(t),v(t)\rangle+\int_0^t\operatorname{Re}\langle c_l^0(x,D)v(\tau),v(\tau)\rangle d\tau\\ &\leq\operatorname{Re}\langle b_l^0(x,D)v_0,v_0\rangle+\int_0^t\operatorname{Re}\langle e(x,D)v(\tau),v(\tau)\rangle d\tau\\ &\quad+\int_0^t(|\langle b_l^0(x,D)(l_1(x,D)v+T_{l_2}v+r_2(u,\bar u)+(l_\theta^1(x,D)+T_{l_\theta^2})\theta_1u\\ &\quad+(\widetilde{l_\theta^1}(x,D)+T_{\widetilde{l_\theta^2}})\overline{\theta_1u}),v\rangle|+|\langle b_l^0(x,D)v,l_1(x,D)v+T_{l_2}v+r_2(u,\bar u)\\ &\quad+(l_\theta^1(x,D)+T_{l_\theta^2})\theta_1u+(\widetilde{l_\theta^1}(x,D)+T_{\widetilde{l_\theta^2}})\overline{\theta_1u}\rangle|)d\tau+C,\end{aligned} \tag{5.57}$$

where $e(x,\xi)$ is in $S(0,2l\eta-2)$ and hence in $S(0,2(l-1)\eta)$. As $\operatorname{supp}b_l^0(x,\xi)\subset\{s_0(x,\xi)\geq\langle x\rangle^{2s-d/2}\}$, Lemma 3.5 and (5.48) yield

$$|\langle b_l^0(x,D)v_0,v_0\rangle|\leq C(\|s_0(x,D)v_0\|_{L^2}^2+\|v_0\|_{L^2}^2)<+\infty. \tag{5.58}$$

As $\theta=1$ on the support of $b_l^0(x,\xi)$, and as the support of $l_\theta^1(x,\xi)$ and of $\widetilde{l_\theta^1}(x,\xi)$ are included in the support of $\nabla\theta$, Proposition 3.2 implies

$$\begin{aligned}&\int_0^t(|\langle b_l^0(x,D)(l_\theta^1(x,D)\theta_1u(\tau)+\widetilde{l_\theta^1}(x,D)\overline{\theta_1u}(\tau)),v(\tau)\rangle|\\ &\quad+|\langle b_l^0(x,D)v(\tau),l_\theta^1(x,D)\theta_1u(\tau)+\widetilde{l_\theta^1}(x,D)\overline{\theta_1u}(\tau)\rangle|)d\tau\\ &\leq C\int_0^T\|u(\tau)\|_{L^2}^2<+\infty.\end{aligned} \tag{5.59}$$

Using Proposition 3.2, we have

$$\begin{aligned}b_l^0(x,D)&=\sqrt{b_l^0}(x,D)^*\sqrt{b_l^0}(x,D)+e_1(x,D),\\ b_l^0(x,D)l_1(x,D)&=\sqrt{b_l^0}(x,D)^*l_1(x,D)\sqrt{b_l^0}(x,D)+e_2(x,D),\end{aligned} \tag{5.60}$$

where $e_1(x,\xi)$ and $e_2(x,\xi)$ are in $S(-1,2l\eta-1)$ and, hence, in $S(0,2(l-1)\eta)$. So

$$
\begin{aligned}
(5.61)\quad & |\langle b_l^0(x,D)((l_1(x,D)+T_{l_2})v+r_2(u,\bar{u})+T_{l_\theta^2}\theta_1 u+T_{\widetilde{l_\theta^2}}\overline{\theta_1 u}),v(\tau)\rangle| \\
& \leq C(\|\sqrt{b_l^0}(x,D)v(\tau)\|_{L^2}^2+\|\sqrt{b_l^0}(x,D)(T_{l_2}v+r_2(u,\bar{u})+T_{l_\theta^2}\theta_1 u+T_{\widetilde{l_\theta^2}}\overline{\theta_1 u})\|_{L^2}^2) \\
& \quad +|\langle e_1(x,D)(T_{l_2}v+r_2(u,\bar{u})+T_{l_\theta^2}\theta_1 u+T_{\widetilde{l_\theta^2}}\overline{\theta_1 u}),v(\tau)\rangle| \\
& \quad +|\langle e_2(x,D)v(\tau),v(\tau)\rangle|.
\end{aligned}
$$

Moreover, as $\sqrt{b_l^0}(x,D)\langle x\rangle^{-(2s-d/2)}$ is bounded on L^2 and $\langle x\rangle^{2s-d/2}r_2(u,\bar{u})$ is in $L^\infty(0,T,L^2)$, we get

$$
\big\|\sqrt{b_l^0}(x,D)r_2(u,\bar{u})\big\|_{L^2}^2\leq C. \tag{5.62}
$$

As l_2, l_θ^2, and $\widetilde{l_\theta^2}$ are in $L^\infty(0,T,\widetilde{\Sigma}_s^0)$, as $\sqrt{b_l^0}(x,\xi)$ is in $S(0,2s-d/2)$ for all l, and as v and $\theta_1 u$ are in $L^\infty(0,T,H_0^s)$, Proposition 4.12 implies

$$
\big\|\sqrt{b_l^0}(x,D)(T_{l_2}v+T_{l_\theta^2}\theta_1 u+T_{\widetilde{l_\theta^2}}\overline{\theta_1 u})\big\|_{L^2}^2\leq C. \tag{5.63}
$$

Finally, Lemma 3.4 with b_l^0 and c_l^0 yields

$$
\begin{aligned}
& \|\sqrt{b_l^0}(x,D)v(t)\|_{L^2}^2+\int_0^t\|\sqrt{c_l^0}(x,D)v(\tau)\|_{L^2}^2d\tau \\
(5.64)\quad & \leq \mathrm{Re}\langle b_l^0(x,D)v(t),v(t)\rangle+\int_0^t\mathrm{Re}\langle c_l^0(x,D)v(\tau),v(\tau)\rangle d\tau \\
& \quad +\mathrm{Re}\langle e_3(x,D)v(t),v(t)\rangle+\int_0^t\mathrm{Re}\langle e_4(x,D)v(\tau),v(\tau)\rangle d\tau+C,
\end{aligned}
$$

where $e_3(x,\xi)$ and $e_4(x,\xi)$ are in $S(0,2(l-1)\eta)$.

When $l=1$, $e(x,D)$, $e_1(x,D)$, $e_2(x,D)$, $e_3(x,D)$, and $e_4(x,\xi)$ are bounded on L^2. Thus, (5.57), (5.58), (5.59), (5.61), (5.62), (5.63), and (5.64) imply

$$
\begin{aligned}
(5.65)\quad & \|\sqrt{b_1^0}(x,D)v(t)\|_{L^2}^2+\int_0^t\|\sqrt{c_1^0}(x,D)v(\tau)\|_{L^2}^2d\tau \\
& \leq C\int_0^t\|\sqrt{b_1^0}(x,D)v(\tau)\|_{L^2}^2d\tau+C.
\end{aligned}
$$

Gronwall's lemma yields

$$
\sup_{0\leq t\leq T}\|\sqrt{b_1^0}(x,D)v(t)\|_{L^2}^2+\int_0^T\|\sqrt{c_1^0}(x,D)v(t)\|_{L^2}^2dt<+\infty. \tag{5.66}
$$

Assume

$$
\sup_{0\leq t\leq T}\|\sqrt{b_{l-1}^0}(x,D)v(t)\|_{L^2}^2+\int_0^T\|\sqrt{c_{l-1}^0}(x,D)v(t)\|_{L^2}^2dt<+\infty \tag{5.67}
$$

for $2 \le l \le L$. As $e_1(x,\xi)$ and $e_2(x,\xi)$ are in $S(0, 2(l-1)\eta)$ with support in $\{b^0_{l-1}(x,\xi) \ge \langle x\rangle^{2(l-1)\eta}\}$ by (5.56), Lemma 3.5 implies

$$\begin{aligned} &|\langle e_1(x,D)(T_{l_2}v + r_2(u,\bar u) + T_{l^2_\theta}\theta_1 u + T_{\widetilde{l^2_\theta}}\overline{\theta_1 u}), v(\tau)\rangle| \\ &\quad + |\langle e_2(x,D)v(\tau), v(\tau)\rangle| \\ &\le C(\|\sqrt{b^0_{l-1}}(x,D)v\|^2_{L^2} + \|\sqrt{b^0_{l-1}}(x,D)r_2(u,\bar u)\|^2_{L^2} \\ &\quad + \|\sqrt{b^0_{l-1}}(x,D)(T_{l_2}v + T_{l^2_\theta}\theta_1 u + T_{\widetilde{l^2_\theta}}\overline{\theta_1 u})\|^2_{L^2}) + C. \end{aligned} \tag{5.68}$$

As for (5.62) and (5.63), we have

$$\begin{aligned} &\big\|\sqrt{b^0_{l-1}}(x,D)r_2(u,\bar u)\big\|^2_{L^2} \le C, \\ &\big\|\sqrt{b^0_{l-1}}(x,D)(T_{l_2}v + T_{l^2_\theta}\theta_1 u + T_{\widetilde{l^2_\theta}}\overline{\theta_1 u})\big\|^2_{L^2} \le C. \end{aligned} \tag{5.69}$$

Inequality (5.67) implies

$$\big\|\sqrt{b^0_{l-1}}(x,D)v(\tau)\big\|^2_{L^2} \le C. \tag{5.70}$$

Then (5.57), (5.58), (5.59), (5.61), (5.62), (5.63), (5.64), (5.68), (5.69), and (5.70) yield

$$\begin{aligned} &\|\sqrt{b^0_l}(x,D)v(t)\|^2_{L^2} + \int_0^t \|\sqrt{c^0_l}(x,D)v(\tau)\|^2_{L^2}d\tau \\ &\qquad \le C\int_0^t \|\sqrt{b^0_l}(x,D)v(\tau)\|^2_{L^2}dt + C + \mathrm{Re}\langle e_3(x,D)v(t), v(t)\rangle \\ &\qquad + \int_0^t \mathrm{Re}\langle e_5(x,D)v(\tau), v(\tau)\rangle dt, \end{aligned} \tag{5.71}$$

where $e_5(x,\xi) = e(x,\xi) + e_4(x,\xi)$ is in $S(0, 2(l-1)\eta)$ with support included in $\{b^0_{l-1}(x,\xi) \ge \langle x\rangle^{2(l-1)\eta}\}$. Lemma 3.5 and (5.67) imply

$$\begin{aligned} &|\langle e_3(x,D)v(t), v(t)\rangle| + \int_0^t |\langle e_5(x,D)v(\tau), v(\tau)\rangle| d\tau \\ &\le C\left(\|\sqrt{b^0_{l-1}}(x,D)v(t)\|^2_{L^2} + \int_0^t \|\sqrt{b^0_{l-1}}(x,D)v(\tau)\|^2_{L^2}d\tau\right) + C < +\infty, \end{aligned} \tag{5.72}$$

which together with (5.71) yields

$$\sup_{0\le t\le T}\|\sqrt{b^0_l}(x,D)v(t)\|^2_{L^2} + \int_0^T \|\sqrt{c^0_l}(x,D)v(t)\|^2_{L^2}dt < +\infty, \tag{5.73}$$

using Gronwall's lemma. Hence, (5.73) is true for $1 \le l \le 2L$ by induction. It suffices to define $(b^0, c^0) = (b^0_l, c^0_l)$ with $l = 2L$. ◻

Proof of Theorem 5.6. Starting with (b^0, c^0) given by Theorem 5.5 and

$$\mathcal{D}_0 = \{(x,\xi)/b^0(x,\xi) \ge \langle x\rangle^{2(2s-d/2)}\} \cap \{(x,\xi)/c^0(x,\xi) \ge \langle x\rangle^{2(2s-d/2)-1}\langle\xi\rangle\}, \tag{5.74}$$

and using Proposition 3.1, we construct a sequence of conic neighborhoods $\mathcal{D}^l \subset \mathcal{D}^{l-1} \subset \cdots \subset \mathcal{D}^0 \subset \mathcal{D}_0 \subset \mathcal{E}^0$ containing $\{\varphi(s, x_0, \xi_0), s \le 0\}$ for $0 \le l \le [2(2s - d/2)]$

and pairs of symbols (b_l, c_l) satisfying (3.1), (3.2), (5.52), and (5.53). For $\mu_l > 0$, (5.47) implies

(5.75)
$$\begin{aligned}
&t^{\mu}\mathrm{Re}\langle b_l(x,D)v(t),v(t)\rangle+\int_0^t\tau^{\mu}\mathrm{Re}\langle c_l(x,D)v(\tau),v(\tau)\rangle d\tau\\
&\leq\int_0^t\tau^{\mu}\mathrm{Re}\langle e(x,D)v(\tau),v(\tau)\rangle d\tau+\int_0^t\mu\tau^{\mu-1}\langle b_l(x,D)v(\tau),v(\tau)\rangle d\tau\\
&+\int_0^t\tau^{\mu}(|\langle b_l(x,D)(l_1(x,D)v+T_{l_2}v+r_2(u,\bar{u})+(l^1_\theta(x,D)+T_{l^2_\theta})\theta_1u+(\widetilde{l^1_\theta}(x,D)\\
&+T_{\widetilde{l^2_\theta}})\overline{\theta_1 u}),v\rangle|+|\langle b_l(x,D)v,l_1(x,D)v+T_{l_2}v+r_2(u,\bar{u})+(l^1_\theta(x,D)+T_{l^2_\theta})\theta_1u\\
&+(\widetilde{l^1_\theta}(x,D)+T_{\widetilde{l^2_\theta}})\overline{\theta_1 u}\rangle|)d\tau+C,
\end{aligned}$$

where $e(x,\xi)$ is in $S(\kappa+l,2(2s-d/2)-l-\kappa-2)$ and hence in $S(\kappa+l,2(2s-d/2)-l-\kappa)$. As $\theta=1$ on the support of $b_l(x,\xi)$, and as the support of $l^1_\theta(x,\xi)$ and of $\widetilde{l^1_\theta}(x,\xi)$ are included in the support of $\nabla\theta$, Proposition 3.2 yields

$$\begin{aligned}
&\int_0^t(|\langle\tau^{\mu_l}b_l(x,D)(l^1_\theta(x,D)\theta_1u(\tau)+\widetilde{l^1_\theta}(x,D)\overline{\theta_1u}(\tau)),v(\tau)\rangle|\\
&\qquad+|\langle\tau^{\mu_l}b_l(x,D)v(\tau),l^1_\theta(x,D)\theta_1u(\tau)\\
&\qquad+\widetilde{l^1_\theta}(x,D)\overline{\theta_1u}(\tau)\rangle|)d\tau\leq C\int_0^T\|u(\tau)\|^2_{L^2}<+\infty.
\end{aligned}\tag{5.76}$$

Moreover, Lemma 3.4 with b_l and c_l implies

$$\begin{aligned}
&t^{\mu_l}\|\sqrt{b_l}(x,D)v(t)\|^2_{L^2}+\int_0^t\tau^{\mu_l}\|\sqrt{c_l}(x,D)v(\tau)\|^2_{L^2}d\tau\\
&\leq\mathrm{Re}\langle t^{\mu_l}b_l(x,D)v(t),v(t)\rangle+\int_0^t\mathrm{Re}\langle\tau^{\mu_l}c_l(x,D)v(\tau),v(\tau)\rangle d\tau\\
&\quad+\mathrm{Re}\langle t^{\mu_l}e_2(x,D)v(t),v(t)\rangle+\int_0^t\mathrm{Re}\langle\tau^{\mu_l}e_3(x,D)v(\tau),v(\tau)\rangle d\tau+C,
\end{aligned}\tag{5.77}$$

where $e_2(x,\xi)$ belongs to $S(l+\kappa-1,2(2s-d/2)-l-\kappa-1)$ and $e_3(x,\xi)$ belongs to $S(l+\kappa,2(2s-d/2)-l-\kappa-2)$.

We define a symbol $d_l(x,\xi)$. If $l=0$, as $b_0(x,\xi)$ is in $S(\kappa,2(2s-d/2)-\kappa)$ and its support is in $\mathcal{D}_0$ defined by (5.74), there exists $d_0(x,\xi)$ in $S(\kappa,2(2s-d/2)-\kappa)$ with $\sqrt{d_0}(x,\xi)\in S(\kappa/2,(2s-d/2)-\kappa/2)$ such that $d_0(x,\xi)$ has its support in $\mathcal{D}_0$ and

$$\mathrm{supp}(b_0(x,\xi))\subset\{(x,\xi)/d_0(x,\xi)\geq\langle x\rangle^{2(2s-d/2)-\kappa}\langle\xi\rangle^{\kappa}\}.\tag{5.78}$$

If $l\geq 1$, as $b_l(x,\xi)$ is in $S(l+\kappa,2(2s-d/2)-l-\kappa)$ and has its support in $\{c_{l-1}(x,\xi)\geq\langle x\rangle^{2(2s-d/2)-l-\kappa}\langle\xi\rangle^{l+\kappa}\}$, there exists $d_l(x,\xi)$ in $S(l+\kappa,2(2s-d/2)-l-\kappa)$ with $\sqrt{d_l}(x,\xi)\in S((l+\kappa)/2,(2s-d/2)-(l+\kappa)/2)$ such that

$$\mathrm{supp}(b_l(x,\xi))\subset\{(x,\xi)\,/\,d_l(x,\xi)\geq\langle x\rangle^{2(2s-d/2)-l-\kappa}\langle\xi\rangle^{l+\kappa}\}\tag{5.79}$$

and

$$\mathrm{supp}(d_l(x,\xi))\subset\{(x,\xi)\,/\,c_{l-1}(x,\xi)\geq\langle x\rangle^{2(2s-d/2)-l-\kappa}\langle\xi\rangle^{l+\kappa}\}.\tag{5.80}$$

By (5.78) and (5.79), Lemma 3.5 implies

(5.81)
$$|\langle b_l(x,D)(l_1(x,D)v + T_{l_2}v + r_2(u,\bar u) + T_{l^2_\theta})\theta_1 u + T_{\widetilde{l^2_\theta}})\overline{\theta_1 u}), v(\tau)\rangle| \\ \leq C(\|\sqrt{d_l}(x,D)v(\tau)\|^2_{L^2} + \|\sqrt{d_l}(x,D)r_2(u,\bar u)\|^2_{L^2} \\ + \|\sqrt{d_l}(x,D)T_{l_2}v\|^2_{L^2} + \|\sqrt{d_l}(x,D)(T_{l^2_\theta}\theta_1 u + T_{\widetilde{l^2_\theta}}\overline{\theta_1 u})\|^2_{L^2}).$$

Moreover, $\sqrt{d_l}(x,D)\langle x\rangle^{(\kappa+l)/2-(2s-d/2)}(1+|D|^2)^{-(\kappa+l)/4}$ is bounded on L^2, and $(1+|D|^2)^{(\kappa+l)/4}\langle x\rangle^{2s-d/2-(\kappa+l)/2}r_2(u,\bar u)$ is in L^2 using Proposition 4.3 and the fact that $r_2(u,\bar u)$ is in $L^\infty(0,T,H_0^{2s-d/2})$. So

$$\|\sqrt{d_l}(x,D)r_2(u,\bar u)\|^2_{L^2} \leq C. \tag{5.82}$$

l^2_θ and $\widetilde{l^2_\theta}$ are in $L^\infty(0,T,\widetilde\Sigma^0_s)$ and they vanish in a neighborhood of the support of $\sqrt{d_l}(x,\xi)$, $\sqrt{d_l}(x,\xi)$ is in $S((l+\kappa)/2,(2s-d/2)-(l+\kappa)/2)$, and v and $\theta_1 u$ are in $L^\infty(0,T,H^s_0)$. Therefore, Proposition 4.12 implies that

$$\|\sqrt{d_l}(x,D)(T_{l^2_\theta}\theta_1 u + T_{\widetilde{l^2_\theta}}\overline{\theta_1 u})\|^2_{L^2} \leq C. \tag{5.83}$$

Finally, Proposition 4.12 implies that

$$\|\sqrt{d_l}(x,D)T_{l_2}v\|^2_{L^2} \leq C\left(\sum_{1\leq j\leq N}\|\sqrt{d^j_l}(x,D)v\|^2_{L^2}+1\right), \tag{5.84}$$

where $d^j_l(x,\xi)$ is in $S((l+\kappa)/2,(2s-d/2)-(l+\kappa)/2)$ and its support is included in the support of $\sqrt{d_l}(x,\xi)$. Inequalities (5.81), (5.82), (5.83), and (5.84) yield

(5.85)
$$|\langle b_l(x,D)(l_1(x,D)v + T_{l_2}v + r_2(u,\bar u) + T_{l^2_\theta}\theta_1 u + T_{\widetilde{l^2_\theta}}\overline{\theta_1 u}), v(\tau)\rangle| \\ \leq C\left(\|\sqrt{d_l}(x,D)v(\tau)\|^2_{L^2} + \sum_{1\leq j\leq N}\|d^j_l(x,D)v(\tau)\|^2_{L^2}\right) + C.$$

First we prove (5.54) when $l=0$. Propositions 3.2 and 3.3 imply the existence of $e_4(x,\xi)$ in $S(\kappa, 2(2s-d/2)-\kappa)$ with support in $\mathcal{D}_0$ such that

$$\sqrt{d_0}(x,D)^*\sqrt{d_0}(x,D) + \sum_{1\leq j\leq N} d^j_0(x,D)^* d^j_0(x,D) = e_4(x,D), \tag{5.86}$$

which together with (5.75), (5.76), (5.77), and (5.85) yields

(5.87)
$$t^{\mu_0}\|\sqrt{b_0}(x,D)v(t)\|^2_{L^2} + \int_0^t \tau^{\mu_0}\|\sqrt{c_0}(x,D)v(\tau)\|^2_{L^2}d\tau \\ \leq \mathrm{Re}\langle t^{\mu_0}e_2(x,D)v(t),v(t)\rangle + \int_0^t \mathrm{Re}\langle \tau^{\mu_0}e_5(x,D)v(\tau),v(\tau)\rangle d\tau \\ + \int_0^t \mu_0\mathrm{Re}\langle\tau^{\mu_0-1}b_0(x,D)v(\tau),v(\tau)\rangle d\tau + C,$$

where $e_5(x,\xi) = e(x,\xi) + e_3(x,\xi) + e_4(x,\xi)$. As $e_2(x,\xi)$ is in $S(\kappa - 1, 2s - d/2 - \kappa - 1)$ and has its support in $\mathcal{D}_0$, Lemma 3.5 and Theorem 5.5 yield

$$\mathrm{Re}\langle e_2(x,D)v(t), v(t)\rangle \le C\|\sqrt{b^0}(x,D)v(t)\|_{L^2}^2 + C < +\infty. \tag{5.88}$$

We choose μ_0 such that $\kappa < \mu_0 < 1$, which implies that $t^{-(1-\mu_0)/(1-\kappa)}$ is locally integrable in t. As $e_5(x,\xi)$ and $b_0(x,\xi)$ have their support in $\mathcal{D}_0$ defined by (5.74), we have

$$\begin{aligned} |t^{\mu_0} e_5(x,\xi)| &\le C(b^0(x,\xi))^{1-\kappa}(t^{\mu_0/\kappa}c^0(x,\xi))^{\kappa} \le C(b^0(x,\xi) + t^{\mu_0/\kappa}c^0(x,\xi)),\\ |t^{\mu_0-1}b_0(x,\xi)| &\le C(t^{-(1-\mu_0)/(1-\kappa)}b^0(x,\xi))^{1-\kappa}(c^0(x,\xi))^{\kappa}\\ &\le C(t^{-(1-\mu_0)/(1-\kappa)}b^0(x,\xi) + c^0(x,\xi)), \end{aligned}$$

and Lemma 3.4 with $C(b^0(x,\xi) + t^{\mu_0/\kappa}c^0(x,\xi)) - t^{\mu_0}e_5(x,\xi)$ and $C(b^0(x,\xi) + t^{(1-\mu_0)/(1-\kappa)}c^0(x,\xi)) - t^{\kappa/(1-\kappa)(1-\mu_0)}b_0(x,\xi)$ implies

$$\begin{aligned} &\int_0^t \mathrm{Re}\langle \tau^{\mu_0} e_5(x,D)v(\tau), v(\tau)\rangle d\tau + \int_0^t \mu_0 \mathrm{Re}\langle \tau^{\mu_0-1} b_0(x,D)v(\tau), v(\tau)\rangle d\tau\\ &\quad \le C \sup_{0\le t\le T} \mathrm{Re}\langle b^0(x,D)v(t), v(t)\rangle + C\int_0^t \mathrm{Re}\langle c^0(x,D)v(\tau), v(\tau)\rangle d\tau\\ &\qquad + C\sup_{0\le t\le T} \mathrm{Re}\langle e_6(t,x,D)v(t), v(t)\rangle + C, \end{aligned} \tag{5.89}$$

where $e_6(t,x,\xi)$ is in $S(0, 2(2s - d/2))$ with seminorms bounded independently of t for $0 \le t \le T$ and with support in $\mathcal{D}_0$. Lemma 3.5 and Theorem 5.5 yield

$$\begin{aligned} \sup_{0\le t\le T} \mathrm{Re}\langle b^0(x,D)v(t), v(t)\rangle &+ \int_0^T \mathrm{Re}\langle c^0(x,D)v(\tau), v(\tau)\rangle d\tau\\ &+ \sup_{0\le t\le T} \mathrm{Re}\langle e_6(t,x,D)v(t), v(t)\rangle < +\infty, \end{aligned} \tag{5.90}$$

which together with (5.87), (5.88), and (5.89) implies

$$\sup_{0\le t\le T} t^{\mu_0}\|\sqrt{b_0}(x,D)v(t)\|_{L^2}^2 + \int_0^T t^{\mu_0}\|\sqrt{c_0}(x,D)v(t)\|_{L^2}^2 dt < +\infty. \tag{5.91}$$

We define $\mu_l = \mu_0 + l$. Assume

$$\sup_{0\le t\le T} t^{\mu_{l-1}}\|\sqrt{b_{l-1}}(x,D)v(t)\|_{L^2}^2 + \int_0^T t^{\mu_{l-1}}\|\sqrt{c_{l-1}}(x,D)v(t)\|_{L^2}^2 dt < +\infty \tag{5.92}$$

for $1 \le l \le [2(2s - d/2)]$. As $\sqrt{d_l}(x,D)^*\sqrt{d_l}(x,D) + \sum_{1\le j\le N} d_l^j(x,D)^* d_l^j(x,D)$ is in $Op(S(l+\kappa, 2(2s-d/2) - l - \kappa))$ and as its symbol is supported in $\{c_{l-1}(x,\xi) \ge \langle x\rangle^{2(2s-d/2)-l-\kappa}\langle\xi\rangle^{l+\kappa}\}$, Lemma 3.5 and (5.92) imply

$$\begin{aligned} &\int_0^t \tau^{\mu_l}\left(\|\sqrt{d_l}(x,D)v(\tau)\|_{L^2}^2 + \sum_{1\le j\le N}\|d_l^j(x,D)v(\tau)\|_{L^2}^2\right)d\tau\\ &\quad \le C\int_0^t \tau^{\mu_{l-1}}\|\sqrt{c_{l-1}}(x,D)v(\tau)\|_{L^2}^2 d\tau + C < +\infty. \end{aligned} \tag{5.93}$$

Then (5.75), (5.76), (5.77), (5.85), and (5.93) yield

$$
\begin{aligned}
(5.94)\qquad & t^{\mu_l}\|\sqrt{b_l}(x,D)v(t)\|_{L^2}^2+\int_0^t \tau^{\mu_l}\|\sqrt{c_l}(x,D)v(\tau)\|_{L^2}^2 d\tau \\
&\le \mathrm{Re}\langle t^{\mu_l}e_2(x,D)v(t),v(t)\rangle+\int_0^t \mathrm{Re}\langle \tau^{\mu_l}e_5(x,D)v(\tau),v(\tau)\rangle d\tau \\
&\quad +\int_0^t \mu_l \mathrm{Re}\langle \tau^{\mu_l-1}b_l(x,D)v(\tau),v(\tau)\rangle d\tau + C.
\end{aligned}
$$

$b_l(x,\xi)$ and $e_5(x,\xi)$ are in $S(l+\kappa,2(2s-d/2)-l-\kappa)$ and have their support in $\{c_{l-1}(x,\xi)\ge \langle x\rangle^{2(2s-d/2)-l-\kappa}\langle\xi\rangle^{l+\kappa}\}$. $e_2(x,\xi)$ is in $S(l+\kappa-1,2(2s-d/2)-l-\kappa-1)$ and has its support in $\{b_{l-1}(x,\xi)\ge\langle x\rangle^{2(2s-d/2)-l-\kappa+1}\langle\xi\rangle^{l+\kappa-1}\}$. Thus, Lemma 3.5 implies

$$
\begin{aligned}
(5.95)\qquad & \mathrm{Re}\langle t^{\mu_l}e_2(x,D)v(t),v(t)\rangle+\int_0^t \mathrm{Re}\langle \tau^{\mu_l}e_5(x,D)v(\tau),v(\tau)\rangle d\tau \\
&+\int_0^t \mu_l\mathrm{Re}\langle \tau^{\mu_l-1}b_l(x,D)v(\tau),v(\tau)\rangle d\tau \le C\sup_{0\le t\le T} t^{\mu_{l-1}}\|\sqrt{b_{l-1}}(x,D)v\|_{L^2}^2 \\
&+C\int_0^t \tau^{\mu_{l-1}}\|\sqrt{c_{l-1}}(x,D)v(\tau)\|_{L^2}^2 d\tau + C,
\end{aligned}
$$

which together with (5.94) and (5.92) implies

$$
(5.96)\qquad t^{\mu_l}\|\sqrt{b_l}(x,D)v(t)\|_{L^2}^2+\int_0^t \tau^{\mu_l}\|\sqrt{c_l}(x,D)v(\tau)\|_{L^2}^2 d\tau \le C.
$$

Thus, (5.54) is true for all $0\le l\le[2(2s-d/2)]$ by induction. □

We can now prove Theorems 2.3 and 2.4.

Proof of Theorem 2.3. We first check that v_0 satisfies (5.48). As $\mathcal{E}$ satisfies (2.7), Proposition 3.1 implies the existence of a conic neighborhood $\mathcal{E}^0\subset\mathcal{E}$ of $\{\varphi(s,x_0,\xi_0), s\le 0\}$ and of $s_0(x,\xi)$ in $S(0,2s-d/2)$ with support in $\mathcal{E}$ such that $(s_0^2,-\{a_2,s_0^2\})$ satisfies (3.1) and (3.2) and $\langle x\rangle^{2(2s-d/2)}\le s_0^2(x,\xi)$ on the set $\mathcal{E}^0\cap\{|\xi|\ge 1\}$. As θu_0 is in H_0^s, as $q_2(0,x,\xi)$ is in $\widetilde{\Sigma}_s^0$ and as $s_0(x,\xi)$ is in $S(0,2s-d/2)$, we have

$$
(5.97)\qquad \|s_0(x,D)T_{q_2(0,.)}\overline{\theta u_0}\|_{L^2}\le C
$$

by Proposition 4.12. Moreover, by the assumptions of Theorem 2.3, $s(x,\xi)$ is in $S(0,2s-d/2)$ and $\langle x\rangle^{2(2s-d/2)}\le s^2(x,\xi)$ on the set $\mathcal{E}$. As θ is in $S(0,0)$, Lemma 3.5 and (2.11) yield

$$
(5.98)\qquad \|s_0(x,D)\theta u_0\|_{L^2}\le C\|s(x,D)u_0\|_{L^2}\le C,
$$

and as $q_1(x,\xi)$ and θ are in $S(0,0)$, Lemma 3.5 and (2.12) yield

$$
(5.99)\qquad \|s_0(x,D)q_1(x,D)\overline{\theta u_0}\|_{L^2}\le C\|s(x,D)\overline{u_0}\|_{L^2}\le C.
$$

Finally, as $v_0=\theta u_0-(q_1(x,D)+T_{q_2(0,.)})\overline{\theta u_0}$, (5.98), (5.97), and (5.99) imply that v_0 satisfies (5.48). So, the assumptions of Theorem 5.6 are satisfied.

For $l=[2(2s-d/2)]$, the symbol $b_l(x,\xi)$ of Theorem 5.6 is in $S(2(2s-d/2),0)$. Therefore, $\sqrt{b_l}(x,\xi)$ is in $S(2s-d/2,0)$ and hence in $S^{2s-d/2}$. Inequality (5.54) yields for all $0<t\le T$,

$$
(5.100)\qquad \|\sqrt{b_l}(x,D)v(t)\|_{L^2}<+\infty,
$$

which implies that $v(t,.)$ is microlocally $H^{2s-d/2}$ in (x_0,ξ_0) for all $0<t\le T$ since $\sqrt{b_l}(x,\xi)$ is elliptic at (x_0,ξ_0).

Moreover, for $l=[2(2s-d/2)]$, the symbol $c_l(x,\xi)$ of Theorem 5.6 is in $S(2(2s-d/2)+1,-1)$. Therefore, $\sqrt{c_l}(x,\xi)$ is in $S(2s-d/2+1/2,-1/2)$ and hence in $S^{2s-d/2+1/2}$. Inequality (5.54) yields for all $0<t\le T$ and for all $\delta>0$

$$\int_0^T t^{2(2s-d/2)+\delta}\|\sqrt{c_l}(x,D)v(t,.)\|_{L^2}^2 dt<+\infty. \tag{5.101}$$

Assume that $u(t,.)$ is microlocally H^σ at $(x_0,-\xi_0)$ for a time t, $0<t\le T$. So, $\overline{\theta u}(t,.)$ is microlocally H^σ at (x_0,ξ_0). As $q_1(x,\xi)$ is in S^{-2} and $q_2(t,x,\xi)$ is in Bony's class $\Sigma^{-2}_{s-d/2}$, $(q_1(x,D)+T_{q_2(t,.)})\overline{\theta u}$ is microlocally $H^{\sigma+2}$ at (x_0,ξ_0). Therefore, $u(t,.)$ is microlocally $H^{\min(\sigma+2,2s-d/2)}$ at (x_0,ξ_0) since $v(t,.)$ is microlocally $H^{2s-d/2}$ at (x_0,ξ_0) and $\theta u=v+(q_1(x,D)+T_{q_2})\overline{\theta u}$.

Assume there exists a pseudodifferential operator whose symbol $c(x,\xi)$ is in S^0 and is elliptic at $(x_0,-\xi_0)$ such that

$$\int_0^T t^{2(2s-d/2)+\delta}\|c(x,D)u(t,.)\|_{H^\sigma}^2 dt<+\infty \tag{5.102}$$

for a $\delta>0$. We may assume that $c(x,\xi)$ is equal to 1 in a conic neighborhood of $(x_0,-\xi_0)$ and that θ is equal to 1 in a neighborhood of the support of $c^J(x,\xi)$. We choose $c_1(x,\xi)$ in $S(0,0)$ with compact support in x, such that $\sqrt{c_l}(x,\xi)$ is elliptic and $c^J(x,\xi)\equiv 1$ in a neighborhood of the support of $c_1(x,\xi)$. Lemma 3.5 and (5.101) yield

$$\begin{aligned}\int_0^T & t^{2(2s-d/2)+\delta}\|c_1(x,D)v(t,.)\|_{H^{2s-d/2+1/2}}^2 dt\\ &\le \int_0^T t^{2(2s-d/2)+\delta}\|\sqrt{c_l}(x,D)v(t,.)\|_{L^2}^2 dt<+\infty.\end{aligned} \tag{5.103}$$

Moreover,

$$\begin{aligned}c_1(x,D)Q\overline{\theta u}&=c_1(x,D)Qc^J(x,D)\overline{\theta u}+c_1(x,D)Q(1-c^J(x,D))\overline{\theta u}\\ &=c_1(x,D)Q\theta\overline{c(x,D)u}+c_1(x,D)Q[c^J(x,D),\theta]\bar u\\ &\quad+c_1(x,D)Q(1-c^J(x,D))\overline{\theta u}.\end{aligned} \tag{5.104}$$

θ is equal to 1 in a neighborhood of the support of $c^J(x,\xi)$ and $c^J(x,\xi)\equiv 1$ in a neighborhood of the support of $c_1(x,\xi)$. Thus, $[c^J(x,D),\theta]$ and $c_1(x,D)q_1(x,D)(1-c^J(x,D))$ are bounded from L^2 to $H^{2s-d/2+1/2}$. Moreover, the symbolic calculus of Bony implies that $c_1(x,D)T_{q_2}(1-c^J(x,D))$ is bounded from L^2 to $H^{2s-d/2+2}$. So, (5.104) yields

$$\begin{aligned}\int_0^T & t^{2(2s-d/2)+\delta}\|c_1(x,D)Q\overline{\theta u}\|_{H^{\min(\sigma+2,2s-d/2+1/2)}}^2 dt\\ &\le \int_0^T t^{2(2s-d/2)+\delta}\|c_1(x,D)Q\theta\overline{c(x,D)u}\|_{H^{\sigma+2}}^2 dt+C\\ &\le C\left(\int_0^T t^{2(2s-d/2)+\delta}\|c(x,D)u\|_{H^\sigma}^2 dt+1\right)<+\infty.\end{aligned} \tag{5.105}$$

As $\theta u = v + (q_1(x,D) + T_{q_2})\overline{\theta u}$, (5.103) and (5.105) imply that

$$\int_0^T t^{2(2s-d/2)+\delta} \|c_1(x,D)u(t,.)\|^2_{H^{\min(\sigma+2,2s-d/2+1/2)}} dt < +\infty, \tag{5.106}$$

since θ is equal to 1 in a neighborhood of the support of $c_1(x,\xi)$.

Finally, assume that $\partial_{\bar{u}} f(0,0) = 0$. Then, we may choose $q_1(x,\xi) \equiv 0$ by Proposition 5.1. Therefore, inequality (5.99) is always satisfied. So, we do not need assumption (2.12). □

Proof of Theorem 2.4. As θu is in $L^\infty(0,T,H^s)$, $u(t,.)$ is microlocally H^s in (x_0,ξ_0) and in $(x_0,-\xi_0)$ for all $0 < t \leq T$. Moreover, there exists $c_1(x,\xi)$ and $c_2(x,\xi)$ in S^0, with $c_1(x,\xi)$ elliptic in (x_0,ξ_0) and $c_2(x,\xi)$ elliptic in $(x_0,-\xi_0)$, such that

$$\int_0^T t^{2(2s-d/2)+\delta} \|c_j(x,D)u(t,.)\|^2_{H^s} dt < +\infty, \quad j = 1,2. \tag{5.107}$$

Starting from $\sigma = s$, it suffices to use Theorem 2.3 with (x_0,ξ_0) and with $(x_0,-\xi_0)$ as many times as needed. □

Acknowledgment. The author would like to thank Professor J. M. Delort for stimulating discussions and careful reading of the manuscript.

REFERENCES

[1] S. Alinhac and P. Gérard, *Opérateurs pseudo-différentiels et théorème de Nash-Moser*, Interéditions, Savoirs actuels, Paris, 1991.

[2] J. M. Bony, *Calcul symbolique et propagation des singularités pour les équations aux dérivées partielles non linéaires*, Ann. Sci. École Norm. Sup (4), 14 (1981), pp. 209–246.

[3] L. Boutet de Monvel, *Propagation des singularités des solutions d'équations analogues à l'équation de Schrödinger* in Fourier Integral Operators and Partial Differential Equations, Lecture Notes in Math. 459, Springer-Verlag, Berlin, 1975, pp. 1–14.

[4] J. Y. Chemin, *Fluides parfaits incompressibles*, Astérisque, 230 (1995).

[5] H. Chihara, *Gain of regularity for semilinear Schrödinger equations*, Math. Ann., 315 (1999), pp. 529–567.

[6] W. Craig, T. Kappeler, and W. Strauss, *Microlocal dispersive smoothing for the Schrödinger equation*, Comm. Pure Appl. Math., 48 (1995), pp. 769–860.

[7] S. Doi, *Smoothing effects for Schrödinger evolution equation and global behaviour of geodesic flow*, Math. Ann., 318 (2000), pp. 355–389.

[8] N. Hayashi, K. Nakamitsu, and M. Tsutsumi, *On solutions of the initial value problem for the nonlinear Schrödinger equations*, J. Funct. Anal., 71 (1987), pp. 218–245.

[9] L. Hörmander, *The Weyl calculus of pseudo-differential operators*, Comm. Pure Appl. Math., 32 (1979), pp. 359–443.

[10] L. Hörmander, *Lectures on Nonlinear Hyperbolic Differential Equations*, Springer-Verlag, Berlin, Heidelberg, 1997.

[11] C. Kenig, G. Ponce, and L. Vega, *Smoothing effects and local existence theory for the generalized nonlinear Schrödinger equations*, Invent. Math., 134 (1998), pp. 489–545.

[12] R. Lascar, *Propagation des singularités des solutions d'équations pseudodifférentielles quasihomogènes*, Ann. Inst. Fourier (Grenoble), 27 (1977), pp. 79–123.

[13] Y. Meyer, *Remarques sur un théorème de J. M. Bony*, Rend. Circ. Mat. Palermo (2), Suppl. 1 (1981), pp. 1–20.

[14] L. Robbiano and C. Zuily, *Microlocal analytic smoothing effect for the Schrödinger equation*, Duke Math. J., 100 (1999), pp. 93–129.

[15] L. Robbiano and C. Zuily, *Analytic theory for the quadratic scattering wave front set and application to the Schrödinger equation*, Astérisque, 283 (2002).

[16] J. Szeftel, *Propagation et réflexion des singularités pour l'équation de Schrödinger non linéaire*, Ann. Inst. Fourier (Grenoble), 55 (2005), pp. 573–671.

[17] J. Wunsch, *Propagation of singularities and growth for Schrödinger operators*, Duke Math. J., 98 (1999), pp. 137–186.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 598–618

EXISTENCE, UNIQUENESS, AND VARIATIONAL METHODS FOR SCATTERING BY UNBOUNDED ROUGH SURFACES*

SIMON N. CHANDLER-WILDE† AND PETER MONK‡

Abstract. In this paper we study, via variational methods, the problem of scattering of time harmonic acoustic waves by an unbounded sound soft surface. The boundary ∂D is assumed to lie within a finite distance of a flat plane and the incident wave is that arising from an inhomogeneous term in the Helmholtz equation whose support lies within some finite distance of the boundary ∂D. Via analysis of an equivalent variational formulation, we provide the first proof of existence of a unique solution to a three-dimensional rough surface scattering problem for an arbitrary wave number. Our method of analysis does not require any smoothness of the boundary which can, for example, be the graph of an arbitrary bounded continuous function. An attractive feature is that all constants in a priori bounds, for example the inf-sup constant of the sesquilinear form, are bounded by explicit functions of the wave number and the maximum surface elevation.

Key words. nonsmooth boundary, radiation condition, a priori estimate, inf-sup constant, Helmholtz equation

AMS subject classifications. 35J05, 35J20, 35J25, 42B10, 78A45

DOI. 10.1137/040615523

1. Introduction. This paper is concerned with the development and analysis of a variational formulation for scattering by unbounded surfaces, in particular, with the study of what are termed *rough surface scattering problems* in the engineering literature. We shall use the phrase *rough surface* to denote surfaces which are a (usually nonlocal) perturbation of an infinite plane surface such that the whole surface lies within a finite distance of the original plane. Such problems arise frequently in applications, for example in modeling acoustic and electromagnetic wave propagation over outdoor ground and sea surfaces, and are the subject of intensive studies in the engineering literature, with a view to developing both rigorous methods of computation and approximate, asymptotic, or statistical methods (see, e.g., the reviews and monographs by Ogilvy [23], Voronovich [26], Saillard and Sentenac [24], Warnick and Chew [27], and de Santo [13]).

In this paper we will focus on a particular, typical problem of the class, which models time harmonic acoustic scattering by a sound soft rough surface. In particular, we seek to solve the Helmholtz equation with wave number $k>0$, $\Delta u + k^2 u = g$, in the perturbed half-plane or half-space $D \subset \mathbb{R}^n$, $n = 2, 3$. We suppose that the homogeneous Dirichlet boundary condition $u = 0$ holds on ∂D, and a suitable radiation condition is imposed to select a unique solution to this problem. We shall give in the next section complete details about our assumptions on D and on the radiation condition, but we now note that the inhomogeneous term g might be in $L^2(D)$ with bounded support, or be a more general distribution. In addition the boundary ∂D may or may not be the graph of a function.

*Received by the editors September 22, 2004; accepted for publication (in revised form) February 1, 2005; published electronically October 27, 2005. This work was supported by the UK Engineering and Physical Sciences Research Council under grant GR/S67401.

http://www.siam.org/journals/sima/37-2/61552.html

†Department of Mathematics, University of Reading, Whiteknights, PO Box 220, Berkshire RG6 6AX, UK (S.N.Chandler-Wilde@reading.ac.uk).

‡Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 (monk@math.udel.edu). This author's research is supported by AFOSR grant F49620-02-1-0071.

The main results of the paper are the following. In the next section we formulate the boundary value problem precisely, in the case when $g \in L^2(D)$ with support lying within a finite distance of ∂D. We also establish the equivalent variational formulation that we use and study in this paper. As part of the boundary value problem formulation we require the radiation condition often used in a formal manner in the rough surface scattering literature (e.g., [13]) that, above the rough surface and the support of g, the solution can be represented in integral form as a superposition of upward traveling and evanescent plane waves. This radiation condition is equivalent to the upward propagating radiation condition proposed for two-dimensional (2D) rough surface scattering problems in [11], and has recently been analyzed carefully in the 2D case by Arens and Hohage [5]. Arens and Hohage also propose a further equivalent radiation condition (a "pole condition").

In section 3 we analyze the variational formulation in the long wavelength case, showing that the sesquilinear form is then elliptic, so that unique existence of solution and explicit bounds on the solution in terms of the data g follow from the Lax–Milgram lemma.

In section 4 we show that, for an arbitrary wave number k, the variational problem and the equivalent boundary value problem remain well-posed in the case when the rough surface has the property that if x lies in D, then every point above x lies in D. Our methods of argument, which depend on an a priori estimate established via a Rellich-type identity, application of the generalized Lax–Milgram theory of Babuška, and results on approximation of nonsmooth by smooth domains, lead to simple, explicit lower bounds on the inf-sup constant of the sesquilinear form and corresponding explicit bounds on the solution in terms of the data g. We note that, in contrast to earlier uniqueness and existence results for rough surface scattering problems, no additional regularity conditions on the boundary are required; our theorem applies, for example, whenever the boundary ∂D is the graph of a bounded continuous function.

The results and methods of our paper are closest to those of Kirsch [20] and Elschner [16]. These authors study the same problem tackled in this paper, but consider the 2D diffraction grating case when $\partial D = \{(x_1, f(x_1)) : x_1 \in \mathbb{R}\}$ with f periodic and g quasi-periodic (i.e., $g(x)e^{i\alpha x_1}$ is periodic with the same period as f for some $\alpha \in \mathbb{R}$). The variational formulation we propose for the rough surface scattering problem is analogous to that considered for the periodic case in [20, 16]. We note, however, that the periodicity simplifies the mathematical arguments considerably compared to the case we study; the variational formulation is over a bounded region, part of a single period of the domain, so that compact embedding arguments can be applied and the sesquilinear form which arises satisfies a Gårding inequality for all wave numbers. We note, moreover, that the methods of [20, 16] require f to be at least Lipschitz continuous, and do not lead to explicit bounds on stability constants.

The methods of argument used to prove uniqueness in [20, 16] derive, in part, from Alber [2] and Cadilhac [7]. In fact the argument outlined in [7] for the 2D diffraction grating problem could be adapted to prove uniqueness of solution for our boundary value problem in the case when ∂D is the graph of a sufficiently smooth function. However, we will prefer to establish uniqueness via an a priori bound which also leads to an existence result.

In the general 2D case when f is not periodic, existence of a unique solution to the boundary value problem we study has been established via integral equation methods in the case that $f \in C^2(\mathbb{R})$ (∂D is C^2), and well-posedness of the integral equation formulation has been established in a variety of function spaces [11, 10, 9, 3, 4]. The extension to the case when ∂D is Lipschitz is outlined in Zhang [28]. To date, however, the only existence result [8] for the three-dimensional rough surface problem, derived via integral equation methods, applies only to the Dirichlet boundary value problem for the Helmholtz equation when the rough surface is the graph of a sufficiently smooth function with sufficiently small surface slope, and deals only with the case when the wave number is sufficiently small.

In another, somewhat related body of work existence of solution to the Dirichlet problem for the Helmholtz equation, with ∂D unbounded, is established by the limiting absorption method, via a priori estimates in weighted Sobolev spaces (see Eidus and Vinnik [14], Vogelsang [25], Minskii [22], and references therein). The results obtained apply to the problem considered in this paper, but only if we assume that the rough surface approaches a flat boundary sufficiently rapidly at infinity and/or that the sign of $x \cdot \nu(x)$ is constant on ∂D outside a large sphere, where $\nu(x)$ denotes the unit normal at $x \in \partial D$. Moreover, this body of work requires that g decrease sufficiently rapidly at infinity so that a Rellich–Sommerfeld radiation condition is satisfied.

An attractive feature of our results is the explicit bounds we obtain on the solution in terms of the data g, which exhibit explicitly dependence of constants on the wave number and on the geometry of the domain. In part our methods of argument to obtain our bounds are inspired by the work of Melenk [21] and by the closely related work of Cummings and Feng [12]. In these publications bounds, exhibiting explicit dependence on the wave number, are developed for the impedance boundary value problem for the Helmholtz equation in a bounded domain which is either convex or smooth and star-like.

In this paper we propose a variational formulation and exploit it as a theoretical tool to study the well-posedness of the boundary value problem. We anticipate that the variational formulation will also be very suitable for numerical solution via finite element discretization, as are similar formulations for the 2D diffraction grating case [6, 15, 16]. Moreover, the explicit bounds we obtain should be helpful in establishing the dependence, on the wave number and the domain, of the constants in a priori error estimates for finite element schemes. These numerical analysis aspects will be considered in a future paper.

2. The boundary value problem and variational formulation. In this section we shall define some notation related to the rough surface scattering problem and write down the boundary value problem and equivalent variational formulation that will be analyzed in later sections. For $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ $(n = 2, 3)$ let $\tilde{x} = (x_1, \ldots, x_{n-1})$ so that $x = (\tilde{x}, x_n)$. For $H \in \mathbb{R}$ let $U_H := \{x : x_n > H\}$ and $\Gamma_H := \{x : x_n = H\}$. Let $D \subset \mathbb{R}^n$ be a connected open set such that for some constants $f_- < f_+$ it holds that

$$U_{f_+} \subset D \subset U_{f_-}. \tag{2.1}$$

This definition of D (the domain of the acoustic field) allows the rough surface $\Gamma = \partial D$ to be more general than the graph of a function. The variational problem will be posed on the open set $S_H := D \setminus \overline{U}_H$, for some $H \geq f_+$, and we denote the unit outward normal to S_H by ν.

Given a source $g \in L^2(D)$ of compact support, the problem we wish to analyze is to find an acoustic field u such that

$$\Delta u + k^2 u = g \quad \text{in } D, \tag{2.2}$$

$$u = 0 \quad \text{on } \Gamma, \tag{2.3}$$

and such that u satisfies an appropriate radiation condition.

This problem has been studied in a rigorous manner by integral equation methods [10, 9, 30, 3, 4, 28, 8] in the case when Γ is the graph of a sufficiently smooth-bounded function f so that

$$\Gamma = \left\{(\tilde{x}, x_n) \ : \ x_n = f(\tilde{x}), \quad \tilde{x} \in \mathbb{R}^{n-1}\right\} \tag{2.4}$$

with f at least bounded and continuous. The most general results are restricted to the 2D case [10, 9, 30, 3, 4, 28]. In the case $n = 2$ with (2.2) understood in a distributional sense, a solution $u \in C^1(D) \cap C(\overline{D})$ is sought such that u is bounded in every strip S_H, $H > f_+$, and such that u satisfies the upward propagating radiation condition (UPRC) proposed in [10], which states that

$$u(x) = 2\int_{\Gamma_H} \frac{\partial\Phi(x,y)}{\partial y_n} u(y)\, ds(y), \quad x \in U_H, \tag{2.5}$$

for all H such that the support of g is contained in S_H. Here the fundamental solution of the Helmholtz equation Φ is given by

$$\Phi(x,y) = \begin{cases} \dfrac{\mathrm{i}}{4} H_0^{(1)}(k|x-y|), & n = 2, \\ \dfrac{\exp(\mathrm{i}k|x-y|)}{4\pi|x-y|}, & n = 3, \end{cases}$$

for $x, y \in \mathbb{R}^n$, $x \neq y$, where $H_0^{(1)}$ is the Hankel function of the first kind of order zero. Under the assumption that Γ is Lipschitz (i.e., that $f \in C^{0,1}(\mathbb{R}^{n-1})$), and that Γ is piecewise Lyapunov, uniqueness of solution is shown in the 2D case in [11].

To show existence of solution to (2.2)–(2.5) one approach is to first convert the boundary value problem to an equivalent Dirichlet boundary value problem. To do this we need to split u into an incident and scattered field. Introducing the Dirichlet Green's function for the half-space U_a, defined by

$$\mathcal{G}_a(x,y) = \Phi(x,y) - \Phi(x, y_a'),$$

where y_a' is the reflection of y in Γ_a, we define the incident field u_a^i, for $a < f_-$, by

$$u_a^i(x) = -\int_D \mathcal{G}_a(x,y) g(y)\, dy.$$

Then $u_a^i \in H^2_{\mathrm{loc}}(U_a)$ and satisfies (2.2) in a distributional sense in U_a. Choosing $a < f_-$, we write u as

$$u = u_a^i + u^s, \tag{2.6}$$

and seek the scattered field $u^s \in C^2(D) \cap C(\overline{D})$ that satisfies

$$\begin{aligned} \Delta u^s + k^2 u^s &= 0 \quad \text{in } D, \\ u^s &= G \quad \text{on } \Gamma, \end{aligned}$$

where $G := -u_a^i|_\Gamma$. Then u, given by (2.6), satisfies (2.2)–(2.5) provided u^s satisfies (2.5).

In the case $n = 2$ and $f \in C^{1,1}(\mathbb{R})$ it has been shown, for arbitrary bounded and continuous data G, that this Dirichlet problem for u^s has exactly one solution that satisfies the radiation condition (2.5) [10]. Moreover, in the case that $G = -u^i_a|_\Gamma$ it holds that $G(x) = O(|x|^{-3/2})$ as $|x| \to \infty$, and it is shown in [9, 3, 4] that u^s and u inherit this property; precisely that $u(x) = O(|x|^{-3/2})$ as $|x_1| \to \infty$ with $x_2 = O(1)$. Thus $G \in L^2(\Gamma)$ and $u \in L^2(S_H)$ for each $H > f_-$. It follows from local regularity estimates up to the boundary that $u \in C^1(\overline{D})$. Further, by an application of Green's theorem, the Helmholtz equation, and the a priori estimates up to the boundary of [11, Theorem 3.1], it follows also that $u \in H^1(S_H)$ for every $H > f_+$. This in turn implies that $u|_{\Gamma_H} \in H^{1/2}(\Gamma_H) \subset L^2(\Gamma_H)$ for every $H \geq f_+$.

In the case that $u|_{\Gamma_H} \in L^2(\Gamma_H)$ we can rewrite (2.5) in terms of the Fourier transform of $u|_{\Gamma_H}$. For $\phi \in L^2(\Gamma_H)$, which we identify with $L^2(\mathbb{R}^{n-1})$, we denote by $\hat{\phi} = \mathcal{F}\phi$ the Fourier transform of ϕ which we define by

$$\mathcal{F}\phi(\xi) = (2\pi)^{-(n-1)/2} \int_{\mathbb{R}^{n-1}} \exp(-\mathrm{i}\tilde{x}\cdot\xi)\phi(\tilde{x})\, d\tilde{x}, \quad \xi \in \mathbb{R}^{n-1}. \tag{2.7}$$

Our choice of normalization of the Fourier transform ensures that $\mathcal{F}$ is a unitary operator on $L^2(\mathbb{R}^{n-1})$ so that, for $\phi, \psi \in L^2(\mathbb{R}^{n-1})$,

$$\int_{\mathbb{R}^{n-1}} \phi\bar{\psi} d\tilde{x} = \int_{\mathbb{R}^{n-1}} \hat{\phi}\bar{\hat{\psi}} d\xi. \tag{2.8}$$

If $F_H := u|_{\Gamma_H} \in L^2(\Gamma_H)$, then (see [11, 5] in the case $n = 2$) (2.5) can be rewritten as

$$u(x) = \frac{1}{(2\pi)^{(n-1)/2}} \int_{\mathbb{R}^{n-1}} \exp(\mathrm{i}[(x_n - H)\sqrt{k^2 - \xi^2} + \tilde{x}\cdot\xi])\hat{F}_H(\xi)\, d\xi, \quad x \in U_H. \tag{2.9}$$

In this equation $\sqrt{k^2 - \xi^2} = \mathrm{i}\sqrt{\xi^2 - k^2}$, when $|\xi| > k$.

Equation (2.9) is a representation for u, in the upper half-plane U_H, as a superposition of upward propagating homogeneous and inhomogeneous plane waves. A requirement that (2.9) holds is commonly used (e.g., [13]) as a formal radiation condition in the physics and engineering literature on rough surface scattering. The meaning of (2.9) is clear when $F_H \in L^2(\mathbb{R}^{n-1})$ so that $\hat{F}_H \in L^2(\mathbb{R}^{n-1})$; indeed the integral (2.9) exists in the Lebesgue sense for all $x \in U_H$. Recently Arens and Hohage [5] have explained, in the case $n = 2$, in what precise sense (2.9) can be understood when $F_H \in BC(\Gamma_H)$, the space of bounded continuous functions on Γ_H, so that $\hat{F}_H$ must be interpreted as a tempered distribution.

The above discussion motivates the following precise formulation of problem (2.2)–(2.3). Let $H^1_0(D)$ denote the standard Sobolev space, the completion of $C^\infty_0(D)$ in the norm $\|\cdot\|_{H^1(D)}$ defined by $\|u\|_{H^1(D)} = \{\int_D(|\nabla u|^2 + |u|^2)dx\}^{1/2}$. The main function space in which we set our problem will be the Hilbert space V_H, defined, for $H \geq f_+$, by

$$V_H := \left\{\phi|_{S_H} \ : \ \phi \in H^1_0(D)\right\},$$

on which we impose the wave number dependent scalar product $(u, v)_{V_H} := \int_{S_H}(\nabla u \cdot \overline{\nabla v} + k^2 u\bar{v})\, dx$ and norm $\|u\|_{V_H} = \{\int_{S_H}(|\nabla u|^2 + k^2|u|^2)dx\}^{1/2}$.

The boundary value problem. Given $g \in L^2(D)$, whose support lies in $\overline{S_H}$, for some $H \geq f_+$, find $u : D \to \mathbb{C}$ such that $u|_{S_a} \in V_a$ for every $a > f_+$,

$$\Delta u + k^2 u = g \quad \text{in } D \tag{2.10}$$

in a distributional sense, and the radiation condition (2.9) holds with $F_H = u|_{\Gamma_H}$.

REMARK 2.1. *We note that, as one would hope, the solutions of the above problem do not depend on the choice of H. Precisely, if u is a solution to the above problem for one value of $H \geq f_+$ for which* $\operatorname{supp} g \subset \overline{S_H}$, *then u is a solution for all $H \geq f_+$ with this property. To see that this is true is a matter of showing that if* (2.9) *holds for one H with* $\operatorname{supp} g \subset \overline{S_H}$, *then* (2.9) *holds for all H with this property. It is shown in Lemma* 2.2 *that if* (2.9) *holds, with $F_H = u|_{\Gamma_H}$, for some $H \geq f_+$, then it holds for all larger values of H. One way to show that* (2.9) *holds also for every smaller value of H, $\tilde{H}$ say, for which $\tilde{H} \geq f_+$ and* $\operatorname{supp} g \subset \overline{S_{\tilde{H}}}$, *is to consider the function*

$$v(x) := u(x) - \frac{1}{(2\pi)^{(n-1)/2}} \int_{\mathbb{R}^{n-1}} \exp(\mathrm{i}[(x_n - \tilde{H})\sqrt{k^2 - \xi^2} + \tilde{x} \cdot \xi]) \hat{F}_{\tilde{H}}(\xi)\, d\xi, \quad x \in U_{\tilde{H}},$$

with $F_{\tilde{H}} := u|_{\Gamma_{\tilde{H}}}$, and show that v is identically zero. To see this we note that, by Lemma 2.2, *v satisfies the above boundary value problem with $D = U_{\tilde{H}}$ and $g = 0$. That $v \equiv 0$, then follows from Theorem* 4.1.

As indicated in the above discussion, it is known that the above boundary value problem has a solution in the case $n = 2$ when Γ is the graph of a sufficiently smooth function. A main result of this paper is to prove that the boundary value problem is uniquely solvable, both in two and three dimensions, under much more general conditions on the boundary Γ. Moreover, we provide explicit estimates of the norm of the solution in the strip S_H as a function of the dimensionless wave number

$$\kappa = k(H - f_-). \tag{2.11}$$

We now derive a variational formulation of the boundary value problem above. To derive this alternative formulation we require a preliminary lemma. In this lemma and subsequently we use standard fractional Sobolev space notation, except that we adopt a wave number dependent norm, equivalent to the usual norm, and reducing to the usual norm if the unit of length measurement is chosen so that $k = 1$. Thus, identifying Γ_H with $\mathbb{R}^{n-1}$, $H^s(\Gamma_H)$, for $s \in \mathbb{R}$, denotes the completion of $C_0^\infty(\Gamma_H)$ in the norm $\|\cdot\|_{H^s(\Gamma_H)}$ defined by

$$\|\phi\|_{H^s(\Gamma_H)} = \left(\int_{\mathbb{R}^{n-1}} (k^2 + \xi^2)^s |\mathcal{F}\phi(\xi)|^2 \, d\xi \right)^{1/2}.$$

We recall [1] that, for all $a > H \geq f_+$, there exist continuous embeddings $\gamma_+ : H^1(U_H \setminus U_a) \to H^{1/2}(\Gamma_H)$ and $\gamma_- : V_H \to H^{1/2}(\Gamma_H)$ (the trace operators) such that $\gamma_\pm \phi$ coincides with the restriction of ϕ to Γ_H when ϕ is C^∞. In the case when $H = f_+$, when Γ_H may not be a subset of the boundary of S_H (if part of ∂D coincides with Γ_H) we understand this trace by first extending $\phi \in V_H$ by zero to $U_{f_-} \setminus \bar{U}_{f_+}$. We recall also that if $u_+ \in H^1(U_H \setminus U_a)$, $u_- \in V_H$, and $\gamma_+ u_+ = \gamma_- u_-$, then $v \in V_a$, where $v(x) := u_+(x)$, $x \in U_H \setminus U_a$, $:= u_-(x)$, $x \in S_H$. Conversely, if $v \in V_a$ and $u_+ := v|_{U_H \setminus U_a}$, $u_- := v|_{S_H}$, then $\gamma_+ v_+ = \gamma_- v_-$. We introduce the operator T, which will prove to be a Dirichlet to Neumann map on Γ_H (see (2.20)) defined by

$$T := \mathcal{F}^{-1} M_z \mathcal{F}, \tag{2.12}$$

where M_z is the operation of multiplying by

$$z(\xi) := \begin{cases} -\mathrm{i}\sqrt{k^2 - \xi^2} & \text{if } |\xi| \leq k, \\ \sqrt{\xi^2 - k^2} & \text{for } |\xi| > k. \end{cases}$$

We shall prove shortly in Lemma 2.4 that $T : H^{1/2}(\Gamma_H) \to H^{-1/2}(\Gamma_H)$ and is bounded.

LEMMA 2.2. *If* (2.9) *holds, with* $F_H \in H^{1/2}(\Gamma_H)$, *then* $u \in H^1(U_H\backslash U_a)\cap C^2(U_H)$, *for every* $a > H$,

$$\Delta u + k^2 u = 0 \quad \text{in } U_H,$$

$\gamma_+ u = F_H$, *and*

$$\int_{\Gamma_H} \bar{v} T\gamma_+ u \, ds + k^2 \int_{U_H} u\bar{v}\, dx - \int_{U_H} \nabla u \cdot \nabla \bar{v}\, dx = 0, \quad v \in C_0^\infty(D). \tag{2.13}$$

Further, the restrictions of u *and* ∇u *to* Γ_a *are in* $L^2(\Gamma_a)$ *for all* $a > H$, *and*

$$\int_{\Gamma_a} \left[\left| \frac{\partial u}{\partial x_n} \right|^2 - |\nabla_{\tilde{x}} u|^2 + k^2 |u|^2 \right] ds \le 2k\Im \int_{\Gamma_a} \bar{u} \frac{\partial u}{\partial x_n}\, ds. \tag{2.14}$$

Moreover, for all $a > H$, *where* $F_a \in H^{1/2}(\Gamma_a)$ *denotes the restriction of* u *to* Γ_a, (2.9) *holds with* H *replaced by* a.

Proof. If $F_H \in L^2(\Gamma_H)$, then, as a function of ξ, $\exp(i[(x_n - H)\sqrt{k^2 - \xi^2} + \tilde{x}\cdot \xi])\hat{F}_H(\xi)(1 + \xi^2)^s \in L^1(\mathbb{R}^{n-1})$ for every $x \in U_H$ and $s \ge 0$. It follows that (2.9) is well-defined for every $x \in U_H$, and that $u \in C^2(U_H)$, with all partial derivatives computed by differentiating under the integral sign, so that $\Delta u + k^2 u = 0$ in U_H. Thus, for $a > H$ and almost all $\xi \in \mathbb{R}^{n-1}$,

$$\mathcal{F}(u|_{\Gamma_a})(\xi) = \exp\left(\mathrm{i}(a - H)\sqrt{k^2 - \xi^2}\,\right)\hat{F}_H(\xi), \tag{2.15}$$

$$\mathcal{F}\left(\left.\frac{\partial u}{\partial x_n}\right|_{\Gamma_a} \right)(\xi) = \mathrm{i}\sqrt{k^2 - \xi^2}\exp\left(\mathrm{i}(a - H)\sqrt{k^2 - \xi^2}\,\right)\hat{F}_H(\xi), \tag{2.16}$$

$$\mathcal{F}(\nabla_{\tilde{x}} u|_{\Gamma_a})(\xi) = \mathrm{i}\xi \exp\left(\mathrm{i}(a - H)\sqrt{k^2 - \xi^2}\,\right)\hat{F}_H(\xi).$$

Therefore, by the Plancherel identity (2.8), $u|_{\Gamma_a}$, $\nabla u|_{\Gamma_a} \in L^2(\Gamma_a)$ with

$$\int_{\Gamma_a} |u|^2 ds = \int_{\mathbb{R}^{n-1}} \left|\exp\left(2\mathrm{i}(a - H)\sqrt{k^2 - \xi^2}\,\right)\right| |\hat{F}_H(\xi)|^2\, d\xi \le \int_{\Gamma_H} |F_H|^2\, ds$$

and

$$\int_{\Gamma_a} |\nabla u|^2 ds \le \int_{\mathbb{R}^{n-1}} [|k^2 - \xi^2| + \xi^2] \left|\exp\left(2\mathrm{i}(a - H)\sqrt{k^2 - \xi^2}\,\right)\right| |\hat{F}_H(\xi)|^2\, d\xi, \tag{2.17}$$

while

$$\int_{\Gamma_a} \left[\left| \frac{\partial u}{\partial x_n} \right|^2 - |\nabla_{\tilde{x}} u|^2 + k^2|u|^2 \right] ds = 2\int_{|\xi|<k} (k^2 - \xi^2)|\hat{F}_H(\xi)|^2\, d\xi$$

and

$$\Im \int_{\Gamma_a} \bar{u} \frac{\partial u}{\partial x_n}\, ds = \int_{|\xi|<k} \sqrt{k^2 - \xi^2}\, |\hat{F}_H(\xi)|^2\, d\xi.$$

Thus (2.14) holds and

$$\int_{U_H\backslash U_a} |u|^2\, dx \le (a - H) \int_{\Gamma_H} |F_H|^2\, ds. \tag{2.18}$$

Further, from (2.17) it follows that

$$\int_{U_H\setminus U_a} |\nabla u|^2\,dx \le (a-H)k^2 \int_{|\xi|<k} |\hat{F}_H(\xi)|^2\,d\xi + \int_{|\xi|>k} \xi^2 \frac{1-\exp(-2[a-H]\sqrt{\xi^2-k^2}\,)}{\sqrt{\xi^2-k^2}} |\hat{F}_H(\xi)|^2\,d\xi \le \int_{\mathbb{R}^{n-1}} (4(a-H)k^2+\sqrt{2}|\xi|)|\hat{F}_H(\xi)|^2\,d\xi, \tag{2.19}$$

since $1-\mathrm{e}^{-z} \le z$ for $z \ge 0$ and $\sqrt{\xi^2-k^2} \ge |\xi|/\sqrt{2}$ for $\xi^2 \ge 2k^2$. Thus $u \in H^1(U_H\setminus U_a)$ if $F_H \in H^{1/2}(\Gamma_H)$. That $u|_{\Gamma_H} = F_H$ is clear when $F_H \in C_0^\infty(\Gamma_H)$, and $\gamma_+ u = F_H$ for all $F_H \in H^{1/2}(\Gamma_H)$ follows from the continuity of γ_+, (2.18) and (2.19), and the density of $C_0^\infty(\Gamma_H)$ in $H^{1/2}(\Gamma_H)$. Similarly, in the case that $F_H \in C_0^\infty(\Gamma_H)$ so that $u \in C^\infty(\overline{U_H})$, it is easily seen that

$$T\gamma_+ u = -\partial u/\partial x_n|_{\Gamma_H}, \tag{2.20}$$

and (2.13) follows by Green's theorem. The same equation for the general case follows from the density of $C_0^\infty(\Gamma_H)$ in $H^{1/2}(\Gamma_H)$, (2.18) and (2.19), and the continuity of the operator T.

That (2.9) holds with H replaced by a, for all $a > H$, is clear from (2.15). □

Now suppose that u satisfies the boundary value problem. Then $u|_{S_a} \in V_a$ for every $a > f_+$ and, by definition, since $\Delta u + k^2 u = g$ in a distributional sense,

$$\int_D [g\bar{v} + \nabla u\cdot\nabla\bar{v} - k^2 u\bar{v}]dx = 0, \quad v \in C_0^\infty(D). \tag{2.21}$$

Applying Lemma 2.2, and defining $w := u|_{S_H}$, it follows that

$$\int_{S_H} [g\bar{v} + \nabla w\cdot\nabla\bar{v} - k^2 w\bar{v}]\,dx + \int_{\Gamma_H} \bar{v}T\gamma_- w\,ds = 0, \quad v \in C_0^\infty(D).$$

From the denseness of $\{\phi|_{S_H} : \phi \in C_0^\infty(D)\}$ in V_H and the continuity of γ_- and T, it follows that this equation holds for all $v \in V_H$.

Let $\|\cdot\|_2$ and $(\cdot,\cdot)$ denote the norm and scalar product on $L^2(S_H)$ so that $\|v\|_2 = \sqrt{\int_{S_H} |v|^2\,dx}$ and

$$(u,v) = \int_{S_H} u\bar{v}\,dx,$$

and define the sesquilinear form $b : V_H \times V_H \to \mathbb{C}$ by

$$b(u,v) = (\nabla u, \nabla v) - k^2(u,v) + \int_{\Gamma_H} \gamma_-\bar{v}T\gamma_- u\,ds. \tag{2.22}$$

Then we have shown that if u satisfies the boundary value problem, then $w := u|_{S_H}$ is a solution of the following variational problem: find $u \in V_H$ such that

$$b(u,v) = -(g,v), \quad v \in V_H. \tag{2.23}$$

Conversely, suppose that w is a solution to the variational problem and define $u(x)$ to be $w(x)$ in S_H and to be the right-hand side of (2.9), with $F_H := \gamma_- w$, in

U_H. Then, by Lemma 2.2, $u \in H^1(U_H \setminus U_a)$ for every $a > H$ with $\gamma_+ u = F_H = \gamma_- w$. Thus $u|_{S_a} \in V_a$, $a \geq f_+$. Further, from (2.13) and (2.23) it follows that (2.21) holds so that $\Delta u + k^2 u = g$ in D in a distributional sense. Thus u satisfies the boundary value problem.

We have thus proved the following theorem.

THEOREM 2.3. *If u is a solution of the boundary value problem, then $u|_{S_H}$ satisfies the variational problem, Conversely, if u satisfies the variational problem, $F_H := \gamma_- u$, and the definition of u is extended to D by setting $u(x)$ equal to the right-hand side of* (2.9), *for $x \in U_H$, then the extended function satisfies the boundary value problem, with g extended by zero from S_H to D.*

It remains to prove the mapping properties of T.

LEMMA 2.4. *The Dirichlet-to-Neumann map T defined by* (2.12) *is a bounded linear map from $H^{1/2}(\Gamma_H)$ to $H^{-1/2}(\Gamma_H)$ with $\|T\| = 1$.*

Proof. From the definitions of T and the Sobolev norms we see that, as a map from $H^{1/2}(\Gamma_H)$ to $H^{-1/2}(\Gamma_H)$,

$$\|T\| = \max_{\xi \in \mathbb{R}^{n-1}} \frac{|\sqrt{k^2 - \xi^2}|}{|\sqrt{k^2 + \xi^2}|} = 1. \qquad \square$$

While the variational formulation (2.23) does not appear to have been studied previously, the analogous weak formulation for the 2D diffraction grating case has recently been studied in [16], as mentioned in the introduction. The diffraction grating case, with f periodic and g quasi-periodic with the same period, is significantly simpler because the variational problem can be formulated on a bounded domain (one period of the strip S_H) and the corresponding sesquilinear form on this bounded domain satisfies a Gårding inequality. Standard methods of analysis thus apply, in particular, existence follows from uniqueness via the Fredholm alternative. But we note that, even in the diffraction grating case, establishing uniqueness for arbitrary Lipschitz domains D (f Lipschitz) requires careful and ingenious arguments [16] which are not required for scattering by bounded domains. Indeed, uniqueness does not hold in all cases in which ∂D is not the graph of a function, as is shown by the example in Gotlib [18].

3. Analysis of the variational problem for low frequency. In this section we shall derive preliminary results and bounds used throughout, and will analyze (2.23) when k is sufficiently small that b is V_H-elliptic (we shall give an explicit bound for k to guarantee this). An attraction of our results for low wave number, in contrast to our results in section 4 for larger wave number, is that we require no additional assumption on the domain, except that κ, given by (2.11), be sufficiently small. We note also that the bounds we establish for κ small in Theorem 3.1 are somewhat sharper than those which can be established as valid for general κ by the techniques of the next section. From the point of view of numerical solution by, e.g., finite element methods, the ellipticity we establish for small k is of course highly desirable, guaranteeing, by Céa's lemma, unique existence and stability of the numerical solution method.

Let V_H^* denote the dual space of V_H, i.e., the space of continuous antilinear functionals on V_H. Then our analysis will also apply to the following slightly more general problem: given $\mathcal{G} \in V_H^*$ find $u \in V_H$ such that

$$b(u, v) = \mathcal{G}(v), \quad v \in V_H. \tag{3.1}$$

We shall prove the following theorem.

THEOREM 3.1. *Suppose the wave number k satisfies $k < \sqrt{2}/(H - f_-)$ (equivalently $\kappa < \sqrt{2}$). Then the sesquilinear form b is V_H-elliptic so that the variational problem* (3.1) *is uniquely solvable, and the solution satisfies the estimate*

$$\|u\|_{V_H} \leq C\|\mathcal{G}\|_{V_H^*}, \tag{3.2}$$

where the constant C satisfies

$$C \leq \frac{1 + \kappa^2/2}{1 - \kappa^2/2}.$$

In particular, the scattering problem (2.23) *is uniquely solvable and the solution satisfies the bound*

$$k\|u\|_{V_H} \leq \frac{\kappa}{\sqrt{2}} \frac{1 + \kappa^2/2}{1 - \kappa^2/2} \|g\|_2. \tag{3.3}$$

In order to prove Theorem 3.1 we establish a sequence of lemmas which are of some independent interest and are used extensively in the rest of the paper. The first two concern the Dirichlet to Neumann map T and the trace operator γ_- and will be proved using the Fourier transform (2.7).

LEMMA 3.2. *For all $\phi, \psi \in H^{1/2}(\Gamma_H)$,*

$$\int_{\Gamma_H} \phi T\psi \, ds = \int_{\Gamma_H} \psi T\phi \;\, ds.$$

For all $\phi \in H^{1/2}(\Gamma_H)$,

$$\Re \int_{\Gamma_H} \bar{\phi}\, T\phi \, ds \geq 0, \quad \Im \int_{\Gamma_H} \bar{\phi}\, T\phi \, ds \leq 0.$$

Proof. Let $\hat{\phi} = \mathcal{F}\phi$, $\hat{\psi} = \mathcal{F}\psi$. Then $\mathcal{F}(T\phi) = z\hat{\phi}$. Thus, using the Plancherel identity (2.8) and since $\hat{\bar{\psi}}(\xi) = \overline{\hat{\psi}(-\xi)}$ and z is even,

$$\int_{\Gamma_H} \psi\, T\phi \, ds = \int_{\mathbb{R}^{n-1}} \hat{\psi}(-\xi) z(\xi) \hat{\phi}(\xi) \, d\xi = \int_{\mathbb{R}^{n-1}} \hat{\psi}(\xi) z(\xi) \hat{\phi}(-\xi) \, d\xi = \int_{\Gamma_H} \phi\, T\psi \, ds.$$

In particular, putting $\psi = \bar{\phi}$,

$$\begin{aligned}\int_{\Gamma_H} \bar{\phi}\, T\phi \, ds &= \int_{\mathbb{R}^{n-1}} z(\xi) |\hat{\phi}(\xi)|^2 \, d\xi \\ &= \int_{|\xi|>k} \sqrt{\xi^2 - k^2} |\hat{\phi}(\xi)|^2 \, d\xi - \mathrm{i} \int_{|\xi|<k} \sqrt{k^2 - \xi^2} |\hat{\phi}(\xi)|^2 \, d\xi,\end{aligned}$$

from which the second result follows. $\square$

The above lemma implies that $b(\cdot,\cdot)$ has the following important symmetry property.

COROLLARY 3.3. *For all $u, v \in V_H$,*

$$b(v, u) = b(\bar{u}, \bar{v}).$$

LEMMA 3.4. *For all* $u \in V_H$,

$$\|\gamma_- u\|_{H^{1/2}(\Gamma_H)} \le \|u\|_{V_H}$$

and

$$\|u\|_2 \le \frac{H - f_-}{\sqrt{2}} \left\| \frac{\partial u}{\partial x_n} \right\|_2 .$$

Proof. For $u \in C_0^\infty(D) \subset C_0^\infty(U_{f_-})$ and defining $\hat{u}(\xi, x_n) = (\mathcal{F}u(\cdot, x_n))(\xi)$, we have

$$|\hat{u}(\xi, H)|^2 = \int_{f_-}^{H} \frac{\partial}{\partial x_n} |\hat{u}(\xi, x_n)|^2 \, dx_n = 2\Re \int_{f_-}^{H} \overline{\hat{u}(\xi, x_n)} \frac{\partial}{\partial x_n} \hat{u}(\xi, x_n) \, dx_n.$$

Thus, if $S = \mathbb{R}^{n-1} \times (f_-, H)$,

$$\begin{aligned}
\|u\|^2_{H^{1/2}(\Gamma_H)} &= \int_{\mathbb{R}^{n-1}} |\sqrt{\xi^2 + k^2}| \, |\hat{u}(\xi, H)|^2 \, d\xi \\
&\le 2 \int_S |\sqrt{\xi^2 + k^2}| \, |\hat{u}(\xi, x_n)| \left| \frac{\partial}{\partial x_n} \hat{u}(\xi, x_n) \right| d\xi \, dx_n \\
&\le 2 \left\{ \int_S |\xi^2 + k^2| \, |\hat{u}(\xi, x_n)|^2 \, d\xi \, dx_n \right\}^{1/2} \left\{ \int_S \left| \frac{\partial}{\partial x_n} \hat{u}(\xi, x_n) \right|^2 d\xi \, dx_n \right\}^{1/2} .
\end{aligned}$$

Now, by Parseval's theorem,

$$\begin{aligned}
\int_S \xi^2 \, |\hat{u}(\xi, x_n)|^2 \, d\xi \, dx_n &= \int_S |\mathcal{F}(\nabla_{\tilde{x}} u(\cdot, x_n))(\xi)|^2 \, d\xi \, dx_n \\
&= \int_S |\nabla_{\tilde{x}} u(x)|^2 \, dx.
\end{aligned}$$

Applying Parseval's theorem again, and since $2\sqrt{ab} \le a + b$ for $a, b \ge 0$,

$$\|u\|^2_{H^{1/2}(\Gamma_H)} \le 2 \left\{ \int_S \left\{ k^2 |u(x)|^2 + |\nabla_{\tilde{x}} u(x)|^2 \right\} dx \int_S \left| \frac{\partial}{\partial x_n} u(x) \right|^2 dx \right\}^{1/2} \le \|u\|^2_{V_H}.$$

Further, using the fact that $u \in C_0^\infty(U_{f_-})$, for $x \in S$,

$$|u(x)|^2 = \left| \int_{f_-}^{x_n} \frac{\partial}{\partial x_n} u(x) \, dx_n \right|^2 \le (x_n - f_-) \int_{f_-}^{H} \left| \frac{\partial}{\partial x_n} u(x) \right|^2 dx_n$$

so that, since $\int_{f_-}^{H} (x_n - f_-) \, dx_n = (H - f_-)^2/2$,

$$\int_S |u(x)|^2 \, dx \le \frac{(H - f_-)^2}{2} \int_S \left| \frac{\partial}{\partial x_n} u(x) \right|^2 dx. \tag{3.4}$$

Since the set $\{v|_{S_H} : v \in C_0^\infty(D)\}$ is dense in V_H these bounds hold for all $u \in V_H$. $\square$

We are now in a position to prove that the sesquilinear form $b(.,.)$ is bounded, establishing an explicit value for the bound.

LEMMA 3.5. *For all $u, v \in V_H$,*

$$|b(u,v)| \le 2\|u\|_{V_H}\|v\|_{V_H}$$

so that the sesquilinear form $b(.,.)$ is bounded.

Proof. From the definition of the sesquilinear form $b(.,.)$ and the Cauchy–Schwarz inequality, we have

$$|b(u,v)| \le \|\nabla u\|_2\|\nabla v\|_2 + k^2\|u\|_2\|v\|_2 + \|\gamma_- u\|_{H^{1/2}(\Gamma_H)}\|T\|\,\|\gamma_- v\|_{H^{1/2}(\Gamma_H)}.$$

Applying the Cauchy–Schwarz inequality and Lemmas 2.4 and 3.4 we obtain the desired estimate. □

Our last lemma of this section shows that the sesquilinear form $b(.,.)$ is V_H-elliptic provided the wave number k is not too large.

LEMMA 3.6. *For all $u \in V_H$,*

$$\Re\, b(u,u) \ge \frac{1-\kappa^2/2}{1+\kappa^2/2}\,\|u\|^2_{V_H}.$$

Proof. By Lemma 3.2,

$$\Re\, b(u,u) \ge \|u\|^2_{V_H} - 2k^2\|u\|^2_2.$$

The result follows from Lemma 3.4, implying that $\|u\|^2_{V_H} \ge k^2(2/\kappa^2+1)\|u\|^2_2$. □

Using Lemmas 3.5 and 3.6 we can now prove Theorem 3.1.

Proof. By Lemma 3.6 and under the assumption of the theorem that $k < \sqrt{2}/(H-f_-)$ we see that $b(.,.)$ is V_H-elliptic. Lemma 3.5 shows that $b(.,.)$ is bounded and hence by the Lax–Milgram lemma the existence of a unique solution u to (3.1) is assured. The estimate (3.2) also follows from the Lax–Milgram lemma. In the particular case that $\mathcal{G}(v) := -(g,v)$, for some $g \in L^2(S_H)$, we have further, by the Cauchy–Schwarz inequality and Lemma 3.4, that

$$\|\mathcal{G}\|_{V_H^*} = \sup_{v\in V_H}\frac{|(v,g)|}{\|v\|_{V_H}} \le \sup_{v\in V_H}\frac{\|v\|_2\|g\|_2}{\|v\|_{V_H}} \le \frac{H-f_-}{\sqrt{2}}\|g\|_2,$$

and (3.3) follows. □

4. Analysis of the variational problem at arbitrary frequency. The sesquilinear form $b(.,.)$ is not V_H-elliptic if the wave number k is large. In this section we shall establish, with no restriction on the wave number but some additional constraint on the domain, that the boundary value problem and the equivalent variational problem are uniquely solvable by using the generalized Lax–Milgram theory of Babuška. The domains D for which we will establish this result are those which, in addition to our assumption throughout that $U_{f_+} \subset D \subset U_{f_-}$, satisfy the condition that

$$x \in D \Rightarrow x + se_n \in D \quad \text{for all } s > 0, \tag{4.1}$$

where e_n denotes the unit vector in the direction x_n. Condition (4.1) is satisfied if Γ is the graph of a continuous function, but certainly does not require that this be the case. Nor does (4.1) impose any regularity on ∂D. Our main result in this section is the following.

THEOREM 4.1. *If* (4.1) *holds, then the variational problem* (3.1) *has a unique solution* $u \in V_H$ *for every* $\mathcal{G} \in V_H^*$ *and*

$$\|u\|_{V_H} \leq C\|\mathcal{G}\|_{V_H^*}, \tag{4.2}$$

where

$$C = 1 + \sqrt{2}\,\kappa\,(\kappa+1)^2.$$

In particular, the boundary value problem and the equivalent variational problem (2.23) *have exactly one solution, and the solution satisfies the bound*

$$k\|u\|_{V_H} \leq \frac{\kappa}{\sqrt{2}}\,(\kappa+1)^2\,\|g\|_2.$$

To apply the generalized Lax–Milgram theorem (e.g., [19, Theorem 2.15]) we need to show that b is bounded, which we have done in Lemma 3.5; to establish the inf-sup condition that

$$\beta := \inf_{0\neq u\in V_H}\sup_{0\neq v\in V_H}\frac{|b(u,v)|}{\|u\|_{V_H}\|v\|_{V_H}} > 0; \tag{4.3}$$

and to establish a "transposed" inf-sup condition. It follows easily from Corollary 3.3 that this transposed inf-sup condition follows automatically if (4.3) holds.

LEMMA 4.2. *If* (4.3) *holds, then for all nonzero* $v \in V_H$,

$$\sup_{0\neq u\in V_H}\frac{|b(u,v)|}{\|u\|_{V_H}} > 0.$$

Proof. If (4.3) holds and $v \in V_H$ is nonzero, then

$$\sup_{0\neq u\in V_H}\frac{|b(u,v)|}{\|u\|_{V_H}} = \sup_{0\neq u\in V_H}\frac{|b(\bar{v},u)|}{\|u\|_{V_H}} \geq \beta\|v\|_{V_H} > 0.$$

This proves the lemma. ∎

The following result follows from [19, Theorem 2.15] and Lemmas 3.5 and 4.2.

COROLLARY 4.3. *If* (4.3) *holds, then the variational problem* (3.1) *has exactly one solution* $u \in V_H$ *for all* $\mathcal{G} \in V_H^*$. *Moreover,*

$$\|u\|_{V_H} \leq \beta^{-1}\|\mathcal{G}\|_{V_H^*}.$$

To show (4.3) we will establish an a priori bound for solutions of (3.1), from which the inf-sup condition will follow by the following easily established lemma (see [19, Remark 2.20]).

LEMMA 4.4. *Suppose that there exists* $C > 0$ *such that for all* $u \in V_H$ *and* $\mathcal{G} \in V_H^*$ *satisfying* (3.1) *it holds that*

$$\|u\|_{V_H} \leq C\|\mathcal{G}\|_{V_H^*}. \tag{4.4}$$

Then the inf-sup condition (4.3) *holds with* $\beta \geq C^{-1}$.

The following lemma reduces the problem of establishing (4.4) to that of establishing an a priori bound for solutions of the special case (2.23).

LEMMA 4.5. *Suppose there exists $\tilde{C} > 0$ such that for all $u \in V_H$ and $g \in L^2(S_H)$ satisfying* (2.23) *it holds that*

$$\|u\|_{V_H} \leq k^{-1}\tilde{C}\,\|g\|_2. \tag{4.5}$$

Then, for all $u \in V_H$ and $\mathcal{G} \in V_H^*$ satisfying (3.1)*, the bound* (4.4) *holds with*

$$C \leq 1 + 2\,\tilde{C}.$$

Proof. Suppose $u \in V_H$ is a solution of

$$b(u,v) = \mathcal{G}(v), \quad v \in V_H, \tag{4.6}$$

where $\mathcal{G} \in V_H^*$. Let $b_0 : V_H \times V_H \to \mathbb{C}$ be defined by

$$b_0(u,v) = (\nabla u, \nabla v) + k^2(u,v) + \int_{\Gamma_H} \gamma_- \overline{v}\, T\gamma_- u\, ds, \quad u, v \in V_H.$$

It follows from Lemma 3.2 that b_0 is V_H-elliptic, in fact that

$$\Re\, b_0(v,v) \geq \|v\|^2_{V_H}, \quad v \in V_H.$$

Thus the problem of finding $u_0 \in V_H$ such that

$$b_0(u_0,v) = \mathcal{G}(v), \quad v \in V_H, \tag{4.7}$$

has a unique solution which satisfies

$$\|u_0\|_{V_H} \leq \|\mathcal{G}\|_{V_H^*}. \tag{4.8}$$

Furthermore, defining $w = u - u_0$ and using (4.6) and (4.7), we see that

$$b(w,v) = b(u,v) - b(u_0,v) = \mathcal{G}(v) - (\mathcal{G}(v) - 2k^2(u_0,v)) = 2k^2(u_0,v)$$

for all $v \in V_H$. Thus w satisfies (2.23) with $g = -2k^2 u_0$. It follows, using (4.8), (4.5), and Lemma 3.4, that

$$\|w\|_{V_H} \leq 2k\tilde{C}\|u_0\|_2 \leq 2\tilde{C}\|\mathcal{G}\|_{V_H^*}. \tag{4.9}$$

The bound (4.4), with $C \leq 1 + 2\,\tilde{C}$, follows from (4.8) and (4.9). ∎

Following these preliminary lemmas we turn now to establishing the a priori bound (4.5), at first just for the case when Γ is the graph of a smooth function. We recall that ν is the outward unit normal to S_H and $\nu_n = \nu \cdot e_n$ is the nth (vertical) component of ν.

LEMMA 4.6. *Suppose Γ is given by* (2.4) *with $f \in C^\infty(\mathbb{R}^{n-1})$. Let $H \geq f_+$, $g \in L^2(S_H)$ and suppose $w \in V_H$ satisfies*

$$b(w,\phi) = -(g,\phi) \quad \textit{for all } \phi \in V_H. \tag{4.10}$$

Then

$$\|w\|_{V_H} \leq k^{-1}\tilde{C}\|g\|_2,$$

where $\tilde{C} = \frac{\kappa}{\sqrt{2}}(\kappa+1)^2$.

Proof. The proof of this lemma is motivated by [21, 12], where a Rellich identity is used to prove estimates for solutions of the Helmholtz equation posed on bounded domains, by the proofs of the basic inequalities for rough surface scattering problems in [11, 29], and by the estimates derived for the diffraction grating problem in [16].

Let $r = |\tilde{x}|$. For $A \geq 1$ let $\phi_A \in C_0^\infty(\mathbb{R})$ be such that $0 \leq \phi_A \leq 1$, $\phi_A(r) = 1$ if $r \leq A$ and $\phi_A(r) = 0$ if $r \geq A+1$ and finally such that $\|\phi_A'\|_\infty \leq M$ for some fixed M independent of A.

Extending the definition of w to D by defining w in U_H by (2.9) with $F_H := \gamma_- w$, it follows from Theorem 2.3 that w satisfies the boundary value problem with g extended by zero from S_H to D. By standard local regularity results [17] it holds, since $g \in L^2(D)$, $w = 0$ on Γ, and the boundary is smooth, that $w \in H^2_{\mathrm{loc}}(D)$. Further, $w \in H^2(U_b \setminus U_c)$ for $c > b > f_+$ (though $w \in H^2(S_c)$ is not clear without some further constraint on the behavior of Γ at infinity). Moreover, by Lemma 2.2, w is given by the right-hand side of (2.9) in U_b for all $b > H$ if H is replaced in (2.9) by b and F_b denotes the restriction of w to Γ_b. Thus w satisfies the boundary value problem with H replaced by b for all $b > H$, and so, by Theorem 2.3,

$$\int_{S_b} (\nabla w \cdot \nabla \bar{v} - k^2 w \bar{v})\, dx = -\int_{\Gamma_b} \gamma_- \bar{v}\, T \gamma_- w\, ds - \int_{S_b} \bar{v} g\, dx \tag{4.11}$$

for all $b \geq H$.

In view of this regularity and since w satisfies the boundary value problem, we have, for all $a > H$,

$$\begin{aligned}
&2\Re \int_{S_a} \phi_A(r)(x_n - f_-) g \frac{\partial \overline{w}}{\partial x_n}\, dx \\
&= 2\Re \int_{S_a} \phi_A(r)(x_n - f_-)(\Delta w + k^2 w) \frac{\partial \overline{w}}{\partial x_n}\, dx \\
&= \int_{S_a} \Bigg\{ 2\Re \left\{ \nabla \cdot \left(\phi_A(r)(x_n - f_-) \frac{\partial \overline{w}}{\partial x_n} \nabla w \right) \right\} - 2\phi_A(r) \left| \frac{\partial w}{\partial x_n} \right|^2 \\
&\quad -(x_n - f_-)\phi_A(r) \frac{\partial |\nabla w|^2}{\partial x_n} \\
&\quad -2\phi_A'(r)(x_n - f_-) \frac{\tilde{x}}{|\tilde{x}|} \cdot \Re \left(\nabla_{\tilde{x}} w \frac{\partial \overline{w}}{\partial x_n} \right) + k^2 (x_n - f_-)\phi_A(r) \frac{\partial |w|^2}{\partial x_n} \Bigg\}\, dx.
\end{aligned}$$

Using the divergence theorem and integration by parts,

$$\begin{aligned}
&2\Re \int_{S_a} \phi_A(r)(x_n - f_-) g \frac{\partial \overline{w}}{\partial x_n}\, dx \\
&= (a - f_-) \int_{\Gamma_a} \phi_A(r) \left\{ \left| \frac{\partial w}{\partial x_n} \right|^2 - |\nabla_{\tilde{x}} w|^2 + k^2 |w|^2 \right\} ds \\
&\quad - \int_\Gamma (x_n - f_-)\phi_A(r) \left\{ \nu_n |\nabla w|^2 - 2\Re \left(\frac{\partial \overline{w}}{\partial x_n} \frac{\partial w}{\partial \nu} \right) \right\} ds \\
&\quad + \int_{S_a} \left\{ \phi_A(r) \left(|\nabla w|^2 - k^2 |w|^2 - 2 \left| \frac{\partial w}{\partial x_n} \right|^2 \right) - 2\phi_A'(r)(x_n - f_-) \Re \left(\frac{\partial \overline{w}}{\partial x_n} \frac{\partial w}{\partial r} \right) \right\} dx.
\end{aligned}$$

Using the fact that $w = 0$ on Γ so that $\nabla w = (\partial w/\partial\nu)\nu$ and

$$\frac{\partial w}{\partial x_n} = e_n \cdot \nabla w = e_n \cdot \nu \frac{\partial w}{\partial \nu} = \nu_n \frac{\partial w}{\partial \nu},$$

and rearranging terms we find that

$$\begin{aligned}
&-\int_\Gamma \phi_A(r)(x_n - f_-)\nu_n \left|\frac{\partial w}{\partial \nu}\right|^2 ds + 2\int_{S_a} \phi_A(r)\left|\frac{\partial w}{\partial x_n}\right|^2 dx \\
&= (a - f_-)\int_{\Gamma_a} \phi_A(r)\left\{\left|\frac{\partial w}{\partial x_n}\right|^2 - |\nabla_{\tilde{x}} w|^2 + k^2|w|^2\right\} ds \\
&\quad + \int_{S_a}\left\{\phi_A(r)\left(|\nabla w|^2 - k^2|w|^2\right) - 2\phi_A'(r)(x_n - f_-)\Re\left(\frac{\partial \overline{w}}{\partial x_n}\frac{\partial w}{\partial r}\right)\right\} dx \\
&\quad -2\Re\int_{S_a} \phi_A(r)(x_n - f_-) g \frac{\partial \overline{w}}{\partial x_n}\, dx.
\end{aligned} \tag{4.12}$$

We now wish to let $A \to \infty$. The only problem is the term involving ϕ_A' which we estimate as follows. Let $S_a^b = \{x \in S_a \ : \ |\tilde{x}| < b\}$ for $b \geq 1$. Then

$$\left|\int_{S_a}\left\{2\phi_A'(r)(x_n - f_-)\Re\left(\frac{\partial \overline{w}}{\partial x_n}\frac{\partial w}{\partial r}\right)\right\} dx\right| \leq 2M(a - f_-)\int_{S_a^{A+1}\setminus \overline{S}_a^A} |\nabla w|^2\, dx \to 0$$

as $A \to \infty$, where the convergence follows from the fact that $w \in H^1(S_H)$. In addition since $w \in H^2(U_b \setminus U_c)$, for $c > a > b > f_+$, $\nabla w|_{\Gamma_H} \in (H^{1/2}(\Gamma_H))^n$. Thus, taking the limit as $A \to \infty$ in (4.12), and applying the Lebesgue dominated convergence and monotone convergence theorems, we see that

$$\begin{aligned}
&-\int_\Gamma (x_n - f_-)\nu_n \left|\frac{\partial w}{\partial \nu}\right|^2 ds + 2\int_{S_a}\left|\frac{\partial w}{\partial x_n}\right|^2 dx \\
&= (a - f_-)\int_{\Gamma_a}\left\{\left|\frac{\partial w}{\partial x_n}\right|^2 - |\nabla_{\tilde{x}} w|^2 + k^2|w|^2\right\} ds \\
&\quad + \int_{S_a}\left(|\nabla w|^2 - k^2|w|^2 - 2\Re\left((x_n - f_-) g \frac{\partial \overline{w}}{\partial x_n}\right)\right) dx.
\end{aligned} \tag{4.13}$$

Now, since w satisfies the boundary value problem, including the radiation condition (2.9), applying Lemma 2.2 it follows that

$$\begin{aligned}
\int_{\Gamma_a}\left\{\left|\frac{\partial w}{\partial x_n}\right|^2 - |\nabla_{\tilde{x}} w|^2 + k^2|w|^2\right\} ds &\leq 2k\Im\int_{\Gamma_a} \overline{w}\frac{\partial w}{\partial x_n}\, ds \\
&= -2k\Im\int_{\Gamma_a} \gamma_-\overline{w} T\gamma_- w\, ds
\end{aligned} \tag{4.14}$$

on applying the Plancherel identity (2.8), noting (2.15) and (2.16). Further, setting $v = w$ in (4.11), we get

$$\int_{S_b}\left(|\nabla w|^2 - k^2|w|^2\right) dx = -\int_{\Gamma_b} \gamma_-\overline{w} T\gamma_- w\, ds - \int_{S_b} g\overline{w}\, dx \tag{4.15}$$

for $b \geq H$, so that, by Lemma 3.2,

$$\int_{S_b} [|\nabla w|^2 - k^2|w|^2]\, dx \leq -\Re \int_{S_b} g\overline{w}\, dx \tag{4.16}$$

and

$$-2k\Im \int_{\Gamma_b} \gamma_- \overline{w} T \gamma_- w\, ds = 2k\Im \int_{S_b} g\overline{w}\, dx. \tag{4.17}$$

Using (4.17) in (4.14) and then using the resulting equation and (4.16) in (4.13) and noting that $\operatorname{supp} g \subset \overline{S_H}$, we get that

$$\begin{aligned} -\int_\Gamma (x_n - f_-)\nu_n \left|\frac{\partial w}{\partial \nu}\right|^2 ds + 2\int_{S_H} \left|\frac{\partial w}{\partial x_n}\right|^2 dx \leq 2(a - f_-)k\Im \int_{S_H} g\bar{w}\, dx \\ -\Re \int_{S_H} \left[g\bar{w} + 2(x_n - f_-)g\frac{\partial \bar{w}}{\partial x_n}\right] dx. \end{aligned}$$

Since this equation holds for all $a > H$ and $\nu_n < 0$ on Γ, it follows by the Cauchy–Schwarz inequality that

$$2\left\|\frac{\partial w}{\partial x_n}\right\|_2^2 \leq \left(2\kappa\|w\|_2 + \|w\|_2 + 2(H - f_-)\left\|\frac{\partial w}{\partial x_n}\right\|_2\right)\|g\|_2.$$

Now using Lemma 3.4 to estimate $\|w\|_2$ we obtain

$$\left\|\frac{\partial w}{\partial x_n}\right\|_2 \leq (H - f_-)\left(\frac{1}{\sqrt{2}}\kappa + \frac{1}{2\sqrt{2}} + 1\right)\|g\|_2 \tag{4.18}$$

and use of Lemma 3.4 again shows that

$$\|w\|_2 \leq (H - f_-)^2\left(\frac{1}{2}\kappa + \frac{1}{4} + \frac{1}{\sqrt{2}}\right)\|g\|_2.$$

Using the above inequality in (4.16) shows that

$$\begin{aligned} \|w\|_{V_H}^2 &\leq 2k^2\|w\|_2^2 + \|g\|_2\|w\|_2 \\ &\leq \frac{(H - f_-)^2}{4}\left(\frac{\kappa^2}{2}(2\kappa + 1 + 2\sqrt{2})^2 + 2\kappa + 1 + 2\sqrt{2}\right)\|g\|_2^2. \end{aligned}$$

Thus, for $\kappa \geq 1$,

$$\|w\|_{V_H}^2 \leq \frac{(H - f_-)^2}{2}(\kappa + 1)^4\|g\|_2^2.$$

The same bound holds for $\kappa < 1$ by Theorem 3.1. ∎

REMARK 4.7. *The above argument works under milder assumptions on the boundary* Γ, *in particular that* Γ *is the graph of a function* $f \in C^2(\mathbb{R}^{n-1})$, *so that* Γ *is of class* C^2. *This assumption is enough* [17] *to deduce the necessary local regularity result that* $w \in H^2_{\mathrm{loc}}(D)$.

Combining Lemmas 4.6, 4.5, and 4.4 with Corollary 4.3, we have the following result.

LEMMA 4.8. *If Γ satisfies the conditions of Lemma* 4.6*, then the variational problem* (3.1) *has a unique solution $u \in V_H$ for every $\mathcal{G} \in V_H^*$ and the solution satisfies the estimate* (4.2).

REMARK 4.9. *The above result, combined with Lemma* 4.4*, implies that β, the* inf-sup *constant for $b(\cdot,\cdot)$, satisfies $\beta^{-1} \leq C = O(k^3)$ as $k \to \infty$. This high power of the wave number is, we suspect, not optimal. For an interior problem in a smooth starlike and bounded domain in $\mathbb{R}^2$ or $\mathbb{R}^3$ with impedance boundary data it is known that the constant in the corresponding bound satisfies the estimate $C = O(k)$ (for example, this can be proved by combining estimate* (2) *of Theorem* 1 *of* [12] *with the argument of Lemma* 4.5*, involving a function corresponding to u_0). For a somewhat analogous one-dimensional problem the* inf-sup *constant is also $O(k)$ as $k \to \infty$ (Theorem* 4.2 *of* [19]).

We proceed now to establish that Lemmas 4.6 and 4.8 hold for much more general boundaries, namely those satisfying (4.1). To establish this we first prove the following technical lemma.

LEMMA 4.10. *If* (4.1) *holds, then, for every $\phi \in C_0^\infty(D)$, there exists $f \in C^\infty(\mathbb{R}^{n-1})$ such that*

$$\operatorname{supp}\phi \subset D' := \{x \in \mathbb{R}^n : x_n > f(\tilde{x}),\, \tilde{x} \in \mathbb{R}^{n-1}\}$$

and $U_{f_+} \subset D' \subset D$.

Proof. Let $S := \operatorname{supp}\phi \setminus U_{f_+}$. Then either $S = \emptyset$, in which case $f(\tilde{x}) \equiv f_+$ has the properties claimed, or $S \neq \emptyset$.

Thus, suppose $S \neq \emptyset$ and let $\delta := \operatorname{dist}(S, \partial D)/2$. Then $\delta > 0$ and, defining $G := \{x + se_n : x \in S, s \geq 0\}$, $\operatorname{dist}(G, \partial D) = \operatorname{dist}(S, \partial D) = 2\delta$. Let $G_\delta := \{x \in \mathbb{R}^n : \operatorname{dist}(x, G) < \delta\}$ and let A and A_δ denote the projections of G and G_δ, respectively, onto the $Ox_1 \cdots x_{n-1}$ plane.

Let $N \in \mathbb{N}$ and $S_j \subset \mathbb{R}^{n-1}$, $j = 1, \ldots, N$, be such that each S_j is measurable and nonempty, $S_j \cap S_m = \emptyset$ for $j \neq m$,

$$A_\delta \subset \bigcup_{j=1}^{N} S_j,$$

and $\operatorname{diam}(S_j) \leq \delta/2$, $j = 1, \ldots, N$. For $j = 1, \ldots, N$ choose $\tilde{x}_j \in S_j$ and let

$$f_j := \inf\{x_n \in \mathbb{R} : x = (\tilde{x}_j, x_n) \in G_\delta \cup \Gamma_{f_+}\}.$$

Then $f_- \leq f_j \leq f_+$, $j = 1, \ldots, N$. Define $\tilde{f} : \mathbb{R}^{n-1} \to \mathbb{R}$ by

$$\tilde{f}(\tilde{x}) := \begin{cases} f_j & \text{if } \tilde{x} \in S_j,\ j = 1, \ldots, N, \\ f_+ & \text{otherwise.} \end{cases}$$

Then $\tilde{f} \in L^\infty(\mathbb{R}^{n-1})$; in fact $\tilde{f}$ is a simple function and $f_- \leq \tilde{f}(\tilde{x}) \leq f_+$, $\tilde{x} \in \mathbb{R}^{n-1}$. Choose ϵ with $0 < \epsilon < \delta/2$ and let $J \in C_0^\infty(\mathbb{R}^{n-1})$ be such that $J \geq 0$, $J(\tilde{x}) = 0$ if $|\tilde{x}| \geq \epsilon$, and $\int_{\mathbb{R}^{n-1}} J(\tilde{x})\, d\tilde{x} = 1$. Define $f \in C^\infty(\mathbb{R}^{n-1})$ by

$$f(\tilde{x}) := \int_{\mathbb{R}^{n-1}} J(\tilde{x} - \tilde{y}) \tilde{f}(\tilde{y})\, d\tilde{y}, \quad \tilde{x} \in \mathbb{R}^{n-1},$$

and let D' be defined as in the statement of the lemma. Then f and D' have the properties claimed.

To see that this is true note first that

$$\min_{|\tilde{y}-\tilde{x}|<\epsilon} |\tilde{f}(\tilde{y})| \le f(\tilde{x}) \le \max_{|\tilde{y}-\tilde{x}|<\epsilon} |\tilde{f}(\tilde{y})|, \quad \tilde{x} \in \mathbb{R}^{n-1}, \tag{4.19}$$

so that $U_{f_+} \subset D'$. If $\tilde{x} \in A$, then $|\tilde{y}-\tilde{x}| < \epsilon$ implies that $\tilde{y} \in A_\delta$ and so (4.19) implies that

$$f(\tilde{x}) \le \max_{j=1,\dots,N,\ |\tilde{x}_j-\tilde{x}|<\epsilon+\delta/2} f_j$$

so that $f(\tilde{x}) \le f_m$ for some m for which $|\tilde{x}_m - \tilde{x}| < \epsilon + \delta/2$. Now let $x = (\tilde{x}, f_m)$, $y = (\tilde{x}_m, f_m)$. Then $|x-y| = |\tilde{x} - \tilde{x}_m| < \epsilon + \delta/2$ and $\mathrm{dist}(y, G) = \delta$ so that

$$\mathrm{dist}(x, G) \ge \mathrm{dist}(y, G) - |x-y| \ge \delta - (\epsilon + \delta/2) > 0.$$

Thus $x \notin G$ and so $(\tilde{x}, f(\tilde{x})) \notin G$. Thus $S \subset G \subset D'$ and so $\mathrm{supp}\,\phi \subset U_{f_+} \cup S \subset D'$.

Arguing similarly, for all $\tilde{x} \in \mathbb{R}^{n-1}$, either $f(\tilde{x}) = f_+$, in which case $(\tilde{x}, x_n) \in D$ for $x_n > f(\tilde{x})$, or $f(\tilde{x}) \ge f_m$ for some m for which $|\tilde{x}_m - \tilde{x}| < \epsilon + \delta/2$. In this latter case, defining $x = (\tilde{x}, f_m)$ and $y = (\tilde{x}_m, f_m)$, it holds that

$$\mathrm{dist}(x, G) \le \mathrm{dist}(y, G) + |x-y| \le \delta + \epsilon + \delta/2 < 2\delta$$

so that $x \in D$ and hence $(\tilde{x}, f(\tilde{x})) \in D$. Thus, for all $\tilde{x} \in \mathbb{R}^{n-1}$, $(\tilde{x}, x_n) \in D$ for $x_n > f(\tilde{x})$, i.e., $D' \subset D$. ∎

With this preliminary lemma we can proceed to show that Lemma 4.6 holds whenever (4.1) holds.

LEMMA 4.11. *Suppose* (4.1) *holds,* $H \ge f_+$, $g \in L^2(S_H)$, *and* $w \in V_H$ *satisfies*

$$b(w, \phi) = -(g, \phi) \quad \textit{for all } \phi \in V_H. \tag{4.20}$$

Then

$$\|w\|_{V_H} \le k^{-1}\tilde{C}\|g\|_2,$$

where $\tilde{C} = \frac{\kappa}{\sqrt{2}}(\kappa+1)^2$.

Proof. Let $\tilde{V} := \{\phi|_{S_H} : \phi \in C_0^\infty(D)\}$. Then $\tilde{V}$ is dense in V_H. Suppose w satisfies (4.20) and choose a sequence $(w_m) \subset \tilde{V}$ such that $\|w_m - w\|_{V_H} \to 0$ as $m \to \infty$. Then $w_m = \phi_m|_{S_H}$, with $\phi_m \in C_0^\infty(D)$, and, by Lemma 4.10, there exists $f_m \in C^\infty(\mathbb{R}^{n-1})$ such that $\mathrm{supp}\,\phi_m \subset D_m$ and $U_{f_+} \subset D_m \subset D$, where $D_m := \{x \in \mathbb{R}^n : x_n > f_m(\tilde{x}),\ \tilde{x} \in \mathbb{R}^{n-1}\}$. Let $V_H^{(m)}$ and b_m denote the space and sesquilinear form corresponding to the domain D_m. That is, where $S_H^{(m)} := D_m \setminus \overline{U_H}$, $V_H^{(m)}$ is defined by $V_H^{(m)} := \{\phi|_{S_H^{(m)}} : \phi \in H_0^1(D_m)\}$ and b_m is given by (2.22) with S_H and V_H replaced by $S_H^{(m)}$ and $V_H^{(m)}$, respectively. Then $S_H^{(m)} \subset S_H$ and, if $v_m \in V_H^{(m)}$ and v denotes v_m extended by zero from $S_H^{(m)}$ to S_H, it holds that $v \in V_H$. Via this extension by zero, we can regard $V_H^{(m)}$ as a subspace of V_H and regard w_m as an element of $V_H^{(m)}$.

For all $v \in V_H^{(m)} \subset V_H$, we have

$$b_m(w_m, v) = b(w_m, v) = -(g, v) - b(w - w_m, v).$$

By Lemma 4.8, there exist unique w_m', $w_m'' \in V_H^{(m)}$ such that

$$b_m(w_m', v) = -(g, v), \quad v \in V_H^{(m)},$$

and

$$b_m(w_m'', v) = -b(w - w_m, v), \quad v \in V_H^{(m)}.$$

Clearly $w_m = w_m' + w_m''$ and, by Lemma 4.6,

$$\|w_m'\|_{V_H^{(m)}} \le k^{-1}\tilde{C}\|g\|_2$$

while, by Lemmas 4.8 and 3.4,

$$\|w_m''\|_{V_H^{(m)}} \le 2C\|w - w_m\|_{V_H}.$$

Thus

$$\|w\|_{V_H} = \lim_{m\to\infty} \|w_m\|_{V_H^{(m)}} \le k^{-1}\tilde{C}\|g\|_2. \qquad \square$$

Theorem 4.1 now follows by combining Lemmas 4.11, 4.5, and 4.4 with Corollary 4.3.

Acknowledgments. The authors are grateful to Andreas Kirsch and Martin Thomas for helpful comments on an earlier version of this paper.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces,* Academic Press, New York, 1975.

[2] H. D. Alber, *A quasi-periodic boundary value problem for the Laplacian and the continuation of its resolvent,* Proc. Roy. Soc. Edinburgh, 82A (1979), pp. 251–272.

[3] T. Arens, S. N. Chandler-Wilde, and K. Haseloh, *Solvability and spectral properties of integral equations on the real line: I. Weighted spaces of continuous functions,* J. Math. Anal. Appl., 272 (2002), pp. 276–302.

[4] T. Arens, S. N. Chandler-Wilde, and K. Haseloh, *Solvability and spectral properties of integral equations on the real line: II. L^p spaces and applications,* J. Integral Equations Appl., 15 (2003), pp. 1–35.

[5] T. Arens and T. Hohage, *On radiation conditions for rough surface scattering problems*, IMA J. Appl. Math., to appear; also available online from http://imamat.oupjournals.org/cgi/content/abstract/hxh065v1.

[6] G. Bao, D. Dobson, and J. Cox, *Mathematical studies in rigorous grating theory,* J. Opt. Soc. Amer. A, 12 (1995), pp. 1029–1042.

[7] M. Cadilhac, *Some mathematical aspects of the grating theory,* in Electromagnetic Theory of Gratings, R. Petit, ed., Springer, Berlin, 1980, pp. 53–62.

[8] S. N. Chandler-Wilde, E. Heinemeyer, and R. Potthast, *Acoustic scattering by mildly rough unbounded surfaces in three dimensions,* submitted.

[9] S. N. Chandler-Wilde, C. R. Ross, and B. Zhang, *Scattering by rough surfaces,* in Proceedings of the Fourth International Conference on Mathematical and Numerical Aspects of Wave Propagation, J. DeSanto, ed., SIAM, Philadelphia, 1998, pp. 164–168.

[10] S. N. Chandler-Wilde, C. R. Ross, and B. Zhang, *Scattering by infinite one-dimensional rough surfaces,* Proc. R. Soc. Lond. Ser. A, 455 (1999), pp. 3767–3787.

[11] S. N. Chandler-Wilde and B. Zhang, *A uniqueness result for scattering by infinite rough surfaces,* SIAM J. Appl. Math., 58 (1998), pp. 1774–1790.

[12] P. Cummings and X. Feng, *Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations*, Math. Models Methods Appl. Sci., to appear.

[13] J. A. De Santo, *Scattering by rough surfaces,* in Scattering: Scattering and Inverse Scattering in Pure and Applied Science, R. Pike and P. Sabatier, eds., Academic Press, New York, 2002, pp. 15–36.

[14] D. Eidus and A. Vinnik, *On radiation conditions for domains with infinite boundaries,* Soviet Math. Dokl., 15 (1974), pp. 12–15.

[15] J. Elschner and G. Schmidt, *Diffraction in periodic structures and optimal design of binary gratings I: Direct problems and gradient formulas,* Math. Methods Appl. Sci., 81 (1998), pp. 1307–1328.

[16] J. Elschner and M. Yamamoto, *An inverse problem in periodic diffractive optics: Reconstruction of Lipschitz grating profiles,* Appl. Anal., 81 (2002), pp. 1307–1328.

[17] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order,* 2nd ed., Springer, Berlin, 1983.

[18] V. Y. Gotlib, *Solutions of the Helmholtz equation, concentrated near a plane periodic boundary,* J. Math. Sci., 102 (2000), pp. 4188–4194.

[19] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering,* Appl. Math. Sci. 132, Springer, Berlin, 1998.

[20] A. Kirsch, *Diffraction by periodic structures*, in Inverse Problems in Mathematical Physics, L. Pävärinta and E. Somersalo, eds., Springer, Berlin, 1993, pp. 87–102.

[21] J. Melenk, *On Generalized Finite Element Methods,* Ph.D. thesis, University of Maryland, College Park, MD, 1995.

[22] V. Minskii, *Sommerfeld radiation conditions for second-order differential operator in a domain with infinite border,* J. Differential Equations, 48 (1983), pp. 157–176.

[23] J. A. Ogilvy, *Theory of Wave Scattering from Random Rough Surfaces,* Adam Hilger, Bristol, 1991.

[24] M. Saillard and A. Sentenac, *Rigorous solutions for electromagnetic scattering from rough surfaces,* Waves Random Media, 11 (2001), pp. R103–R137.

[25] V. Vogelsang, *Das Ausstrahlungsproblem fur elliptische Differentialgleichungen in Gebieten mit unbeschranktem Rand,* Math. Z., 144 (1975), pp. 101–124.

[26] A. G. Voronovich, *Wave Scattering from Rough Surfaces,* Springer, Berlin, 1994.

[27] K. Warnick and W. C. Chew, *Numerical simulation methods for rough surface scattering,* Waves Random Media, 11 (2001), pp. R1–R30.

[28] B. Zhang, *An integral equation method for scattering by an infinite Lipschitz surface,* in Proc. JEE '02, European Symp. Numerical Methods in Electromagnetics, Toulouse, 2002, pp. 31–35.

[29] B. Zhang and S. N. Chandler-Wilde, *Acoustic scattering by an inhomogeneous layer on a rigid plate,* SIAM J. Appl. Math., 58 (1998), pp. 1931–1950.

[30] B. Zhang and S. N. Chandler-Wilde, *Integral equation methods for scattering by infinite rough surfaces,* Math. Methods Appl. Sci., 26 (2003), pp. 463–488.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 619–650

WEAK SEQUENTIAL STABILITY OF THE SET OF ADMISSIBLE VARIATIONAL SOLUTIONS TO THE NAVIER–STOKES–FOURIER SYSTEM*

EDUARD FEIREISL† AND ANTONÍN NOVOTNÝ‡

Abstract. We define and investigate variational (weak) solutions of the initial-boundary-value problem for the Navier–Stokes–Fourier system with the general pressure law. We prove that the set of these solutions is weakly sequentially stable.

Key words. Navier–Stokes–Fourier system, variational weak solutions, weak compactness

AMS subject classifications. 35Q30, 76N10

DOI. 10.1137/04061458X

1. Introduction. The state of a fluid in classical continuum mechanics is characterized through the value of three *macroscopic* quantities: the *density* $\varrho = \varrho(t,x)$, the *velocity* field $\mathbf{u} = \mathbf{u}(t,x)$, and the absolute *temperature* $\vartheta = \vartheta(t,x)$, where $t \in (0,T) \subset R$ is the time, and $x \in \Omega \subset R^3$ denotes the coordinate in the ambient Euclidean space. The basic axioms of physics, written in the spatial (Eulerian) reference system, assert the following:

(i) conservation of mass

$$\partial_t \varrho + \mathrm{div}_x(\varrho \mathbf{u}) = 0; \tag{1.1}$$

(ii) balance of (linear) momentum

$$\partial_t(\varrho \mathbf{u}) + \mathrm{div}_x(\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla_x p = \mathrm{div}_x\, \mathbb{S} + \varrho \mathbf{f}; \tag{1.2}$$

(iii) balance of entropy

$$\partial_t(\varrho s) + \mathrm{div}_x(\varrho s \mathbf{u}) + \mathrm{div}_x\left(\frac{\mathbf{q}}{\vartheta}\right) = \sigma. \tag{1.3}$$

In accordance with the commonly accepted mathematical definition of a fluid, the total stress is expressed by Stokes' law $\mathbb{S} - p\mathbb{I}$, where $\mathbb{S}$ denotes the *viscous stress* tensor, and p is the *pressure*. The symbol s stands for the specific *entropy*, $\mathbf{q}$ is the *heat flux*, and σ denotes the *entropy production* rate (see, for example, Chapter 1 in [21]). The quantity $\mathbf{f}$ represents the external *body force* density—a given function of the time t and the position x.

One of the fundamental assumptions of the present theory is the requirement that the system is both thermally and mechanically *isolated*. One of the possible ways to meet this stipulation is to postulate

(iv) no-flux boundary conditions

$$\mathbf{q} \cdot \mathbf{n}|_{\partial\Omega} = 0, \ \text{ where } \mathbf{n} \text{ denotes the outer normal vector,} \tag{1.4}$$

*Received by the editors September 7, 2004; accepted for publication (in revised form) April 7, 2005; published electronically October 27, 2005. This work was supported by grant 201/02/0854 of GA ČR and by the Project of French-Czech Collaboration in Applied Mathematics of USTV in France.

http://www.siam.org/journals/sima/37-2/61458.html

†Mathematical Institute AS CR, Žitná 25, 115 67 Praha 1, Czech Republic (feireisl@math.cas.cz).

‡Université du Sud Toulon-Var, BP 20132, 839 57 La Garde, France (novotny@univ-tln.fr).

together with

(v) no-slip boundary conditions

$$\mathbf{u}|_{\partial\Omega} = 0. \tag{1.5}$$

Consequently, the *total energy*

$$E \equiv \frac{1}{2}\varrho|\mathbf{u}|^2 + \varrho e, \text{ with } e \text{ being the specific internal energy,}$$

has to be a conserved quantity; more precisely, we have

(vi) total energy balance

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega E\,\mathrm{d}x = \int_\Omega \varrho\mathbf{f}\cdot\mathbf{u}\,\mathrm{d}x. \tag{1.6}$$

The *field equations* (1.1)–(1.3), together with the boundary conditions (1.4) and (1.5), supplemented with the total energy balance (1.6), form a suitable platform for the mathematical theory based on the concept of variational (weak) solutions. This concept will be developed in section 2. Note that we have deliberately opted for the "entropy form" (1.3) of the energy equation. Moreover, the total energy balance (1.6), which follows from (1.1) to (1.5) provided s and σ are related to the state variables through the basic thermodynamic equations and all fields are *smooth*, has been appended to the system (cf. section 2.1). Note that for weak solutions this equality may provide new information. The reason for this particular choice of the governing equations will become clear in section 2.

Of course, the general physical laws introduced above do not suffice to determine the motion of a specific fluid subject to the given loading. Before a concrete mathematical problem can be formulated, it is necessary to specify the *constitutive equations* relating the pressure p, the viscous stress $\mathbb{S}$, the heat flux $\mathbf{q}$, the entropy density s, the entropy production rate σ, and the internal energy density e to the state variables ϱ, $\mathbf{u}$, and ϑ. Here again, the reader will have noticed, there is freedom of choice of the basic state variables, yielding a variety of "equivalent" formulations provided the motion is smooth.

The constitutive equations, reflecting the *microscopic* structure of a given fluid, may be regarded as averages or expected values of molecular actions and as such they can be evaluated by the methods of statistical mechanics. When doing so, one often tacitly assumes the changes on the microscopic level to be so fast that the system at any "macroscopic" time t can be considered as being in thermodynamic equilibrium. In particular, the constitutive equations coincide with those obtained by the methods of thermostatics (cf. [20]).

The most important constitutive relation in fluid dynamics—the *state equation*—relates the pressure p to the values of the density ϱ and the absolute temperature ϑ. The best-known example is provided by Boyle's law for the perfect gas:

$$p = R\varrho\vartheta \text{ with a material constant } R > 0. \tag{1.7}$$

In light of arguments of modern statistical mechanics, however, relation (1.7) cannot be accepted without criticism even in the case of a "real perfect" gas, which means a large ensemble of identical particles with no interactive forces. Indeed there is a threshold temperature $\Theta_c = \Theta_c(\varrho)$ below which the gas exhibits *degeneration phenomena* due to quantum effects (see section 2.2 and Chapter 3 in [20]). Accordingly, for gases in low-temperature (or high-density) regimes, the dominant component of the macroscopic pressure results from the phenomenon of *pressure ionization*, with

$$p = a\varrho^{\frac{5}{3}} \text{ for } \vartheta < \Theta_c(\varrho) \tag{1.8}$$

the pressure of a completely degenerate electron gas (see, for instance, Chapter 15 in [14]).

These observations can be graphically visualized in Figure 1.1.

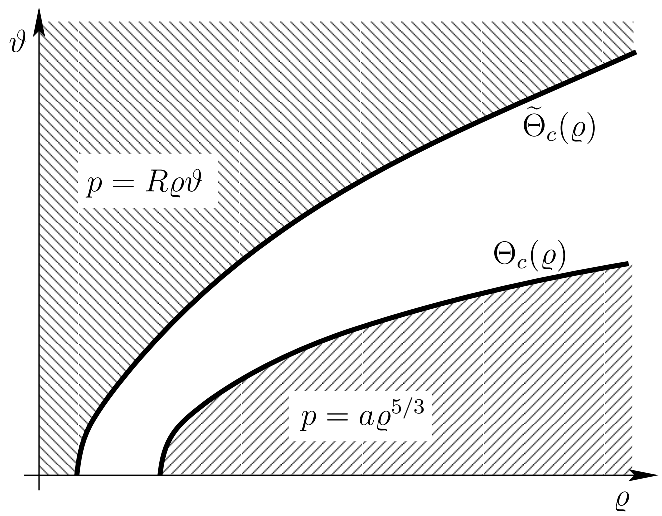


FIG. 1.1. $p = p(\varrho, \vartheta)$.

Motivated by the previous example, the main objective of this paper is to identify the class of physically relevant constitutive equations for which the set of all suitable variational solutions of problem (1.1)–(1.6) enjoys the property of *weak sequential stability.* That is to say that any sequence of (physically) admissible solutions bounded by the available a priori estimates possesses an accummulation point—another admissible solution of problem (1.1)–(1.6). Such a property was established in [16] under the main hypothesis that the pressure $p = p(\varrho, \vartheta)$ is an *affine* function of the temperature, which means

$$p(\varrho, \vartheta) = p_e(\varrho) + \vartheta p_\vartheta(\varrho)$$

for suitable functions p_e, p_ϑ. In particular, the internal energy density e can be written in the form

$$e(\varrho, \vartheta) = P_e(\varrho) + Q(\vartheta), \text{ where } P_e(\varrho) \equiv \int_1^\varrho \frac{p_e(z)}{z^2}\,\mathrm{d}z, \text{ and } Q \text{ is a given function.}$$

The possibility of separating the "elastic" part P_e from the thermal component Q gives rise to a family of rescaled (renormalized) energy equations yielding (i) suitable a priori estimates on the temperature ϑ and (ii) compatibility of the corresponding Navier–Stokes–Fourier system with the so-called biting limit procedure applied to the thermal energy equation (see Chapter 6 in [16]). Such an approach is of no use in the case of the more realistic and physically relevant pressure–density–temperature state equations considered in this paper.

Leaving aside the vast amount of literature devoted to problems in the one-dimensional space geometry (see, for instance, the monograph by Antontsev, Kazhikhov, and Monakhov [1]) as well as the "small data" problems and the problems of existence of "local in time smooth solutions" (cf. the papers by Hoff [23], Matsumura and Nishida [26], [25], among others), the most relevant references to be quoted here are represented by the monograph by Lions [24] together with the results of Vaigant and Kazhikhov [38]. In both cases, the "large data" problems are addressed concerning the so-called barotropic flows for which p can be taken as a function of ϱ only. In particular, the fundamental property of "weak continuity" of the *effective viscous flux* established in [24] (cf. also [22], [33]) will play a significant role in the analysis of the present problem (see section 5).

After an introductory section 2 defining the class of admissible variational solutions to problem (1.1)–(1.6), the paper is divided into three major parts as follows.

- Section 3 presents a general class of constitutive equations, based upon physical principles, for which the main result—the *weak sequential stability* of the solution set of problem (1.1)–(1.6)—is established (see Theorem 3.1).
- Section 4 concerns the available estimates on the state variables ϱ, $\mathbf{u}$, and ϑ resulting from boundedness of the total energy and entropy. The role of these bounds is to prevent possible *concentration phenomena* that may develop in a sequence of admissible solutions. As a by-product, we deduce the weak sequential stability of the convective terms $\varrho\mathbf{u}$, $\varrho\mathbf{u} \otimes \mathbf{u}$, and $\varrho s\mathbf{u}$.
- The central issue of compatibility of the (nonlinear) constitutive equations with weak convergence is treated in section 5. In particular, we establish a weak "continuity" property of the effective viscous flux for a general linearly viscous fluid with temperature-dependent shear and bulk viscosity coefficients. Moreover, a new class of *renormalized* solutions to the continuity equation is introduced in order to show strong L^1-compactness of the temperature field. Finally, the technique of [16] based on the concept of "oscillations defect measure" is adopted in order to show strong L^1-compactness of the density.

2. Admissible variational solutions. The purpose of this section is to introduce a suitable concept of admissible solutions to problem (1.1)–(1.6) to be dealt with in what follows. Unfortunately, the most natural choice of *smooth solutions* lies out of reach of the presently available mathematical tools. More specifically, the set of smooth solutions is not (known to be) closed with respect to a priori bounds resulting mostly from the rather poor energy or entropy estimates.

Instead our analysis is based on the concept of variational (weak) solutions, where the original system of partial differential equations is replaced by a family of *integral identities* to be satisfied when the equations are "tested" (multiplied) on a suitable smooth function. Although such an approach seems to be much closer to the underlying physical principles, the fundamental question whether or not it yields a definite theory in the sense of existence, uniqueness, and stability of solutions of meaningful initial-value problems is far from being settled.

Nonetheless, our aim is to develop a mathematical theory of (1.1)–(1.6) which conforms with the following stipulations, the rigorous verification of which is the objective of this paper and/or of future research.

- *Compatibility.* Any smooth solution of problem (1.1)–(1.6) supplemented with a suitable set of constitutive equations is an admissible solution in the sense specified later in this section. Conversely, any admissible solution which is smooth represents a classical solution of the problem.

- *Physical admissibility.* The class of admissible solutions meets all restrictions imposed by the underlying physical principles, in particular, the second law of thermodynamics.
- *Existence.* The results obtained in this paper combined with the approximation scheme developed in Chapter 7 in [16] should yield rigorous existence results (in the class of weak solutions) for the corresponding initial-boundary-value problem without any essential restriction on the size of initial data.
- *Equilibrium states (Prigogine's principle).* Minimization of the entropy production rate σ over the set of all admissible states of the system yields an equilibrium solution (cf. [30]).
- *Large time relaxation.* Any admissible solution resulting from the action of a *conservative* body force $\mathbf{f} = \nabla_x F$ tends to an equilibrium for large values of time.

2.1. General thermodynamics relations. In accordance with the common strategy delineated in section 1, the thermodynamic functions p, e, and s will be expressed in terms of the instantaneous values of the scalar state variables ϱ and ϑ. Consequently, by virtue of the basic principles of thermodynamics, the entropy density s is determined, up to an additive constant, through

$$\frac{\partial s}{\partial \varrho} = \frac{1}{\vartheta}\Big(\frac{\partial e}{\partial \varrho} - \frac{p}{\varrho^2}\Big), \quad \frac{\partial s}{\partial \vartheta} = \frac{1}{\vartheta}\frac{\partial e}{\partial \vartheta} \tag{2.1}$$

imposing Maxwell's relation (a compatibility condition) on e and p in the form

$$\frac{\partial e}{\partial \varrho} = \frac{1}{\varrho^2}\Big(p - \vartheta\frac{\partial p}{\partial \vartheta}\Big). \tag{2.2}$$

By the same token, the entropy production rate σ is a nonnegative quantity given by

$$\sigma = \mathbb{S} : \left(\frac{\nabla_x \mathbf{u}}{\vartheta}\right) - \mathbf{q} \cdot \left(\frac{\nabla_x \vartheta}{\vartheta^2}\right) \geq 0 \tag{2.3}$$

provided the flow is *smooth* (see, for instance, Chapter 1 in [21]).

Here, the assumption of smoothness is crucial since it is well known that discontinuous solutions (shock waves) of the inviscid system do increase entropy even if $\mathbb{S} = \mathbf{q} = 0$ (see, for example, the classical text by Smoller [34]). As a matter of fact, a kind of "mechanical energy defect" cannot be legitimately excluded even in the case of a viscous incompressible flow described by the classical Navier–Stokes system in three space dimensions (see Duchon and Robert [11], Eyink [15], Nagasawa [28], Caffarelli, Kohn, and Nirenberg [6], among others). If such a scenario really occurs, and if we want, at the same time, to keep the total energy balance (1.6) in force, it is necessary to replace the equality in (2.3) by a more general stipulation:

$$\sigma \geq \mathbb{S} : \left(\frac{\nabla_x \mathbf{u}}{\vartheta}\right) - \mathbf{q} \cdot \left(\frac{\nabla_x \vartheta}{\vartheta^2}\right). \tag{2.4}$$

On the other hand, it is a routine matter to check that (2.4) *together with* (1.1)–(1.6) implies, in fact, (2.3), provided s, e satisfy (2.1) and (2.2) and all quantities in question are smooth (see also section 2.3).

As the field equations are supposed to describe a system far from equilibrium, a more natural way would be to *postulate* the existence of the entropy s as a function of ϱ and e and to *define* the absolute temperature $\vartheta^{-1} \equiv \partial_e s > 0$ (see, for example, [21]). However, we have chosen the more "classical" family of the state variables as the constitutive equations we discuss below seem much more transparent in this setting.

2.2. Admissible solutions. Motivated by the previous discussion we are now ready to introduce the class of admissible solutions of problem (1.1)–(1.6).

DEFINITION 2.1. *Let $\Omega \subset R^3$ be a domain. We shall say that a quantity $\{\varrho, \mathbf{u}, \vartheta\}$ represents an admissible solution of problem* (1.1)–(1.6) *on a time interval* $(0, T)$ *provided the following conditions are fulfilled.*

- *The density ϱ and the velocity $\mathbf{u}$,*

$$\varrho \geq 0, \ \ \varrho \in L^\infty(0, T; L^\gamma(\Omega)), \ \ \mathbf{u} \in L^2(0, T; W_0^{1,2}(\Omega; R^3)), \ \ \gamma > \frac{3}{2},$$

solve a renormalized continuity equation in $\mathcal{D}'((0, T) \times R^3)$ provided they were extended to be zero outside Ω. Specifically, the integral identity

$$\int_0^T \int_\Omega (\varrho H(\varrho) \partial_t \varphi + \varrho H(\varrho) \mathbf{u} \cdot \nabla_x \varphi) \, \mathrm{d}x \, \mathrm{d}t = \int_0^T \int_\Omega h(\varrho) \mathrm{div}_x \mathbf{u} \ \varphi \, \mathrm{d}x \, \mathrm{d}t \tag{2.5}$$

holds for any test function $\varphi \in \mathcal{D}((0, T) \times R^3)$ and any

$$h \in BC[0, \infty), \ \ H(\varrho) \equiv H(1) + \int_1^\varrho \frac{h(z)}{z^2} \, \mathrm{d}z. \tag{2.6}$$

- *The momentum equation* (1.2) *is satisfied in $\mathcal{D}'((0, T) \times \Omega)$.*
- *The absolute temperature ϑ,*

$$\vartheta, \ \log(\vartheta) \in L^2(0, T; W^{1,2}(\Omega)),$$

is positive a.a. on $(0, T) \times \Omega$. The specific entropy $s = s(\varrho, \vartheta)$, related to p and e through (2.1) *and* (2.2)*, satisfies* (1.3) *in $\mathcal{D}'((0, T) \times \Omega)$ with the entropy production rate σ—a positive measure on $[0, T] \times \overline{\Omega}$ such that inequality* (2.4) *holds. More specifically, the integral identity*

$$\int_0^T \int_\Omega \left(\varrho s \ \partial_t \varphi + \varrho s \mathbf{u} \cdot \nabla_x \varphi + \frac{\mathbf{q}}{\vartheta} \cdot \nabla_x \varphi \right) \mathrm{d}x \, \mathrm{d}t + \int_0^T \int_\Omega \left[\mathbb{S} : \left(\frac{\nabla_x \mathbf{u}}{\vartheta} \right) - \mathbf{q} \cdot \left(\frac{\nabla_x \vartheta}{\vartheta^2} \right) \right] \varphi \, \mathrm{d}x \, \mathrm{d}t \leq 0 \tag{2.7}$$

is valid for any test function

$$\varphi \in \mathcal{D}((0, T) \times R^3), \quad \varphi \geq 0.$$

- *The total energy balance* (1.6) *is satisfied in $\mathcal{D}'(0, T)$.*

- *The "thermal" energy inequality*

$$\frac{\mathrm{d}}{\mathrm{dt}}\int_\Omega \varrho(e-e_c)\mathrm{d}x \geq \int_\Omega (\mathbb{S}:\nabla_x\mathbf{u} - (p-p_c)\ \mathrm{div}_x\mathbf{u})\,\mathrm{d}x \tag{2.8}$$

holds in $\mathcal{D}'(0,T)$ *with*

$$p_c(\varrho)\equiv p(\varrho,0),\ \ e_c(\varrho)\equiv \int_0^\varrho \frac{p_c(z)}{z^2}\ \mathrm{d}z.$$

2.3. Remarks. A few comments are in order.

(i) To begin with, the reader will have noticed that all quantities appearing in the integral identities introduced in Definition 2.1, and this is to be viewed as an inseparable part of the definition, are tacitly assumed to be at least locally integrable on $(0,T)\times\Omega$.

(ii) The renormalized solutions in the spirit of (2.5) were introduced by DiPerna and Lions in [10]. Here, it is easy to check that the choice $h\equiv 0$ yields the validity of (1.1) in $\mathcal{D}'((0,T)\times\Omega)$ together with *the total mass conservation law*

$$\int_\Omega \varrho(t)\,\mathrm{d}x = M \text{ for any } t\in[0,T]. \tag{2.9}$$

As a matter of fact, we shall make use of a slightly stronger formulation of (2.5), namely,

$$\begin{aligned}\partial_t(\varrho H(\varrho,\Theta)) + \mathrm{div}_x(\varrho H(\varrho,\Theta)\mathbf{u}) + h(\varrho,\Theta)\,\mathrm{div}_x\mathbf{u}\\ = \varrho\frac{\partial H(\varrho,\Theta)}{\partial\Theta}(\partial_t\Theta + \nabla_x\Theta\cdot\mathbf{u})\end{aligned} \tag{2.10}$$

in $\mathcal{D}'((0,T)\times\Omega)$ for any $\Theta\in C^1([0,T]\times\overline{\Omega})$, where

$$\begin{aligned}&\frac{\partial H}{\partial\Theta},\ h\in BC([0,\infty)\times\Theta([0,T]\times\overline{\Omega})),\\ &H(\varrho,\Theta) = H(1,\Theta)+\int_1^\varrho \frac{h(z,\Theta)}{z^2}\,\mathrm{d}z.\end{aligned} \tag{2.11}$$

However, as we show in section 5, (2.5) and (2.10) are *equivalent* in the class of admissible solutions.

(iii) Seemingly, the "correct" statement of (2.8) should read as

$$\frac{\mathrm{d}}{\mathrm{dt}}\int_\Omega \varrho e\,\mathrm{d}x \geq \int_\Omega(\mathbb{S}:\mathbb{D}_x\mathbf{u} - p\ \mathrm{div}_x\mathbf{u})\,\mathrm{d}x. \tag{2.12}$$

Unfortunately, the available a priori estimates are not strong enough to render the "cold" pressure term p_c square integrable—in other words, to legalize the product $p_c\mathrm{div}_x\mathbf{u}$. Note, however, that for smooth solutions, both (2.8) and (2.12) (with equality sign) are equivalent to (1.3).

(iv) In order to conclude, we shall show that any smooth admissible solution solves problem (1.1)–(1.6) in the classical sense. Obviously, the only nonstandard part is to show that (1.3) holds with σ given by (2.3) and that $\mathbf{q}$ satisfies the boundary condition (1.4).

If ϑ is smooth, one can take the product $\varphi\vartheta$ as a test function in (2.7) to deduce

$$\int_0^T \int_\Omega \vartheta(\partial_t(\varrho s) + \mathrm{div}_x(\varrho s \mathbf{u}))\varphi \,\mathrm{d}x\,\mathrm{d}t - \int_0^T \int_\Omega \mathbf{q}\cdot\nabla_x\varphi\,\mathrm{d}x\,\mathrm{d}t$$
$$\geq \int_0^T \int_\Omega \mathbb{S} : \nabla_x \mathbf{u}\ \varphi\,\mathrm{d}x\,\mathrm{d}t \text{ for any } \varphi \in \mathcal{D}((0,T)\times R^3),\ \varphi \geq 0,$$

where, thanks to (1.1) and (2.1),

$$\vartheta\ \partial_t(\varrho s) + \vartheta \mathrm{div}_x(\varrho s\mathbf{u}) = \vartheta\ \varrho\partial_t s + \vartheta\varrho\mathbf{u}\cdot\nabla_x s = \partial_t(\varrho e) + \mathrm{div}_x(\varrho e\mathbf{u}) + p\ \mathrm{div}_x\mathbf{u}.$$

Thus, we have obtained the *internal energy balance*

$$\int_0^T \int_\Omega \varrho e\ \partial_t\varphi + \varrho e\mathbf{u}\cdot\nabla_x\varphi + \mathbf{q}\cdot\nabla_x\varphi\,\mathrm{d}x\,\mathrm{d}t$$
$$\leq \int_0^T \int_\Omega (p\ \mathrm{div}_x\mathbf{u} - \mathbb{S}:\nabla_x\mathbf{u})\varphi\mathrm{d}x\,\mathrm{d}t \text{ for any } \varphi \in \mathcal{D}((0,T)\times R^3),\ \varphi\geq 0. \tag{2.13}$$

On the other hand, we have the *kinetic energy equation*

$$\partial_t\left(\frac{1}{2}\varrho|\mathbf{u}|^2\right) + \mathrm{div}_x\left(\frac{1}{2}\varrho|\mathbf{u}|^2\mathbf{u}\right) + \mathrm{div}_x(p\mathbf{u})$$
$$= \mathrm{div}_x(\mathbb{S}\mathbf{u}) + p\ \mathrm{div}_x\mathbf{u} - \mathbb{S}:\nabla_x\mathbf{u} + \varrho\mathbf{f}\cdot\mathbf{u}, \tag{2.14}$$

which can be deduced from (1.1) and (1.2).

Now it is easy to observe that (2.13) and (2.14) are compatible with the total energy balance (1.6) if and only if (2.13) holds as *equality* for any choice of φ. Consequently, it is a routine matter to deduce the internal energy equation

$$\partial_t(\varrho e) + \mathrm{div}_x(\varrho e\mathbf{u}) + \mathrm{div}_x\ \mathbf{q} = \mathbb{S}:\nabla_x\mathbf{u} - p\ \mathrm{div}_x\mathbf{u} \text{ in } (0,T)\times\Omega \tag{2.15}$$

together with the boundary condition (1.4) to be satisfied by the heat flux $\mathbf{q}$. Of course, (2.15) is equivalent to (1.3) provided all the relevant quantities are smooth.

3. Hypotheses and main results. The *field equations* (1.1)–(1.3), being *linear* with respect to partial derivatives, can be viewed as a family of constraints in the Fourier variable $\mathbf{x}\to\xi$ to be satisfied by any physically admissible flow while the *constitutive equations* represent, in general, *nonlinear* functional relations associated with the material properties of a given fluid. The fundamental idea advocated in the celebrated work of Tartar [35] and DiPerna [8] and [9], to name only a few, asserts that compatibility or rather incompatibility of these stipulations may result in substantial restrictions imposed on the whole set of possible solutions. One can say, very roughly indeed, that the problem will enjoy the property of *weak sequential stability*, which means that any sequence of admissible solutions in the sense of Definition 2.1 possesses an accumulation point, with respect to the topologies induced by the available a priori estimates, that represents another solution of the same problem (see Theorem 3.1). Given the rather poor estimates currently available, this is a nontrivial and usually decisive piece of information that can be directly used to build up a rigorous existence theory.

3.1. The pressure–density–temperature state equation. The constitutive equation relating the pressure p to the state variables ϱ and ϑ will play a central role in our analysis. In accordance with general principles of statistical mechanics, the pressure should obey the following hypotheses.

- In the regime of moderate densities and high temperatures, the perfect gas law (1.7) is always applicable.
- The pressure $p = p(\varrho, \vartheta)$ is a monotone nondecreasing function of ϱ for any fixed ϑ (see Chapter 5 in [20]).
- There is a critical temperature $\Theta_c(\varrho)$ below which the fluid exhibits *degeneration phenomena*. Equivalently, one can say that for any ϑ there is a critical density $\varrho_c(\vartheta)$ so that the "remaining" part $\varrho - \varrho_c$ corresponds to fluid particles with no contribution to the pressure nor to the specific heat at constant volume (see section 2.2 and Chapter 3 in [20]). On the other hand, one can assume the pressure in the "supercritical region" to be that of a degenerate (cold) electron gas given by (1.8) (see Chapter 15 in [14]).
- The pressure is augmented by an intensive *radiation* component p_R when the temperature is extremely high and the density is low (see, for instance, Chapter 15 in [14], [5], and [27]).

Accordingly, the principal mathematical hypotheses imposed on the total pressure p can be formulated as follows.

The total pressure p can be decomposed as

$$p(\varrho, \vartheta) \equiv p_G(\varrho, \vartheta) + p_R(\vartheta), \tag{3.1}$$

where p_R is the pressure due to radiation,

$$p_R(\vartheta) = \frac{d}{3}\vartheta^4, \quad d > 0. \tag{3.2}$$

The *extensive* component $p_G \in C^2((0,\infty) \times (0,\infty))$ satisfies

$$\frac{\partial p_G(\varrho, \vartheta)}{\partial \varrho} \geq 0, \quad \left| \frac{\partial p_G(\varrho, \vartheta)}{\partial \vartheta} \right| \leq c(M) \text{ for all } 0 < \varrho, \vartheta < M, \tag{3.3}$$

where $c(M)$ is bounded for bounded M.

Moreover,

$$\lim_{\varrho \to 0+} p_G(\varrho, \vartheta) = 0 \text{ for any } \vartheta > 0, \quad \lim_{\vartheta \to 0+} p_G(\varrho, \vartheta) = p_c(\varrho) \text{ for any } \varrho > 0. \tag{3.4}$$

The "cold" pressure p_c dominates its thermal complement in the neighborhood of the ϱ-axis bounded above by a graph of a continuous function $\Theta_c = \Theta_c(\varrho) : [0,\infty) \to [0,\infty)$; more specifically,

$$\left| \frac{\partial p_G(\varrho, \vartheta)}{\partial \vartheta} \right| \leq c\big(1 + p_c^{\frac{1}{3}}(\varrho) + \vartheta^3\big) \text{ for all } 0 < \vartheta < \Theta_c(\varrho). \tag{3.5}$$

Finally, we impose a rather technical but, as we have seen above, physically relevant growth condition:

$$a\varrho^\gamma - k \leq p_c(\varrho) \leq a\varrho^\gamma + k \text{ for any } \varrho \geq 0 \quad \text{with } a > 0,\ k > 0,\ \gamma > \frac{3}{2}. \tag{3.6}$$

An overall picture of hypotheses imposed on the state equation can be seen in Figure 3.1.

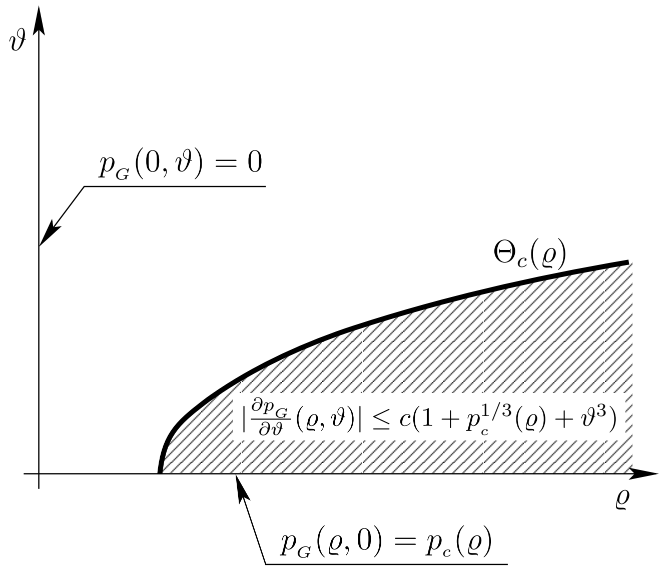


FIG. 3.1. $p_G = p_G(\varrho, \vartheta)$.

3.2. The internal energy and entropy. The specific internal energy e is uniquely determined, modulo a function depending solely on ϑ, by the pressure p through Maxwell's equation (2.2). In addition, writing

$$e(\varrho,\vartheta) = e_G(\varrho,\vartheta) + e_R(\varrho,\vartheta), \text{ with the radiation component } e_R = \frac{d\vartheta^4}{\varrho}, \tag{3.7}$$

we require e_G to satisfy

$$e_G(\varrho,\vartheta) \geq 0, \qquad \lim_{[\varrho,\vartheta]\to[0,0]} e_G(\varrho,\vartheta) = 0, \tag{3.8}$$

$$c_v \equiv \frac{\partial e_G(\varrho,\vartheta)}{\partial \vartheta} \geq 0, \quad c_v \in C([0,\infty) \times [0,\infty)), \tag{3.9}$$

$$c_1(1 + \vartheta^\omega) \leq c_v(\varrho,\vartheta) \leq c_2(1 + \vartheta^\omega), \quad \text{with } c_1 > 0, \ \omega \geq 0, \tag{3.10}$$

for all $\varrho > 0$, $\vartheta > 0$.

The growth condition (3.10) is in agreement with physical experiments predicting $\omega \in [0, 1/2]$ (cf. [40]). The quantities e_G, p_G being interrelated through (2.2) and hypotheses (3.8)–(3.10) implicitly impose certain restrictions on p_G as well.

Now it is easily seen, with p and e given, that the specific entropy s is determined, up to an additive constant, by the thermodynamic equations (2.1).

3.3. Viscosity and thermal conductivity. The fluids considered in this paper are *Newtonian*, which means that the viscous stress tensor $\mathbb{S}$ is a linear function of the velocity gradient $\nabla_x \mathbf{u}$. As a consequence of the principle of *material frame-indifference*, the only physically admissible form of $\mathbb{S}$ must read

$$\mathbb{S} \equiv 2\mu\left(\mathbb{D}_x\mathbf{u} - \frac{1}{N}\,\mathrm{div}_x\mathbf{u}\,\mathbb{I}\right) + \zeta\,\mathrm{div}_x\mathbf{u}\,\mathbb{I} \tag{3.11}$$

with

$$\mathbb{D}_x \mathbf{u} = \frac{1}{2}(\nabla_x \mathbf{u} +^t \nabla_x \mathbf{u}) \text{ being the symmetric part of the velocity gradient,}$$

where μ and ζ are the *shear* and *bulk viscosity coefficients* (see section C.1 in [37]).

In general, the quantities μ and ζ are functions of both ϱ and ϑ. Here, mainly because of technical difficulties, we restrict ourselves to the (physically relevant—see, for instance, [2]) case when μ and ζ depend only on the absolute temperature ϑ. More specifically, we suppose $\mu = \mu(\vartheta)$, $\zeta = \zeta(\vartheta) \in C^1[0,\infty)$ such that

$$m_1(1+\vartheta^\beta) \leq \mu(\vartheta), \quad |\mu'(\vartheta)| \leq m_2(\vartheta^{\beta-1}+1), \quad m_1 > 0,\ \beta \geq 0, \tag{3.12}$$

$$\zeta \geq 0, \quad m_1\vartheta^\beta - 1 \leq \zeta(\vartheta), \quad |\zeta'(\vartheta)| \leq m_2(\vartheta^{\beta-1}+1). \tag{3.13}$$

The requirement that the bulk viscosity coefficient ζ must be positive at least for large values of ϑ may be viewed as slightly restrictive from the physical viewpoint. As a matter of fact, this hypothesis could be dropped at the expense of more severe restrictions imposed on the pressure p. Note, however, that Stokes' relation $\zeta \equiv 0$ inferred from questionable hypotheses is considered unsustainable in light of current theory (cf. section E.1 in [37]).

By virtue of the second law of thermodynamics, expressed through (2.3), the scalar product $\mathbf{q} \cdot \nabla_x \vartheta$ must be nonpositive. For simplicity, we shall assume the heat flux vector $\mathbf{q}$ to be given by Fourier's law

$$\mathbf{q} = -\kappa \nabla_x \vartheta \tag{3.14}$$

with the *heat conductivity coefficient* $\kappa > 0$. To be more precise, the quantity κ itself is allowed to be a continuous function of ϱ and ϑ satisfying the growth condition

$$k_1(1+\vartheta^\alpha) \leq \kappa(\varrho,\vartheta) \leq k_2(1+\vartheta^\alpha), \text{ with } k_1 > 0, \text{ for a certain } \alpha \geq 3. \tag{3.15}$$

The lower bound on the exponent α is motivated by the presence of "radiative" heat conductivity κ_R proportional to ϑ^3 (cf. [5] or [31], [32]). On the other hand, even larger values of $\alpha \approx 4.5$–5.5 can be considered as physically relevant (see [40]).

In agreement with the previous discussion, we shall suppose that the coefficients α and β satisfy

$$0 \leq \beta \leq \frac{4}{3}, \quad \alpha \geq \frac{16}{3} - \beta.$$

These (technical) assumptions make it possible to derive the bounds for dissipation from the thermal energy balance. So far, it is an open question whether they can be relaxed.

3.4. The main result. Having introduced all the necessary hypotheses we are now in a position to state our main result.

THEOREM 3.1. *Let $\Omega \subset R^3$ be a bounded domain with a Lipschitz boundary. Assume that the quantities p, e, s are the given functions of ϱ, ϑ satisfying hypotheses* (3.1)–(3.10)*. Furthermore, let $\mathbb{S}$ and $\mathbf{q}$ be determined through* (3.11) *and* (3.14)*, respectively, where μ, ζ satisfy* (3.12)*,* (3.13) *and κ obeys* (3.15)*. Finally, let*

$$\gamma > \frac{3}{2}, \quad 0 \leq \omega \leq \frac{1}{2}, \quad 0 \leq \beta \leq \frac{4}{3}, \quad \alpha \geq \frac{16}{3} - \beta, \tag{3.16}$$

and

$$\Theta_c(\varrho) \geq c\varrho^{\frac{\gamma}{4}} - 1 \quad \textit{for a certain } c > 0, \tag{3.17}$$

where Θ_c is the critical temperature appearing in (3.5).

Suppose that

$$\left\{ \begin{array}{c} \{\varrho_n\}_{n=1}^{\infty} \subset L^{\infty}(0,T;L^{\gamma}(\Omega)) \cap C([0,T];L^1(\Omega)), \\ \\ \{\mathbf{u}_n\}_{n=1}^{\infty} \subset L^2(0,T;W_0^{1,2}(\Omega;R^3)), \\ \\ \{\vartheta_n\}_{n=1}^{\infty} \subset L^2(0,T;W^{1,2}(\Omega)) \end{array} \right\}$$

is a sequence of admissible solutions of problem (1.1)–(1.6) *on the time interval* $(0,T)$ *in the sense of Definition* 2.1 *such that*

$$\varrho_n(0,\cdot) \to \varrho_0 \ \textit{in } L^1(\Omega), \tag{3.18}$$

$$\operatorname{ess}\limsup_{t\to 0+} \int_{\Omega} \varrho_n(|\mathbf{u}_n|^2 + e(\varrho_n,\vartheta_n))(t)\,\mathrm{d}x \leq E_0, \tag{3.19}$$

$$\operatorname{ess}\liminf_{t\to 0+} \int_{\Omega} \varrho_n s_n(t)\mathrm{d}x \geq S_0,$$

and with

$$\mathbf{f}_n \to \mathbf{f} \ \textit{weakly (*) in } L^{\infty}((0,T)\times\Omega), \tag{3.20}$$

where E_0 and S_0 are constants independent of n.

Then, passing to a subsequence as the case may be, we have

$$\left\{ \begin{array}{c} \varrho_n \to \varrho \ \textit{in } C([0,T];L^1(\Omega)), \\ \\ \mathbf{u}_n \to \mathbf{u} \ \textit{weakly in } L^2(0,T;W_0^{1,2}(\Omega;R^3)), \\ \\ \vartheta_n \to \vartheta \ \textit{weakly in } L^2(0,T;W^{1,2}(\Omega)) \textit{ and strongly in } L^2((0,T)\times\Omega), \end{array} \right\}$$

where the limit quantity $\{\varrho,\mathbf{u},\vartheta\}$ represents another admissible solution of problem (1.1)–(1.6) *on the time interval* $(0,T)$.

One of the first weak stability results for the reduced *barotropic model*, where $p = p(\varrho)$ and the temperature is completely eliminated from the system, was obtained by Lions [24]. Some of the ingredients of his approach can be traced back to the work of Hoff [22] and Serre [33]. Another interesting result in this direction is due to Vaigant and Kazhikhov [38].

The same question as well as the closely related problem of global existence for the full Navier–Stokes–Fourier system with constant viscosity coefficients is studied in [16]; the case of temperature sensitive viscosity is dealt with in [17]; and, finally, these results are extended in [13] in order to include the radiation phenomena.

As already pointed out in section 1, the common feature of the above-mentioned results is that the fluid pressure p_G is always taken to be an affine function of the absolute temperature, whence the internal energy density e admits a decomposition into purely "elastic" and "thermal" components yielding a family of "renormalized" energy equations (see section 4.3.3 in Chapter 4 in [16]). Since these rescaled equations provide additional a priori estimates on the temperature, the technique based on Chacon's biting limit can be used in order to compensate for the lack of integrability of the heat flux $\mathbf{q}$ (see section 6.8.1 in Chapter 6 in [16]). Apparently, such an approach is no longer applicable in the present setting, where the pressure is a general function of ϱ and ϑ, which means that the elastic and thermal contributions to e act simultaneously.

In this work, we introduce a method which makes it possible to treat this general situation. It ranges from nontrivial modifications of approaches introduced in [12], [13], [16], [17], [18] (cf. section 4) to delicate new issues including the temperature-dependent renormalized continuity equation (treated in section 5), which is needed to pass to the limit in the general pressure term.

Regardless of the rather technical hypotheses (3.16) and (3.17), the applicability of the present approach leans essentially on the following characteristic features of the problem.

- The total pressure p is *coercive* at both extremes of the "phase space," which means that on both axes $\{\varrho = 0\}$, $\{\vartheta = 0\}$. From the physical viewpoint, this amounts to taking both radiation and the ionization pressure components into account.
- The viscosity coefficients *depend effectively* on the temperature; in particular,

$$\mu(\vartheta) \to \infty \text{ for } \vartheta \to \infty.$$

  This is a very natural hypothesis, especially for gases.
- The heat conductivity coefficient κ is a superquadratic ("supercubic," as a matter of fact) function of ϑ under large temperature regimes. Such a stipulation seems to be in good agreement with experimental results.

As already pointed out, satisfaction of these conditions does not seem to be at odds with the constitutive laws obtained by the methods of statistical mechanics or empirically, in particular, for real gases. We refer the interested reader to the monographs [2], [20], [27], [39], [40], among others, for the physical background of the present theory.

4. Estimates. The bounds imposed on the family of admissible solutions by hypotheses (3.18)–(3.20) are, very often but rather incorrectly, referred to as a priori estimates. Let us stress, to begin with, that all estimates discussed in this section are "real" estimates to be satisfied by any admissible solution of problem (1.1)–(1.6) in the sense of Definition 2.1 while the term a priori estimates is usually related to *smooth* solutions of a given problem.

4.1. Total mass conservation. As already observed in (2.9), the total mass M of the fluid is a constant of motion thanks to the conservative boundary conditions (1.5). Seeing that ϱ is always nonnegative we infer, in particular, that

$$\{\varrho_n\}_{n=1}^{\infty} \text{ is bounded in } L^\infty(0,T;L^1(\Omega)). \tag{4.1}$$

4.2. Energy estimates. Since the internal energy density e obeys Maxwell's equation (2.2), one can write

$$e_G(\varrho,\vartheta) = P_c(\varrho) + Q(\vartheta) + \int_1^{\varrho} \left(p_G(z,\vartheta) - p_G(z,0) - \vartheta \frac{\partial p_G(z,\vartheta)}{\partial \vartheta} \right) \frac{1}{z^2}\,\mathrm{d}z \tag{4.2}$$

with

$$P_c(\varrho) \equiv \int_0^{\varrho} \frac{p_c(z)}{z^2}\,\mathrm{d}z, \quad Q(\vartheta) \equiv \int_0^{\vartheta} c_v(1,s)\,\mathrm{d}s. \tag{4.3}$$

Note that $e_G(\rho,0) = P_c(\rho)$, and therefore the integral $\int_0^1 p_c(z)/z^2\,\mathrm{d}z$ is convergent in order to comply with hypotheses (3.8) and (3.9).

The total energy balance (1.6) integrated with respect to time yields

$$\int_\Omega E_n(\tau)\mathrm{d}x \le E_0 + \int_0^\tau \int_\Omega \varrho_n \mathbf{f}_n \cdot \mathbf{u}_n \,\mathrm{d}x \text{ for a.a. } \tau \in (0,T)$$

with

$$E_n \equiv \varrho_n \Big(\frac{1}{2}|\mathbf{u}_n|^2 + e(\varrho_n,\vartheta_n)\Big) \text{ and } E_0 \text{ independent of } n.$$

Thus, a straightforward application of Gronwall's lemma gives rise to

$$\{\sqrt{\varrho_n}\mathbf{u}_n\}_{n=1}^\infty \text{ bounded in } L^\infty(0,T;L^2(\Omega;R^3)) \tag{4.4}$$

together with the estimate

$$\sup_{n=1,2,\dots} \Big(\operatorname{ess}\sup_{t\in(0,T)} \int_\Omega \varrho_n e(\varrho_n,\vartheta_n)\mathrm{d}x\Big) < \infty. \tag{4.5}$$

Now, in accordance with hypotheses (3.7) and (3.8), we get

$$\{\vartheta_n\}_{n=1}^\infty \text{ bounded in } L^\infty(0,T;L^4(\Omega)). \tag{4.6}$$

Moreover, writing $e_G(\rho,\vartheta) = e_G(\rho,0) + \int_0^\vartheta \frac{\partial e_G(\rho,s)}{\partial\vartheta}\,\mathrm{d}s$, by virtue of (3.9) and (3.10), we obtain

$$\{\varrho_n\vartheta_n^{1+\omega}\}_{n=1}^\infty \text{ bounded in } L^\infty(0,T;L^1(\Omega)). \tag{4.7}$$

On the other hand, hypotheses (3.8) and (3.9) and formula (4.2) imply

$$Q(\vartheta) + \int_1^\varrho \Big(p_G(z,\vartheta) - p_G(z,0) - \vartheta\frac{\partial p_G(z,\vartheta)}{\partial\vartheta}\Big)\frac{1}{z^2}\,\mathrm{d}z \ge 0,$$

whence (4.5) together with (4.2) and hypothesis (3.6) yields

$$\{\varrho_n\}_{n=1}^\infty \text{ bounded in } L^\infty(0,T;L^\gamma(\Omega)). \tag{4.8}$$

Finally, as the pressure p_G is a nondecreasing function of the density, we obtain

$$0 \le p_G(\varrho,\vartheta) \le p_G(\varrho_c(\vartheta),\vartheta) \ \text{ for any } 0 \le \varrho \le \varrho_c(\vartheta) \le c(1+\vartheta^{\frac{4}{\gamma}}), \tag{4.9}$$

where ϱ_c is the critical density evaluated on the basis of Θ_c (cf. hypothesis (3.17)).

On the other hand, making use of hypotheses (3.5) and (3.17), we deduce

$$\begin{aligned} p_G(\varrho,\vartheta) &= p_c(\varrho) + \int_0^\vartheta \frac{\partial p_G(\varrho,s)}{\partial\vartheta}\,\mathrm{d}s \\ &\le c_1(\varrho^\gamma + \vartheta^4 + \vartheta\varrho^{\frac{\gamma}{3}}) \le c_2(1+\vartheta^4+\varrho^\gamma) \end{aligned} \tag{4.10}$$

whenever

$$0 < \vartheta < \Theta_c(\varrho) \text{ or, equivalently, } \varrho > \varrho_c(\vartheta). \tag{4.11}$$

Relation (4.10) together with (4.9) implies

$$0 \le p_G(\varrho,\vartheta) \le c(1+\varrho^\gamma+\vartheta^4) \ \text{ for all } \ \varrho \ge 0,\ \vartheta \ge 0; \tag{4.12}$$

therefore, estimates (4.6) and (4.8) can be used in combination with (4.12) in order to conclude that

$$\{p(\varrho_n,\vartheta_n)\}_{n=1}^\infty \text{ is bounded in } L^\infty(0,T;L^1(\Omega)). \tag{4.13}$$

4.3. Dissipation estimates I. The active dissipative mechanism manifested through viscosity and thermal conductivity of the fluid under consideration represents an important source of additional estimates involving the spatial derivatives of the velocity field $\mathbf{u}$ and the temperature ϑ.

The entropy production inequality (2.7), integrated with respect to time, yields

$$\int_\Omega \varrho_n s_n(\tau)\mathrm{d}x \geq S_0 + \int_0^\tau \int_\Omega \left(\frac{\mathbb{S}_n : \nabla_x \mathbf{u}_n}{\vartheta_n} + \frac{\kappa(\varrho_n, \vartheta_n)}{\vartheta_n^2} |\nabla_x \vartheta_n|^2 \right) \mathrm{d}x\, \mathrm{d}t$$

for a.a. $\tau \in (0, T)$, where Fourier's law (3.14) and hypothesis (3.19) have been taken into account.

On the other hand, in accordance with the thermodynamic equations (2.1), the specific entropy s may be written in the form

$$s(\varrho, \vartheta) = s_R(\varrho, \vartheta) + s_G(\varrho, \vartheta), \quad s_G(\varrho, \vartheta) \equiv s_G(\varrho, 1) + \int_1^\vartheta \frac{c_v(\varrho, s)}{s} \,\mathrm{d}s, \tag{4.14}$$

where

$$s_R(\varrho, \vartheta) \equiv \frac{4}{3} d \frac{\vartheta^3}{\varrho}, \quad s_G(\varrho, 1) = -\int_1^\varrho \frac{\partial p_G(z, 1)}{\partial \vartheta} \frac{1}{z^2} \,\mathrm{d}z. \tag{4.15}$$

Parallel to (4.10), hypotheses (3.3) and (3.5) can be used in order to deduce

$$|s_G(\varrho, 1)| \leq c\big(1 + \varrho^{-1} + \varrho^{\frac{\gamma}{3} - 1}\big), \tag{4.16}$$

while (3.9) and (3.10) yield

$$\left[\int_1^\vartheta \frac{c_v(\varrho, s)}{s} \,\mathrm{d}s \right]^+ \leq \left\{ \begin{matrix} c(1 + \vartheta^\omega), \ (\omega > 0) \\ c(1 + [\ln \vartheta]^+), \ (\omega = 0) \end{matrix} \right\} \text{ for all } \vartheta > 0.$$

Consequently, having already proved (4.6)–(4.8) we are allowed to conclude that

$$\operatorname*{ess\,sup}_{t \in (0,T)} \int_\Omega \varrho_n s_n(t) \mathrm{d}x \leq c(T), \tag{4.17}$$

whence

$$\left\{ \frac{\mathbb{S}_n : \nabla_x \mathbf{u}_n}{\vartheta_n} \right\}_{n=1}^\infty, \qquad \left\{ \frac{\kappa(\varrho_n, \vartheta_n)}{\vartheta_n^2} |\nabla_x \vartheta_n|^2 \right\}_{n=1}^\infty \text{ are bounded in } L^1((0,T) \times \Omega) \tag{4.18}$$

and

$$\left\{ \varrho_n \left[\int_1^{\vartheta_n} \frac{c_v(\varrho_n, s)}{s} \,\mathrm{d}s \right]^- \right\}_{n=1}^\infty \text{ is bounded in } L^\infty(0, T; L^1(\Omega)). \tag{4.19}$$

In particular, by virtue of hypothesis (3.15), estimate (4.18) yields

$$\{\nabla_x \log(\vartheta_n)\}_{n=1}^\infty, \ \big\{\nabla_x \vartheta_n^{\frac{\alpha}{2}}\big\}_{n=1}^\infty \text{ bounded in } L^2(0, T; L^2(\Omega, R^3)), \tag{4.20}$$

while by virtue of hypothesis (3.10), (4.19) implies

$$\{\varrho_n \log(\vartheta_n)\}_{n=1}^\infty \text{ bounded in } L^\infty(0, T; L^1(\Omega)). \tag{4.21}$$

At this stage, we pause to report an auxiliary result which may be viewed as a straightforward modification of the Poincaré inequality (see, for instance, Lemma 4.1 in [12]).

LEMMA 4.1. *Let $\Omega \subset R^N$, $N \geq 2$, be a bounded Lipschitz domain, and $\Gamma \geq 1$ be a constant. Let $\varrho \geq 0$ be a given function such that*

$$0 < M \leq \int_\Omega \varrho \, \mathrm{d}x, \quad \int_\Omega \varrho^\gamma \mathrm{d}x \leq K,$$

where

$$\gamma > \frac{2N}{N+2}.$$

Then there exists a constant $c = c(M, K, \gamma, \Gamma)$ such that the inequality

$$\|v\|_{L^2(\Omega)} \leq c(M,K)\left(\|\nabla_x v\|_{L^2(\Omega)} + \left[\int_\Omega \varrho |v|^{\frac{1}{\Gamma}} \mathrm{d}x \right]^\Gamma \right)$$

holds for any $v \in W^{1,2}(\Omega)$.

The total mass being conserved, more specifically,

$$M_n \equiv \int_\Omega \varrho_n(t) \mathrm{d}x = \int_\Omega \varrho_n(0) \mathrm{d}x \quad \to \quad \int_\Omega \varrho_0 \mathrm{d}x \equiv M > 0,$$

Lemma 4.1 applies, together with estimates (4.7), (4.20), and (4.21), in order to conclude

$$\{\log(\vartheta_n)\}_{n=1}^\infty, \ \{\vartheta_n^{\frac{\alpha}{2}}\}_{n=1}^\infty \text{ bounded in } L^2(0,T;W^{1,2}(\Omega)). \tag{4.22}$$

4.4. Dissipation estimates II. A short examination of the "thermal" energy inequality (2.8) yields

$$\begin{aligned} \int_0^\tau \int_\Omega \mathbb{S}_n : \nabla_x \mathbf{u}_n \, \mathrm{d}x \, \mathrm{d}t \\ \leq 2E_0 + \int_\Omega \varrho_n (e_n - e_c(\varrho_n))(\tau) \, \mathrm{d}x + \int_0^\tau \int_\Omega (p_n - p_c(\varrho_n)) \, \mathrm{div}_x \mathbf{u}_n \, \mathrm{d}x \, \mathrm{d}t \end{aligned} \tag{4.23}$$

for a.a. $\tau \in (0,T)$.

We claim, keeping in mind hypotheses (3.4) and (3.5), that there is a positive constant c such that

$$|p(\varrho,\vartheta) - p_c(\varrho)| \leq c(1 + \vartheta^4 + \vartheta \varrho^{\frac{\gamma}{3}}) \text{ for all } \varrho > 0, \ \vartheta > 0. \tag{4.24}$$

Indeed, similar to (4.10), we have

$$|p_G(\varrho,\vartheta) - p_c(\varrho)| = \left| \int_0^\vartheta \frac{\partial p_G(\varrho, s)}{\partial \vartheta} \, \mathrm{d}s \right| \leq c(1 + \vartheta^4 + \vartheta \varrho^{\frac{\gamma}{3}})$$

provided $\varrho \geq \varrho_c(\vartheta)$. Moreover, as p_G is a nondecreasing function of ϱ,

$$|p_G(\varrho,\vartheta) - p_c(\varrho)| \leq p_G(\varrho_c(\vartheta),\vartheta) - p_c(\varrho_c(\vartheta)) + 2p_c(\varrho_c(\vartheta)) \leq c(1 + \vartheta^4)$$

for all $0 \leq \varrho \leq \varrho_c(\vartheta)$; cf. estimate (4.9).

Now, by virtue of Hölder's inequality and the imbedding theorem $W^{1,2}(\Omega) \subset L^6(\Omega)$, the integral in the extreme right of (4.23) can be estimated as

$$\left| \int_0^\tau \int_\Omega (p_n - p_c(\varrho_n)) \mathrm{div}_x \mathbf{u}_n \, \mathrm{d}x \, \mathrm{d}t \right| \leq c\big(\|(1+\vartheta_n^4)\mathrm{div}_x \mathbf{u}_n\|_{L^1((0,T)\times\Omega)} + \|\vartheta_n \varrho_n^{\frac{\gamma}{3}} \mathrm{div}_x \mathbf{u}_n\|_{L^1((0,T)\times\Omega)}\big), \tag{4.25}$$

where

$$\|\vartheta_n \varrho_n^{\frac{\gamma}{3}} \mathrm{div}_x \mathbf{u}_n\|_{L^1(\Omega)} \leq \|\vartheta_n\|_{W^{1,2}(\Omega)} \|\varrho_n^\gamma\|_{L^1(\Omega)}^{\frac{1}{3}} \|\nabla_x \mathbf{u}_n\|_{L^2(\Omega;R^N)}.$$

Furthermore, writing

$$\vartheta^4 \mathrm{div}_x \mathbf{u} = \vartheta^{4-\frac{\beta}{2}} \vartheta^{\frac{\beta}{2}} \mathrm{div}_x \mathbf{u},$$

we get, by interpolation,

$$\|\vartheta^{4-\frac{\beta}{2}}\|_{L^2(\Omega)}^2 \leq \|\vartheta\|_{L^{3\alpha}(\Omega)}^{\frac{3\alpha(\beta-4)}{4-3\alpha}} \|\vartheta\|_{L^4(\Omega)}^{8-\beta-\frac{3\alpha(\beta-4)}{4-3\alpha}}, \tag{4.26}$$

where, in accordance with hypothesis (3.16), $3\alpha(\beta-4)/(4-3\alpha) \leq \alpha$. Consequently, by virtue of estimates (4.6) and (4.22),

$$\{\vartheta_n^{4-\frac{\beta}{2}}\}_{n=1}^\infty \text{ is bounded in } L^2((0,T)\times\Omega). \tag{4.27}$$

Due to (3.11), $\mathbb{S} : \nabla_x \mathbf{u} = 2\mu \langle \mathbb{D}_x \mathbf{u} \rangle : \langle \mathbb{D}_x \mathbf{u} \rangle + \zeta |\mathrm{div}_x \mathbf{u}|^2$, where $\langle \mathbb{D}_x \mathbf{u} \rangle_{i,j} = \frac{1}{2}(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}) - \frac{1}{3} \mathrm{div}_x \mathbf{u} \delta_{i,j}$. Moreover, the Korn-type inequality

$$\|\nabla_x \mathbf{u}\|_{L^p(\Omega;R^{3\times3})} \leq c(p) \|\langle \mathbb{D}_x(\mathbf{u}) \rangle\|_{L^p(\Omega;R^{3\times3}_{sym})} \tag{4.28}$$

holds true for any $1 < p < \infty$ and for any $\mathbf{u} \in W_0^{1,p}(\Omega, R^N)$ (cf. Proposition 2.4 in [36] or section 5.1 in [18]). Finally, by virtue of (4.25)–(4.27), (4.8), and (4.22),

$$\left| \int_0^T \int_\Omega (p_n - p_c(\rho_n)) \mathrm{div}_x \mathbf{u}_n \, \mathrm{d}x \, \mathrm{d}t \right| \leq c(\|\vartheta_n^{\frac{\beta}{2}} \mathrm{div}_x \mathbf{u}_n\|_{L^2((0,T)\times\Omega)} + \|\nabla \mathbf{u}_n\|_{L^2((0,T)\times\Omega)}). \tag{4.29}$$

Moreover, by virtue of (4.5) and (4.8), integral $\int_\Omega \varrho_n(e_n - e_c(\varrho_n)) \, \mathrm{d}x$ is bounded in $L^\infty(0,T)$.

Thus, combining estimate (4.23) with inequalities (4.28) and (4.29) and with hypotheses (3.12), (3.13), and (3.16), we obtain, in particular,

$$\{\mathbf{u}_n\}_{n=1}^\infty \text{ bounded in } L^2(0,T;W_0^{1,2}(\Omega;R^3)). \tag{4.30}$$

Then, of course,

$$\{\mathbb{S}_n : \nabla_x \mathbf{u}_n\}_{n=1}^\infty \text{ is bounded in } L^1((0,T)\times\Omega). \tag{4.31}$$

Finally, from (3.11), taking into account bounds (4.27) and (4.30) and hypotheses (3.12), (3.13), and (3.16), we can easily verify that

$$\{\mathbb{S}_n\}_{n=1}^\infty \text{ is bounded in } L^{\frac{5}{3}}(0,T;L^{\frac{5}{3}}(\Omega;R^{3\times3}_{sym})). \tag{4.32}$$

4.5. Refined pressure and density estimates. Refined density estimates up to the boundary can be obtained by using the quantities

$$\varphi_j = \psi\left[\mathcal{B}\left(b(\varrho) * \phi_\epsilon - \frac{1}{|\Omega|}\int_\Omega b(\varrho) * \phi_\epsilon \, \mathrm{d}x\right)\right]_j$$

as test functions in the variational formulation of the momentum equation (1.2). Here ψ is a convenient function in $C_0^\infty(I)$, $b \in C([0,\infty)$ is a function with convenient growth at infinity (see later), the symbol $*$ denotes convolution, $\phi_\epsilon \to \delta(0)$ is a regularizing sequence in the variable t, and $\mathcal{B}$ is the Bogovskii operator satisfying

$$\begin{aligned}&\mathrm{div}\mathcal{B}(v) = v, \quad \|\mathcal{B}(v)\|_{L^q(\Omega)} \le c(q)\|\mathbf{g}\|_{L^q(\Omega)}, \quad \|\nabla\mathcal{B}(v)\|_{L^p(\Omega)} \le c(p)\|v\|_{L^p(\Omega)},\\ &\text{with any } 1 < q, p < \infty,\ \mathbf{g} \in L^q(\Omega; R^N),\ v = \mathrm{div}\mathbf{g} \in L^p(\Omega), \text{ and } \mathbf{g}\cdot\mathbf{n}|_{\partial\Omega} = 0\end{aligned}$$

(cf., for instance, [3], [4], [19], or Lemma 3.17 in [29]).

With estimates (4.4), (4.8), (4.13), (4.30), and (4.32) at hand, such a procedure yields, after a tedious but straightforward computation, a bound

$$\sup_{n=1,2,\dots} \int_0^T \int_\Omega p(\varrho_n, \vartheta_n) b(\varrho_n) \, \mathrm{d}x \, dt < \infty, \tag{4.33}$$

provided $b(\varrho) \approx \varrho^\nu$, with $\nu = \nu(\gamma) > 0$ sufficiently small. Local (interior) pressure estimate of type (4.33) was first established for the barotropic flow in [24], where the optimal (largest) value of $\nu = (2/N)\gamma - 1$ was obtained. For more details concerning the techniques of the calculus leading to (4.33) see, for instance, section 7.9.5 in [29].

Now, writing

$$p(\varrho_n, \vartheta_n) = p_R(\vartheta_n) + \big(p_G(\varrho_n, \vartheta_n) - p_c(\varrho_n)\big) + p_c(\varrho_n),$$

we infer, with the help of (4.24), that the sequence

$$\int_0^T \int_\Omega \varrho_n^{\gamma+\nu} \, \mathrm{d}x \, dt \text{ is bounded.} \tag{4.34}$$

Indeed, by virtue of (4.6), (4.22) and a simple interpolation argument,

$$\{\vartheta_n\}_{n=1}^\infty \text{ is bounded in } L^p((0,T)\times\Omega) \text{ for a certain } p > 4, \tag{4.35}$$

whence (4.34) follows easily from (4.24) and (4.33). Now, by using (3.1), (3.2), and (4.12) it is easy to conclude that[1]

$$p(\rho_n, \vartheta_n) \text{ is bounded in } L^p((0,T)\times\Omega) \text{ for some } p > 1. \tag{4.36}$$

[1] Note that, in contrast with (4.13), estimates (4.34) and (4.35) render the total pressure $\{p(\varrho_n, \vartheta_n)\}_{n=1}^\infty$ *equi-integrable* and thus precompact in the weak topology of the Lebesgue space $L^1((0,T)\times\Omega)$. This is the least information which can replace (4.36) in our proof. In the light of the technical difficulties connected with the construction of $\mathcal{B}$, it seems worth noting that a weaker, but for our purposes still sufficient, result can be obtained via multipliers

$$w_j,\ j = 1, \dots, N, \quad \mathrm{div}_x[\mathbf{w}] = h, \qquad \lim_{\mathrm{dist}(x,\partial\Omega)\to 0} h(x) = \infty,$$

where $h \in L^p(\Omega)$ with $p = p(\gamma) >> 1$ large enough. Indeed in such a way we can show that

$$\{\varrho_n^\gamma\}_{n=1}^\infty \text{ is equi-integrable in } L^1((0,T)\times\Omega),$$

which means

$$\lim_{k\to\infty} \int_{\{\varrho_n \ge k\}} \varrho_n^\gamma \, \mathrm{d}x \, \mathrm{d}t \to 0 \text{ uniformly for } n = 1, 2, \dots.$$

5. Weak sequential compactness. As ϱ_n satisfies (1.1) in the sense of distributions, and since (4.4) and (4.8) hold, we may assume

$$\varrho_n \to \varrho \text{ in } C([0,T]; L^{\gamma}_{weak}(\Omega)) \tag{5.1}$$

passing to a subsequence as the case may be. (Here and in what follows, $L^p_{weak}(\Omega)$ means $L^p(\Omega)$ endowed with the weak topology.)

Similarly, by virtue of (4.30),

$$\mathbf{u}_n \to \mathbf{u} \text{ weakly in } L^2(0,T; W^{1,2}_0(\Omega; R^3)), \tag{5.2}$$

and, in accordance with (4.6) and (4.22),

$$\vartheta_n \to \vartheta \text{ weakly in } L^2(0,T; W^{1,2}(\Omega)) \text{ and weakly-(*) in } L^{\infty}(0,T; L^4(\Omega)) \tag{5.3}$$

at least for a suitable subsequence.

The main goal of this section is to show, in accordance with the conclusion of Theorem 3.1, that ϱ, $\mathbf{u}$, and ϑ represent another admissible solutions of problem (1.1)–(1.6) on the time interval $(0,T)$.

5.1. Convergence of the convective terms. The space $L^{\gamma}(\Omega)$ is compactly imbedded into the dual space $W^{-1,2}(\Omega)$ as soon as $\gamma > 6/5$. Consequently, by virtue of (5.1) and (5.2) and hypothesis (3.16),

$$\varrho_n \mathbf{u}_n \to \varrho \mathbf{u} \text{ weakly-(*) in } L^{\infty}(0,T; L^{\frac{2\gamma}{\gamma+1}}(\Omega; R^3)). \tag{5.4}$$

Moreover, since $(\rho_n, \mathbf{u}_n, \vartheta_n)$ satisfies (1.2) in the sense of distributions, keeping in mind bounds (4.8), (4.30), (4.32), and (4.36), one can sharpen (5.4) to

$$\varrho_n \mathbf{u}_n \to \varrho \mathbf{u} \text{ in } C([0,T]; L^{\frac{2\gamma}{\gamma+1}}_{weak}(\Omega; R^3)). \tag{5.5}$$

Now, since $\gamma > 3/2$, we get, similar to the above, $L^{\frac{2\gamma}{\gamma+1}}(\Omega; R^3)$ compactly imbedded into the dual space $W^{-1,2}(\Omega; R^N)$, whence we are allowed to repeat the same arguments as in (5.4) in order to deduce

$$\varrho_n \mathbf{u}_n \otimes \mathbf{u}_n \to \varrho \mathbf{u} \otimes \mathbf{u} \text{ weakly in } L^2(0,T; L^{\frac{6\gamma}{4\gamma+3}}(\Omega; R^{3\times 3}_{sym})). \tag{5.6}$$

In particular, it is possible to pass to the limit for $n \to \infty$ in (1.1) and (1.2) to obtain

$$\partial_t \varrho + \operatorname{div}_x(\varrho \mathbf{u}) = 0 \text{ in } \mathcal{D}'((0,T) \times R^3), \tag{5.7}$$

provided ϱ, $\mathbf{u}$ were extended to be zero outside Ω, and

$$\partial_t(\varrho \mathbf{u}) + \operatorname{div}_x(\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla_x \overline{p} = \operatorname{div}_x \overline{\mathbb{S}} + \overline{\varrho \mathbf{f}} \text{ in } \mathcal{D}'((0,T) \times R^3)^3, \tag{5.8}$$

where

$$p(\varrho_n, \vartheta_n) \to \overline{p(\varrho, \vartheta)} \text{ weakly in } L^1((0,T) \times \Omega),$$

$$\mathbb{S}_n = \mathbb{S}\,(\mathbb{D}_x \mathbf{u}_n, \vartheta_n) \to \overline{\mathbb{S}(\mathbb{D}_x \mathbf{u}, \vartheta)} \text{ weakly in } L^{\frac{5}{3}}(0,T; L^{\frac{5}{3}}(\Omega; R^3)),$$

$$\varrho_n \mathbf{f}_n \to \overline{\varrho \mathbf{f}} \text{ weakly in } L^1(0,T; L^1(\Omega; R^3))$$

conformably to (4.32) and (4.36).

5.2. Strong convergence of the temperature. Our aim is to show that the sequence $\{\vartheta_n\}_{n=1}^\infty$ converges strongly in the Lebesgue space $L^1((0,T)\times\Omega)$, in particular, almost anywhere on $(0,T)\times\Omega$ for a suitable subsequence. This step represents one of the most delicate issues to be discussed in this paper.

To begin with, let us record the following version of the celebrated Aubin–Lions lemma (see Lemma 6.3 in Chapter 6 in [16]).

LEMMA 5.1. *Let $\Omega \subset R^N$, $N \geq 2$, be a bounded Lipschitz domain. Let $\{v_n\}_{n=1}^\infty$ be a sequence of functions bounded in*

$$L^2(0,T;L^q(\Omega)) \cap L^\infty(0,T;L^1(\Omega)) \text{ with } q > \frac{2N}{N+2}.$$

Furthermore, assume that

$$\partial_t v_n \geq g_n \text{ in } \mathcal{D}'((0,T)\times\Omega),$$

where the distributions g_n are bounded in the space $L^1(0,T;W^{-m,p}(\Omega))$ for certain $m \geq 1$, $p > 1$.

Then we have

$$v_n \to v \text{ in } L^2(0,T;W^{-1,2}(\Omega))$$

passing to a subsequence as the case may be.

Since $\{\vartheta_n\}_{n=1}^\infty$ admits the bound established in (4.6), we have

$$\{\varrho_n s_R(\varrho_n,\vartheta_n)\}_{n=1}^\infty \text{ bounded in } L^\infty(0,T;L^{\frac{4}{3}}(\Omega)). \tag{5.9}$$

Moreover, relation (4.16) yields

$$|\varrho s_G(\varrho,1)| \leq c(1+\varrho+\varrho^{\frac{\gamma}{3}}),$$

so we obtain, by virtue of (4.8), that

$$\{\varrho_n s_G(\varrho_n,1)\}_{n=1}^\infty \text{ is bounded in } L^\infty(0,T;L^p(\Omega)) \text{ with } p=\min(3,\gamma) \tag{5.10}$$

(cf. (4.14) and (4.15)).

Finally, hypotheses (3.9) and (3.10) together with estimates (4.6), (4.21), and (4.22) can be used in order to conclude that

$$\{\varrho_n \Phi(\varrho_n,\vartheta_n)\}_{n=1}^\infty \text{ is bounded in } L^\infty(0,T;L^1(\Omega)) \cap L^2(0,T;L^q(\Omega)) \tag{5.11}$$

with at least any q such that $\frac{q}{\gamma}+\frac{q}{6} \leq 1$, where we have denoted

$$\Phi(\varrho,\vartheta) \equiv \int_1^\vartheta \frac{c_v(\varrho,s)}{s}\,\mathrm{d}s. \tag{5.12}$$

Then, of course,

$$\{\varrho_n s_n \mathbf{u}_n\}_{n=1}^\infty \text{ is bounded in } L^1(0,T;L^p(\Omega,R^3)) \text{ with some } p>1. \tag{5.13}$$

Consequently, the bounds obtained in (5.9)–(5.13) together with (4.18) and (4.22) imply that the sequence $v_n = \varrho_n s_n$, satisfying the entropy production inequality (2.7), complies with the hypotheses of Lemma 5.1, where we have taken $g_n = -\mathrm{div}_x(\rho_n s_n \mathbf{u}_n) - \mathrm{div}_x(\kappa(\varrho_n,\vartheta_n)\frac{\nabla\vartheta_n}{\vartheta_n}) + \mathbb{S}_n : \frac{\nabla\mathbf{u}_n}{\vartheta_n} + \kappa(\varrho_n,\vartheta_n)\frac{|\nabla\vartheta_n|^2}{\vartheta_n^2}$.

Thus, we may assume

$$\varrho_n s_n \to \frac{4}{3} d\overline{\vartheta^3} + \overline{\varrho s_G(\varrho, 1)} + \overline{\varrho \Phi(\varrho, \vartheta)} \text{ in } L^2(0, T; W^{-1,2}(\Omega)), \tag{5.14}$$

where, as always, we have used the bar to denote the weak limits of composed functions.

The functions ϱ_n, $\mathbf{u}_n$ solve the renormalized equation (2.5). Therefore, in particular, (5.10) implies

$$\varrho_n s_G(\varrho_n, 1) \to \overline{\varrho s_G(\varrho, 1)} \text{ in } C([0, T]; L^q_{weak}(\Omega)),\ q = \min(3, \gamma).$$

Consequently, since ϑ_n are bounded by (4.22), we get

$$\varrho_n s_G(\varrho_n, 1)\vartheta_n \to \overline{\varrho s_G(\varrho, 1)}\vartheta \text{ weakly in } L^1((0, T) \times \Omega). \tag{5.15}$$

Furthermore, the function Φ is nondecreasing with respect to ϑ; in particular,

$$\int_0^T \int_\Omega \varrho_n (\Phi(\varrho_n, \vartheta_n) - \Phi(\varrho_n, \vartheta))(\vartheta_n - \vartheta)\varphi \, \mathrm{d}x \, \mathrm{d}t \geq 0 \tag{5.16}$$

for any nonnegative test function $\varphi \in \mathcal{D}((0, T) \times \Omega)$.

As we will show below,

$$\int_0^T \int_\Omega \varrho_n \Phi(\varrho_n, \vartheta)(\vartheta_n - \vartheta)\varphi \, \mathrm{d}x \, \mathrm{d}t \to 0 \text{ as } n \to \infty; \tag{5.17}$$

therefore, (5.16) reduces to

$$\liminf_{n\to\infty} \int_0^T \int_\Omega \varrho_n \Phi(\varrho_n, \vartheta_n)\vartheta_n \ \varphi \, \mathrm{d}x \, \mathrm{d}t \geq \int_0^T \int_\Omega \overline{\varrho \Phi(\varrho, \vartheta)}\vartheta \ \varphi \mathrm{d}x \, \mathrm{d}t \tag{5.18}$$

for any nonnegative $\varphi \in \mathcal{D}((0, T) \times \Omega)$.

Taking, for a moment, relation (5.17) for granted, one can deduce easily from (5.3), (5.14), (5.15), and (5.18) that

$$\begin{aligned} &\int_0^T \int_\Omega \left(\frac{4}{3} d\overline{\vartheta^3} + \overline{\varrho s_G(\varrho, 1)} + \overline{\varrho \Phi(\varrho, \vartheta)} \right) \vartheta\varphi \mathrm{d}x \, \mathrm{d}t \\ &= \lim_{n\to\infty} \int_0^T \int_\Omega \varrho_n s_n \vartheta_n \varphi \, \mathrm{d}x \, \mathrm{d}t \\ &= \lim_{n\to\infty} \int_0^T \int_\Omega \left(\frac{4}{3} d\vartheta_n^3 + \varrho_n s_G(\varrho_n, 1) + \varrho_n \Phi(\varrho_n, \vartheta_n) \right) \vartheta_n \varphi \, \mathrm{d}x \, \mathrm{d}t \\ &\geq \limsup_{n\to\infty} \int_0^T \int_\Omega \frac{4}{3} d\vartheta_n^4 \varphi \mathrm{d}x \, \mathrm{d}t \\ &+ \int_0^T \int_\Omega (\overline{\varrho s_G(\varrho, 1)} + \overline{\varrho \Phi(\varrho, \vartheta)})\vartheta\varphi \, \mathrm{d}x \, \mathrm{d}t. \end{aligned} \tag{5.19}$$

In other words

$$\limsup_{n\to\infty} \int_0^T \int_\Omega \vartheta_n^4 \varphi \, \mathrm{d}x \, \mathrm{d}t \leq \int_0^T \int_\Omega \overline{\vartheta^3}\vartheta\varphi \mathrm{d}x \, \mathrm{d}t, \quad \varphi \in \mathcal{D}((0, T) \times \Omega), \quad \varphi \geq 0.$$

On the other hand, passing to the limit $n \to \infty$ in the evident inequality $\int_0^T \int_\Omega (\vartheta_n^3 - \vartheta^3)(\vartheta_n - \vartheta)\varphi \, \mathrm{d}x \, \mathrm{d}t \geq 0$, by using (5.3) one obtains

$$\liminf_{n\to\infty} \int_0^T \int_\Omega \vartheta_n^4 \varphi \mathrm{d}x \, \mathrm{d}t \leq \int_0^T \int_\Omega \overline{\vartheta^3}\vartheta\varphi \mathrm{d}x \, \mathrm{d}t, \ \varphi \in \mathcal{D}((0,T)\times\Omega), \ \varphi \geq 0.$$

From the last two inequalities we deduce

$$\overline{\vartheta^4} = \overline{\vartheta^3}\vartheta,$$

and we conclude, making use of the standard argument of Minty (see, e.g., Lemmas 3.35 and 3.39 in [29]), that

$$\vartheta_n \to \vartheta \text{ in } L^4((0,T)\times\Omega). \tag{5.20}$$

Consequently, by virtue of (5.2) and (5.20), one can write in (5.8)

$$\overline{\mathbb{S}} = \mathbb{S}, \tag{5.21}$$

where $\mathbb{S}$ is given by (3.11) with $\mu = \mu(\vartheta)$, $\zeta = \zeta(\vartheta)$.

It remains to show (5.17). To this end, we need the generalized form (2.10) of the renormalized continuity equation (2.5). The following auxiliary result may be of independent interest.

LEMMA 5.2. *Let*

$$\varrho \geq 0, \ \varrho \in L^\infty(0,T;L^\gamma(\Omega)), \ \mathbf{u} \in L^2(0,T;W_0^{1,2}(\Omega;R^N)), \ \gamma > \frac{2N}{N+2},$$

be a renormalized solution of (1.1) *in the sense specified in Definition* 2.1.

Then ϱ, $\mathbf{u}$ *satisfy* (2.10) *in* $\mathcal{D}'((0,T)\times\Omega)$ *for any* h, H *as in* (2.11), *and* $\Theta \in C^1((0,T)\times\Omega)$.

Proof. We shall use the truncation technique introduced in [16] along with the regularizing procedure due to DiPerna and Lions [10]. To this end, consider the cutoff functions $T_k \in C^\infty(R)$,

$$T_k(\varrho) = k\mathcal{T}\Big(\frac{\varrho}{k}\Big), \quad k \geq 1, \tag{5.22}$$

where $\mathcal{T} \in C^\infty(R)$,

$$\left\{\begin{array}{c} \mathcal{T}(-z) = -\mathcal{T}(z) \text{ for all } z \in R, \\ \\ \mathcal{T}(z) = z \text{ for } 0 \leq z \leq 1, \ \mathcal{T}''(z) \leq 0 \text{ if } 1 \leq z \leq 3, \ \mathcal{T}(z) = 2 \text{ for all } z \geq 3. \end{array}\right\}$$

As ϱ, $\mathbf{u}$ satisfy (2.5), we get

$$\partial_t T_k(\varrho) + \mathrm{div}_x(T_k(\varrho)\mathbf{u}) + \big(T_k'(\varrho)\varrho - T_k(\varrho)\big)\mathrm{div}_x\mathbf{u} = 0 \tag{5.23}$$

in $\mathcal{D}'((0,T)\times\Omega)$.

Similar to [10], consider a family of Friedrichs' mollifiers $\eta^\varepsilon = \eta^\varepsilon(x) \in \mathcal{D}(R^N)$, $\varepsilon > 0$,

$$\left\{\begin{array}{c} \eta^\varepsilon \geq 0, \ \eta^\varepsilon \text{ radially symmetric and radially decreasing,} \\ \displaystyle\int_{R^N} \eta^\varepsilon \, \mathrm{d}x = 1, \ \mathrm{supp}[\eta^\varepsilon] \subset \{|x| < \varepsilon\}. \end{array}\right\}$$

Using η^ε as a test function in the variational formulation of (5.23), we obtain

$$\partial_t [T_k(\varrho)]^\varepsilon + \mathrm{div}_x([T_k(\varrho)]^\varepsilon \mathbf{u}) + [(T_k'(\varrho)\varrho - T_k(\varrho))\mathrm{div}_x \mathbf{u}]^\varepsilon = r_\varepsilon \tag{5.24}$$

with

$$[h]^\varepsilon \equiv \eta^\varepsilon * h \text{ and } r_\varepsilon \equiv \mathrm{div}_x([T_k(\varrho)]^\varepsilon \mathbf{u}) - [\mathrm{div}_x(T_k(\varrho)\mathbf{u})]^\varepsilon,$$

where the symbol $*$ denotes the convolution with respect to x. Note that (5.24) holds, in the sense of strong derivatives, for a.a. $(t,x) \in (0,T) \times \Omega$ such that $\mathrm{dist}[x, \partial\Omega] > \varepsilon$.

Now we are allowed to multiply (5.24) on

$$H([T_k(\varrho)]^\varepsilon, \Theta) + [T_k(\varrho)]^\varepsilon \frac{\partial H([T_k(\varrho)]^\varepsilon, \Theta)}{\partial \varrho},$$

with $\Theta \in C^1((0,T) \times \Omega)$, H as in (2.11), in order to obtain

$$\begin{aligned}
&\partial_t([T_k(\varrho)]^\varepsilon H([T_k(\varrho)]^\varepsilon, \Theta)) + \mathrm{div}_x([T_k(\varrho)]^\varepsilon H([T_k(\varrho)]^\varepsilon, \Theta)\mathbf{u}) + h([T_k(\varrho)]^\varepsilon, \Theta)\ \mathrm{div}_x \mathbf{u} \\
&\quad = [T_k(\varrho)]^\varepsilon \frac{\partial H([T_k(\varrho)]^\varepsilon, \Theta)}{\partial \Theta} \big(\partial_t \Theta + \nabla_x \Theta \cdot \mathbf{u}\big) \\
&\qquad + \{r_\varepsilon - [(T_k'(\varrho)\varrho - T_k(\varrho))\mathrm{div}_x \mathbf{u}]^\varepsilon\} \left\{ H([T_k(\varrho)]^\varepsilon, \Theta) + [T_k(\varrho)]^\varepsilon \frac{\partial H([T_k(\varrho)]^\varepsilon, \Theta)}{\partial \varrho} \right\}.
\end{aligned}$$

Seeing that, for k fixed,

$$r_\varepsilon \to 0 \text{ in } L^p((0,T) \times K) \text{ for any compact } K \subset \Omega \text{ as } \varepsilon \to 0,\ p \in [1,2)$$

(see, for instance, [10]), we deduce

$$\begin{aligned}
&\partial_t(T_k(\varrho) H(T_k(\varrho), \Theta)) + \mathrm{div}_x(T_k(\varrho) H(T_k(\varrho), \Theta)\mathbf{u}) \\
&\quad + h(T_k(\varrho), \Theta)\ \mathrm{div}_x \mathbf{u} = T_k(\varrho) \frac{\partial H(T_k(\varrho), \Theta)}{\partial \Theta} (\partial_t \Theta + \nabla_x \Theta \cdot \mathbf{u}) \\
&\quad + (T_k(\varrho) - T_k'(\varrho)\varrho) \left\{ H(T_k(\varrho), \Theta) + T_k(\varrho) \frac{\partial H(T_k(\varrho), \Theta)}{\partial \varrho} \right\} \mathrm{div}_x \mathbf{u}
\end{aligned} \tag{5.25}$$

in $\mathcal{D}'((0,T) \times \Omega)$. Assume, in addition to (2.11), that

$$H(\varrho, \Theta) = 0 \text{ for all } \varrho \geq M,\ \Theta \in R,\ \text{and a certain } M > 0. \tag{5.26}$$

Now it is easy to let $k \to \infty$ in (5.25) to conclude, with the help of the Lebesgue theorem, that

$$\begin{aligned}
&\partial_t\big(\varrho H(\varrho, \Theta)\big) + \mathrm{div}_x\big(\varrho H(\varrho, \Theta)\mathbf{u}\big) + h(\varrho, \Theta)\mathrm{div}_x \mathbf{u} \\
&\quad = \varrho \frac{\partial H(\varrho, \Theta)}{\partial \Theta} \big(\partial_t \Theta + \nabla_x \Theta \cdot \mathbf{u}\big) \text{ in } \mathcal{D}'((0,T) \times \Omega).
\end{aligned} \tag{5.27}$$

Finally, approximating

$$H(\varrho, \Theta) \approx H_m(\varrho, \Theta) = \chi\left(\frac{\varrho}{m}\right) H(\varrho, \Theta),\ \text{with } \chi \in \mathcal{D}(R),\ \chi(z) = 1 \text{ for } |z| \leq 1,$$

and letting $m \to \infty$ in (5.27), we obtain (2.10) for any choice of h, H, and Θ compatible with (2.11). $\square$

Now we come back to (5.17). We can easily verify that the function $H(\rho,\tilde{\vartheta}) = \Phi(T_k(\rho),\tilde{\vartheta})$, where

$$\tilde{\vartheta} \in C^1([0,T]\times\overline{\Omega}), \qquad \inf_{t\in(0,T),x\in\Omega} \tilde{\vartheta}(t,x) > 0, \tag{5.28}$$

satisfies assumptions (2.11). Consequently, it verifies renormalized continuity equation (5.27) with $\tilde{\vartheta}$ in place of Θ. From this fact, by using (5.12), (3.9), (3.10), (5.22), and (4.8) we conclude that

$$\begin{gathered}\rho_n \Phi(T_k(\varrho_n),\tilde{\vartheta}) \to \overline{\rho\Phi(T_k(\varrho),\tilde{\vartheta})} \\ \text{in } C([0,T]; L^\gamma_{weak}(\Omega)) \text{ as well as in } L^2(0,T;W^{-1,2}(\Omega)).\end{gathered} \tag{5.29}$$

Now we can write

$$\begin{aligned}\int_0^T\int_\Omega \varrho_n\Phi(\varrho_n,\vartheta)(\vartheta_n-\vartheta)\varphi\,\mathrm{d}x\,\mathrm{d}t &= \int_0^T\int_\Omega [\varrho_n\Phi(\varrho_n,\vartheta)-\varrho_n\Phi(T_k(\varrho_n),\tilde{\vartheta})](\vartheta_n-\vartheta)\varphi\,\mathrm{d}x\,\mathrm{d}t \\ &+ \int_0^T\int_\Omega [\varrho_n\Phi(T_k(\varrho_n),\tilde{\vartheta}) - \overline{\varrho\Phi(T_k(\varrho),\tilde{\vartheta})}(\vartheta_n-\vartheta)\varphi\,\mathrm{d}x\,\mathrm{d}t \\ &+ \int_0^T\int_\Omega \overline{\varrho\Phi(T_k(\varrho),\tilde{\vartheta})}(\vartheta_n-\vartheta)\varphi\,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

Consequently, in view of (5.3), (5.20), and (5.29), in order to show (5.17), it is enough to find, for arbitrary $\varepsilon>0$, a function $\tilde{\vartheta}$ satisfying (5.28) and k large enough so that

$$\sup_{n=1,2,\dots} \big\|\varrho_n\big(\Phi(\varrho_n,\vartheta)-\Phi(T_k(\varrho_n),\tilde{\vartheta})\big)\big\|_{L^2(0,T;W^{-1,2}(\Omega))} < \varepsilon. \tag{5.30}$$

To this end, we compute

$$\Phi(T_k(\varrho_n),\vartheta)-\Phi(T_k(\varrho_n),\tilde{\vartheta}) = \int_{\tilde{\vartheta}}^{\vartheta} \frac{c_v(T_k(\varrho_n),s)}{s}\,\mathrm{d}s,$$

whence, in accordance with hypothesis (3.10),

$$|\Phi(T_k(\varrho_n),\vartheta)-\Phi(T_k(\varrho_n),\tilde{\vartheta})| \le c\big(|\vartheta-\tilde{\vartheta}| + |\log(\vartheta)-\log(\tilde{\vartheta})|\big)$$

with the constant c independent of n, k. From the last estimate, we infer

$$\begin{aligned}&\|\varrho_n(\Phi(T_k(\varrho_n),\vartheta)-\Phi(T_k(\varrho_n),\tilde{\vartheta}))\|_{L^q(\Omega)} \le c\|\rho_n\|_{L^\gamma(\Omega)}(\|\vartheta-\tilde{\vartheta}\|_{L^6(\Omega)} \\ &\quad + \|\log(\vartheta)-\log(\tilde{\vartheta})\|_{L^6(\Omega)}), \qquad \frac{q}{\gamma}+\frac{q}{6}\le 1.\end{aligned}$$

Consequently, as ϑ, $\log(\vartheta)$ belong to the space $L^2(0,T;W^{1,2}(\Omega))$ induced by estimates (4.22), and ϱ_n satisfy (4.8) with $\gamma>3/2$, it is a routine matter to find $\tilde{\vartheta}$ such that, for a given $\varepsilon>0$,

$$\sup_{n=1,2,\dots} \big\|\varrho_n\big(\Phi(T_k(\varrho_n),\vartheta)-\Phi(T_k(\varrho_n),\tilde{\vartheta})\big)\big\|_{L^2(0,T;W^{-1,2}(\Omega))} < \frac{\varepsilon}{2}. \tag{5.31}$$

On the other hand, by virtue of Hölder's inequality,

$$\begin{aligned}&\big\|\varrho_n\big(\Phi(\varrho_n,\vartheta)-\Phi(T_k(\varrho_n),\vartheta)\big)\big\|_{L^q(\Omega)} \\ &\quad \le c\|\varrho_n\|_{L^p(\{\varrho_n>k\})}\|(1+\vartheta+|\log(\vartheta)|)\|_{L^r(\Omega)}, \qquad \frac{1}{q}=\frac{1}{p}+\frac{1}{r},\end{aligned} \tag{5.32}$$

where we can choose

$$q > \frac{6}{5},\ p < \gamma,\ r < 6 \text{ provided } \gamma > \frac{3}{2}.$$

Since

$$\|\varrho_n\|_{L^p(\{\varrho_n > k\})} \le k^{\frac{p-\gamma}{p}} \|\varrho_n\|^{\frac{\gamma}{p}}_{L^\gamma(\Omega)}, \quad W^{1,2}(\Omega) \hookrightarrow L^r(\Omega), \quad \text{and} \quad L^q(\Omega) \hookrightarrow W^{-1,2}(\Omega),$$

relations (5.31) and (5.32) yield (5.30) under the condition that $k = k(\varepsilon)$ is large enough.

Thus, we have shown (5.17) and, consequently, the proof of (5.20) is now complete.

5.3. The effective viscous flux. In order to show pointwise convergence of the sequence of densities $\{\varrho_n\}_{n=1}^\infty$, we evoke the celebrated "weak continuity" property of the so-called *effective viscous flux*,

$$p - (\lambda + 2\mu)\mathrm{div}_x \mathbf{u}, \text{ with the standard notation } \lambda \equiv \zeta - \frac{2}{3}\mu,$$

established for the barotropic fluids with constants μ and λ in [24] and extended to the general case $\mu = \mu(\vartheta)$, $\lambda = \lambda(\vartheta)$ in [17].

We start with an integral identity

$$\int_0^T \int_\Omega \psi\big(\eta\ p(\varrho_n, \vartheta_n) - \mathcal{R}_{i,j}[\eta\ S_n^{i,j}]\big)\xi T_k(\varrho_n)\mathrm{d}x\,\mathrm{d}t = \sum_{j=1}^7 I_j \tag{5.33}$$

$$+ \int_0^T \int_\Omega \psi u_n^j\big(\xi T_k(\varrho_n)\mathcal{R}_{i,j}[\eta\varrho_n u_n^i] - \eta\varrho_n u_n^i \mathcal{R}_{i,j}[\xi T_k(\varrho_n)]\big)\mathrm{d}x\,\mathrm{d}t,$$

where

$$I_1 = \int_0^T \int_\Omega \psi\ \partial_{x_j}\eta\ S_n^{i,j}(\partial_{x_i}\Delta^{-1})[\xi T_k(\varrho_n)]\mathrm{d}x\,\mathrm{d}t,$$

$$I_2 = -\int_0^T \int_\Omega \psi\eta\ \varrho_n f_n^i(\partial_{x_i}\Delta^{-1})[\xi T_k(\varrho_n)]\mathrm{d}x\,\mathrm{d}t,$$

$$I_3 = -\int_0^T \int_\Omega \psi\ \partial_{x_j}\eta\ \varrho_n u_n^i u_n^j(\partial_{x_i}\Delta^{-1})[\xi T_k(\varrho_n)]\mathrm{d}x\,\mathrm{d}t,$$

$$I_4 = -\int_0^T \int_\Omega \partial_t\psi\eta\ \varrho_n u_n^i(\partial_{x_i}\Delta^{-1})[\xi T_k(\varrho_n)]\mathrm{d}x\,\mathrm{d}t,$$

$$I_5 = -\int_0^T \int_\Omega \psi\eta\varrho_n u_n^i(\partial_{x_i}\Delta^{-1})[T_k(\varrho_n)\mathbf{u}_n \cdot \nabla_x\xi]\mathrm{d}x\,\mathrm{d}t,$$

$$I_6 = \int_0^T \int_\Omega \psi\eta\ \varrho_n u_n^i(\partial_{x_i}\Delta^{-1})\big[\xi\big(T_k'(\varrho_n)\varrho_n - T_k(\varrho_n)\big)\mathrm{div}_x\mathbf{u}_n\big]\mathrm{d}x\,\mathrm{d}t,$$

$$I_7 = -\int_0^T \int_\Omega \psi\ p_n\partial_{x_i}\eta(\partial_{x_i}\Delta^{-1})[\xi T_k(\varrho_n)]\mathrm{d}x\,\mathrm{d}t,$$

$$S_n^{i,j} \equiv \mu(\vartheta_n)\big(\partial_{x_i}u_n^j + \partial_{x_j}u_n^i\big) + \lambda(\vartheta_n)\,\mathrm{div}_x\mathbf{u}_n\ \delta_{i,j},$$

and $\mathcal{R}_{i,j} = \partial_{x_i}\Delta^{-1}\partial_{x_j}$ is a pseudodifferential operator corresponding to the Fourier symbol $\xi_i\xi_j/|\xi|^2$. Formula (5.33) can be obtained through the choice of test functions

$$\varphi_i(t,x) = \psi(t)\eta(x)\partial_{x_i}\Delta^{-1}[\xi T_k(\varrho_n)],\ \psi \in \mathcal{D}(0,T),\ \eta,\ \xi \in \mathcal{D}(\Omega)$$

in the variational formulation of (1.2) (see Lemma 5.3 in Chapter 5 in [16]).

On the other hand, one can take

$$\varphi(t,x) = \psi\eta(\partial_{x_i}\Delta^{-1})[\overline{T_k(\varrho)}], \quad i = 1,2,3,$$

as test functions for (5.8), where

$$\overline{T_k(\rho)} \text{ is a limit of } T_k(\rho_n) \text{ in } C([0,T]; L^p_{weak}(\Omega)), \quad 1 < p < \infty \tag{5.34}$$

(cf. (5.22), (5.23)), to deduce

$$\begin{aligned} &\int_0^T \int_\Omega \psi\big(\eta\ \overline{p(\varrho,\vartheta)} - \mathcal{R}_{i,j}[\eta\ S^{i,j}]\big)\xi\overline{T_k(\varrho)}\,\mathrm{d}x\,\mathrm{d}t \\ &\quad = \sum_{j=1}^7 I_j \int_0^T \int_\Omega \psi u^j \big(\xi\overline{T_k(\varrho)}\mathcal{R}_{i,j}[\eta\varrho u^i] - \eta\varrho u^i \mathcal{R}_{i,j}[\xi\overline{T_k(\varrho)}]\big)\,\mathrm{d}x\,\mathrm{d}t, \end{aligned} \tag{5.35}$$

where

$$\begin{aligned} I_1 &= \int_0^T \int_\Omega \psi\ \partial_{x_j}\eta\ S^{i,j}(\partial_{x_i}\Delta^{-1})[\xi\overline{T_k(\varrho)}]\mathrm{d}x\ \mathrm{d}t, \\ I_2 &= -\int_0^T \int_\Omega \psi\eta\ \overline{\varrho f^i}(\partial_{x_i}\Delta^{-1})[\xi\overline{T_k(\varrho)}]\mathrm{d}x\,\mathrm{d}t, \\ I_3 &= -\int_0^T \int_\Omega \psi\ \partial_{x_j}\eta\ \varrho u^i u^j(\partial_{x_i}\Delta^{-1})[\xi\overline{T_k(\varrho)}]\mathrm{d}x\,\mathrm{d}t, \\ I_4 &= -\int_0^T \int_\Omega \partial_t\psi\eta\ \varrho u^i(\partial_{x_i}\Delta^{-1})[\xi\overline{T_k(\varrho)}]\mathrm{d}x\,\mathrm{d}t, \\ I_5 &= -\int_0^T \int_\Omega \psi\eta\varrho u^i(\partial_{x_i}\Delta^{-1})[\overline{T_k(\varrho)}\mathbf{u}\cdot\nabla_x\xi]\mathrm{d}x\,\mathrm{d}t, \\ I_6 &= \int_0^T \int_\Omega \psi\eta\ \varrho u^i(\partial_{x_i}\Delta^{-1})\big[\xi\overline{\big(T_k'(\varrho)\varrho - T_k(\varrho)\big)\mathrm{div}_x\mathbf{u}}\big]\mathrm{d}x\,\mathrm{d}t, \\ I_7 &= -\int_0^T \int_\Omega \psi\ \overline{p(\varrho,\vartheta)}\partial_{x_i}\eta(\partial_{x_i}\Delta^{-1})[\xi\overline{T_k(\varrho)}]\mathrm{d}x\,\mathrm{d}t. \end{aligned}$$

Note that

$$\partial_t\overline{T_k(\varrho)} + \mathrm{div}_x(\overline{T_k(\varrho)}\mathbf{u}) + \overline{\big(T_k'(\varrho)\varrho - T_k(\varrho)\big)\mathrm{div}_x\mathbf{u}} = 0 \text{ in } \mathcal{D}'((0,T)\times\Omega),$$

and, in accordance with (5.21),

$$S^{i,j} = \mu(\vartheta)\big(\partial_{x_i}u^j + \partial_{x_j}u^i\big) + \lambda(\vartheta)\mathrm{div}_x\mathbf{u}\ \delta_{i,j}.$$

It can be checked, in view of the results established in sections 5.2 and 5.3, that all the integrals $I_1,\dots,I_7$ on the right-hand side of (5.33) converge for $n\to\infty$ to their counterparts in (5.35) (see, for example, section 6.3 in Chapter 6 in [16]).

Moreover, by virtue of Corollary 6.1 in [16], the bilinear form

$$\left\{ \begin{array}{c} [\mathbf{v},\mathbf{w}] \mapsto v^i\mathcal{R}_{i,j}[w^j] - w^i\mathcal{R}_{i,j}[v^j] \text{ is weakly} \\ \text{sequentially continuous with values in } [L^s(R^N)]^N_{weak} \text{ on the} \\ \text{product } [L^p(R^N)]^N \times [L^q(R^N)]^N\ ,\ \frac{1}{p}+\frac{1}{q}=\frac{1}{s}<1,\ s \text{ finite.} \end{array} \right\} \tag{5.36}$$

Thus one can use (5.5), (5.34), and (5.36) in order to obtain

$$\begin{aligned}&\big(\xi T_k(\varrho_n)\mathcal{R}_{i,j}[\eta\varrho_n u_n^i] - \eta\varrho_n u_n^i\mathcal{R}_{i,j}[\xi T_k(\varrho_n)]\big)\\ &\quad\to \big(\xi\overline{T_k(\varrho)}\mathcal{R}_{i,j}[\eta\varrho u^i] - \eta\varrho u^i\mathcal{R}_{i,j}[\xi\overline{T_k(\varrho)}]\big)(t)\\ &\quad\text{weakly in } L^s(\Omega), \quad 1 < s < \frac{2\gamma}{\gamma+1}, \quad t \in [0,T].\end{aligned}$$

As the imbedding $L^s(\Omega) \hookrightarrow W^{-1,2}(\Omega)$ is compact, we infer

$$\begin{aligned}&\big(\xi T_k(\varrho_n)\mathcal{R}_{i,j}[\eta\varrho_n u_n^i] - \eta\varrho_n u_n^i\mathcal{R}_{i,j}[\xi T_k(\varrho_n)]\big)\\ &\quad\to \big(\xi\overline{T_k(\varrho)}\mathcal{R}_{i,j}[\eta\varrho u^i] - \eta\varrho u^i\mathcal{R}_{i,j}[\xi\overline{T_k(\varrho)}]\big) \text{ in } L^2(0,T;W^{-1,2}(\Omega)).\end{aligned}$$

Subtracting the limit as $n \to \infty$ of (5.33) and (5.35), we finally get

$$\begin{aligned}\lim_{n\to\infty}&\int_0^T\int_\Omega \psi\big(\eta\, p(\varrho_n,\vartheta_n) - \mathcal{R}_{i,j}[\eta\, S_n^{i,j}]\big)\xi T_k(\varrho_n)\mathrm{d}x\,\mathrm{d}t\\ &= \int_0^T\int_\Omega \psi\big(\eta\,\overline{p(\varrho,\vartheta)} - \mathcal{R}_{i,j}[\eta\, S^{i,j}]\big)\xi\overline{T_k(\varrho)}\mathrm{d}x\,\mathrm{d}t.\end{aligned} \tag{5.37}$$

In order to conclude, let us recall the following result that may be viewed as a variant of the abstract theory developed by Coifman et al. [7] (see Lemma 4.2 in [17]).

LEMMA 5.3. *Let*

$$\mathbf{V} \in L^2(R^N;R^N), \quad w \in W^{1,r}(R^N), \quad r > \frac{2N}{N+2}.$$

Then there exist constants $c = c(r) > 0$, $\omega = \omega(r) \in (0,1)$, $p = p(r) > 1$ such that

$$\|\mathcal{R}_{i,j}[wV_j] - w\mathcal{R}_{i,j}[V_j]\|_{W^{\omega,p}(R^N;R^N)} \le c(r)\|w\|_{W^{1,r}(R^N)}\|\mathbf{V}\|_{L^2(R^N;R^N)}, \; i = 1,\dots,N.$$

Due to (3.12), (3.13), (3.16), and (4.22) and (5.3) and (5.20), we have at least for a chosen subsequence

$$a(\vartheta_n) \to a(\vartheta) \text{ weakly in } L^2(0,T;W^{1,2}(\Omega)), \tag{5.38}$$

$$a(\vartheta_n) \to a(\vartheta) \text{ in } L^q((0,T)\times\Omega), \quad 1 \le q < 3, \tag{5.39}$$

where a stands for μ and λ. Recalling that $\mathcal{R}_{i,j}$ is a continuous linear operator from $L^q(R^N)$ to $L^q(R^N)$, $1 < q < \infty$, and taking into account (4.6) and (4.30), we verify that

$$\mathcal{R}_{i,j}[a(\vartheta_n)\partial_j\mathbf{u}_n] - a(\vartheta_n)\mathcal{R}_{i,j}[\partial_j\mathbf{u}_n] \text{ is bounded in } L^2(0,T;L^{\frac{6}{5}}(\Omega,R^3)). \tag{5.40}$$

On the other hand, using Lemma 5.3 with $w = a(\vartheta)$, $\mathbf{V} = \nabla u_k$, $k = 1,2,3$, we get that

$$\mathcal{R}_{i,j}[a(\vartheta_n)\partial_j\mathbf{u}_n] - a(\vartheta_n)\mathcal{R}_{i,j}[\partial_j\mathbf{u}_n] \text{ is bounded in } L^1(0,T;W^{\omega',s}(\Omega,R^3)) \tag{5.41}$$

with suitable numbers $\omega' \in (0,1)$ and $s > 1$. Now we interpolate (5.40) and (5.41) to get the boundedness of the sequence

$$\mathcal{R}_{i,j}[a(\vartheta_n)\partial_j\mathbf{u}_n] - a(\vartheta_n)\mathcal{R}_{i,j}[\partial_j\mathbf{u}_n]_; \text{ in } L^p(0,T;W^{\omega,p}(\Omega,R^3))$$

with suitable exponents $p > 1$ and $\omega \in (0,1)$. Finally, since according to (5.2), (5.38), and (5.40)

$$\mathcal{R}_{i,j}[a(\vartheta_n)\partial_j \mathbf{u}_n] - a(\vartheta_n)[\mathcal{R}_{i,j}\partial_j \mathbf{u}_n]$$
$$\to \mathcal{R}_{i,j}[a(\vartheta)\partial_j \mathbf{u}] - a(\vartheta)\mathcal{R}_{i,j}[\partial_j \mathbf{u})] \text{ weakly in } L^2(0,T;L^{\frac{6}{5}}(\Omega,R^3)),$$

from the last bound we deduce

$$\mathcal{R}_{i,j}[a(\vartheta_n)\partial_j \mathbf{u}_n] - a(\vartheta_n)\mathcal{R}_{i,j}[\partial_j \mathbf{u}_n]$$
$$\to \mathcal{R}_{i,j}[a(\vartheta)\partial_j \mathbf{u}] - a(\vartheta)\mathcal{R}_{i,j}[\partial_j \mathbf{u}] \text{ weakly in } L^p(0,T;W^{\omega,p}(\Omega;R^3)) \tag{5.42}$$

at least for a suitably chosen subsequence. On the other hand, as for a suitable $q \in (1,\infty)$ the imbedding $L^q(\Omega) \hookrightarrow W^{-\omega,p'}(\Omega)$ is compact, formula (5.34) implies

$$T_k(\rho_n) \to \overline{T_k(\rho)} \text{ in } L^{p'}(0,T;W^{-\omega,p'}(\Omega)). \tag{5.43}$$

Employing the obvious properties of the Riesz operator, namely, $\mathcal{R}_{i,j} = \mathcal{R}_{j,i}$ and $\mathcal{R}_{i,j}(\frac{\partial u_j}{\partial x_i}) = \operatorname{div}_x \mathbf{u}$, and using (5.42) and (5.43) in (5.37), we arrive at

$$\lim_{n\to\infty} \int_0^T \int_\Omega \varphi\big(p(\varrho_n,\vartheta_n) - (\lambda(\vartheta_n) + 2\mu(\vartheta_n)) \operatorname{div}_x \mathbf{u}_n\big) T_k(\varrho_n) \mathrm{d}x\, \mathrm{d}t \tag{5.44}$$
$$= \int_0^T \int_\Omega \varphi\big(\overline{p(\varrho,\vartheta)} - (\lambda(\vartheta) + 2\mu(\vartheta))\operatorname{div}_x \mathbf{u}\big)\overline{T_k(\varrho)} \mathrm{d}x\, \mathrm{d}t \times \Omega)$$

for any $\varphi \in \mathcal{D}(0,T)$. For more details see section 4 in [17].

5.4. Strong convergence of the density. To begin with, we show that the limit functions ϱ, $\mathbf{u}$ solve the renormalized continuity equation (2.5). Note that such a result cannot be obtained via the regularization procedure developed in [10] as the density is not (known to be) square integrable. Instead, following Chapter 6 in [16], we introduce the *oscillation defect measure*

$$\mathbf{osc}_p[\varrho_n \to \varrho](Q) \equiv \sup_{k\geq 1} \Big\{ \limsup_{n\to\infty} \int_Q |T_k(\varrho_n) - T_k(\varrho)|^p \,\mathrm{d}x \ \mathrm{d}t \Big\}.$$

We report the following assertion (see Proposition 6.3 in Chapter 6 in [16]).

LEMMA 5.4. *Let $\Omega \subset R^N$, $N = 2,3$, be a domain. Assume that $\{\varrho_n\}_{n=1}^\infty$ is a sequence of nonnegative functions such that*

$$\varrho_n \to \varrho \textit{ weakly (*) in } L^\infty(0,T;L^\gamma(\Omega)),\ \gamma > \frac{2N}{N+2}.$$

Furthermore, let ϱ_n satisfy the renormalized equation (2.5) *with* $\mathbf{u}_n$,

$$\mathbf{u}_n \to \mathbf{u} \textit{ weakly in } L^2(0,T;W_0^{1,2}(\Omega;R^N)).$$

Finally, suppose

$$\mathbf{osc}_p[\varrho_n \to \varrho](Q) < c(Q) < \infty \textit{ for some } p > 2,$$

for any bounded $Q \subset (0,T) \times \Omega$.

Then ϱ, $\mathbf{u}$ solve the renormalized equation (2.5) *for any h, H satisfying* (2.6).

In the following lemma we show that the assumptions of Lemma 5.4 are satisfied.

LEMMA 5.5. *Under the hypotheses of Theorem* 3.1*, we have*

$$\mathbf{osc}_{\gamma+1}[\varrho_n \to \varrho]((0,T)\times\Omega) < \infty.$$

Proof. As already observed in (4.24), one can write

$$p(\varrho_n,\vartheta_n) = a\varrho_n^\gamma + p_b(\varrho_n,\vartheta_n), \text{ where } |p_b(\varrho,\vartheta)| \le c(1+\vartheta^4+\vartheta\varrho^{\frac{\gamma}{3}}). \tag{5.45}$$

Now relation (5.44) may be rephrased as

$$\begin{aligned} a\lim_{n\to\infty}\int_0^T\int_\Omega &\varrho_n^\gamma T_k(\varrho_n) - \overline{\varrho^\gamma}\;\overline{T_k(\varrho)}\mathrm{d}x\,\mathrm{d}t \\ \le \lim_{n\to\infty}\int_0^T\int_\Omega &p_b(\varrho_n,\vartheta_n)(\overline{T_k(\varrho)} - T_k(\varrho_n))\mathrm{d}x\,\mathrm{d}t \\ + \lim_{n\to\infty}\int_0^T\int_\Omega &(\lambda(\vartheta_n)+2\mu(\vartheta_n))\mathrm{div}_x\mathbf{u}_n(T_k(\varrho_n)-\overline{T_k(\varrho)})\mathrm{d}x\,\mathrm{d}t. \end{aligned} \tag{5.46}$$

Exactly as in the proof of Proposition 6.2 in [16], via arguments based on the convexity of $\varrho\mapsto\varrho^\gamma$, the left-hand side of (5.46) is bounded from below as follows:

$$\begin{aligned} \limsup_{n\to\infty}\int_0^T\int_\Omega &|T_k(\varrho_n)-T_k(\varrho)|^{\gamma+1}\,\mathrm{d}x\,\mathrm{d}t \\ \le \lim_{n\to\infty}\int_0^T\int_\Omega &\varrho_n^\gamma T_k(\varrho_n) - \overline{\varrho^\gamma}\;\overline{T_k(\varrho)}\,\mathrm{d}x\,\mathrm{d}t. \end{aligned} \tag{5.47}$$

On the other hand, by virtue of hypothesis (3.16), estimates (4.8), (4.22), (4.32), and relation (4.26),

$$\sup_{n=1,2,\dots}\|p_b(\varrho_n,\vartheta_n)\|_{L^{\frac{5}{3}}((0,T)\times\Omega)} < \infty$$

and

$$\sup_{n=1,2,\dots}\|(\lambda(\vartheta_n)+2\mu(\vartheta_n))\mathrm{div}_x\mathbf{u}_n\|_{L^{\frac{5}{3}}((0,T)\times\Omega)} < \infty.$$

Thus,

$$\begin{aligned} \lim_{n\to\infty}\int_0^T\int_\Omega &p_b(\varrho_n,\vartheta_n)(\overline{T_k(\varrho)}-T_k(\varrho_n))\,\mathrm{d}x\,\mathrm{d}t \\ + \lim_{n\to\infty}\int_0^T\int_\Omega &(\lambda(\vartheta_n)+2\mu(\vartheta_n))\mathrm{div}_x\mathbf{u}_n(T_k(\varrho_n)-\overline{T_k(\varrho)})\,\mathrm{d}x\,\mathrm{d}t \\ \le c\limsup_{n\to\infty}&\|T_k(\varrho_n)-T_k(\varrho)\|_{L^{\frac{5}{2}}((0,T)\times\Omega)}, \end{aligned}$$

and, consequently, relations (5.46) and (5.47) yield the desired conclusion since $\gamma+1>5/2$. ∎

Combining Lemmas 5.4 and 5.5 we infer that ϱ, $\mathbf{u}$ solve (2.5); in particular,

$$\partial_t L_k(\varrho) + \mathrm{div}_x(L_k(\varrho)\mathbf{u}) + T_k(\varrho)\mathrm{div}_x\mathbf{u} = 0 \text{ in } \mathcal{D}'((0,T)\times R^3) \tag{5.48}$$

provided ϱ, $\mathbf{u}$ were extended to be zero outside Ω with

$$L_k(\varrho) \equiv \varrho \int_1^{\varrho} \frac{T_k(z)}{z^2}\,\mathrm{d}z.$$

Moreover, as ϱ_n, $\mathbf{u}_n$ solve (2.5) as well, it is easy to deduce

$$\partial_t \overline{L_k(\varrho)} + \mathrm{div}_x(\overline{L_k(\varrho)}\mathbf{u}) + \overline{T_k(\varrho)\,\mathrm{div}_x\mathbf{u}} = 0 \text{ in } \mathcal{D}'((0,T)\times R^3). \tag{5.49}$$

Equations (5.48) and (5.49) together with hypothesis (3.18) give rise to

$$\begin{aligned}
&\int_\Omega \big(\overline{L_k(\varrho)} - L_k(\varrho)\big)(\tau)\mathrm{d}x \\
&\quad + \int_0^\tau \int_\Omega \big(\overline{T_k(\varrho)\,\mathrm{div}_x\mathbf{u}} - \overline{T_k(\varrho)}\,\mathrm{div}_x\mathbf{u}\big)\,\mathrm{d}x\,\mathrm{d}t \\
&\quad = \int_0^\tau \int_\Omega \big(T_k(\varrho) - \overline{T_k(\varrho)}\big)\mathrm{div}_x\mathbf{u}\,\mathrm{d}x\,\mathrm{d}t \quad \text{for all } \tau \in [0,T].
\end{aligned} \tag{5.50}$$

Now it can be shown, with the help of Lemma 5.5, that

$$\int_0^T \int_\Omega \big(T_k(\varrho) - \overline{T_k(\varrho)}\big)\mathrm{div}_x\mathbf{u}\,\mathrm{d}x\,\mathrm{d}t \to 0 \text{ for } k \to \infty$$

(see section 6.6 in [16]).

Moreover, in accordance with (5.44),

$$\big(\overline{T_k(\varrho)\,\mathrm{div}_x\mathbf{u}} - \overline{T_k(\varrho)}\mathrm{div}_x\mathbf{u}\big) = (\lambda(\vartheta) + 2\mu(\vartheta))^{-1}\big(\overline{p(\varrho,\vartheta)T_k(\varrho)} - \overline{p(\varrho,\vartheta)}\;\overline{T_k(\varrho)}\big),$$

where, by virtue of (3.1) and (5.20),

$$\big(\overline{p(\varrho,\vartheta)T_k(\varrho)} - \overline{p(\varrho,\vartheta)}\;\overline{T_k(\varrho)}\big) = \big(\overline{p_G(\varrho,\vartheta)T_k(\varrho)} - \overline{p_G(\varrho,\vartheta)}\;\overline{T_k(\varrho)}\big). \tag{5.51}$$

Since p_G is a nondecreasing function of ϱ and $\{\vartheta_n\}_{n=1}^\infty$ converges almost everywhere on $(0,T)\times\Omega$, the most right expression in (5.51) is nonnegative. Thus, we conclude, using (5.50), that

$$\overline{\varrho\log(\varrho)} = \varrho\log(\varrho) \text{ on } (0,T)\times\Omega,$$

which means

$$\varrho_n \to \varrho \;\text{ in } L^1((0,T)\times\Omega). \tag{5.52}$$

5.5. Convergence. Having established the pointwise convergence of the sequences $\{\varrho_n\}_{n=1}^\infty$, $\{\vartheta_n\}_{n=1}^\infty$, and with the estimates obtained in section 4, it is a routine matter to complete the proof of Theorem 3.1, that is, to pass to the limit for $n\to\infty$ in all the integral identities appearing in the variational formulation of problem (1.1)–(1.6) introduced in Definition 2.1.

Let us only remark that we can write

$$\frac{\mathbf{q}}{\vartheta} = -\kappa(\varrho,\vartheta)\nabla_x\log(\vartheta) \quad \text{and} \quad \frac{\mathbf{q}\cdot\nabla_x\vartheta}{\vartheta^2} = -\kappa(\varrho,\vartheta)|\nabla_x\log(\vartheta)|^2.$$

In particular, by virtue of hypothesis (3.15), and estimates (4.6) and (4.22), the quantities

$$\left|\frac{\mathbf{q}_n}{\vartheta_n}\right| = \sqrt{\kappa(\varrho_n,\vartheta_n)}\sqrt{\kappa(\varrho_n,\vartheta_n)}|\nabla_x\log(\vartheta_n)| \text{ are bounded in } L^p((0,T)\times\Omega)$$

for a certain $p > 1$ and independently of $n = 1, 2, \ldots$.

In order to carry out the limit passage in the entropy production inequality (2.7), we report the following observation (see Lemmas 5.3 and 5.4 in [13]).

LEMMA 5.6.

(i) *Let $\Omega \subset R^N$ be a bounded Lipschitz domain. Suppose that ϱ is a given non-negative function satisfying*

$$0 < M \le \int_\Omega \varrho \,\mathrm{d}x, \quad \int_\Omega \varrho^\gamma \,\mathrm{d}x < K, \quad \gamma > \frac{2N}{N+2}.$$

Furthermore, let $\vartheta \in W^{1,2}(\Omega)$.

Then the following two statements are equivalent.

- *The function ϑ is strictly positive a.a. on Ω,*

$$\varrho|\log(\vartheta)| \in L^1(\Omega) \text{ and } \frac{\nabla_x \vartheta}{\vartheta} \in L^2(\Omega).$$

- *The function $\log(\vartheta)$ belongs to the Sobolev space $W^{1,2}(\Omega)$.*

Moreover, if it is the case, then

$$\nabla_x \log(\vartheta) = \frac{\nabla_x \vartheta}{\vartheta} \text{ a.a. on } \Omega.$$

(ii) *Let*

$$\vartheta_n \to \vartheta \text{ in } L^2((0,T)\times\Omega), \text{ and } \log(\vartheta_n) \to \overline{\log(\vartheta)} \text{ weakly in } L^2((0,T)\times\Omega).$$

Then ϑ is strictly positive a.a. on $(0,T)\times\Omega$, and $\log(\vartheta) = \overline{\log(\vartheta)}$.

Acknowledgment. The authors thank T. Bodnár for his help with the typesetting of figures.

REFERENCES

[1] S. N. ANTONTSEV, A. V. KAZHIKHOV, AND V. N. MONAKHOV, *Krajevyje zadaci mechaniki neodnorodnych zidkostej,* Novosibirsk, 1983.
[2] E. BECKER, *Gasdynamik,* Teubner-Verlag, Stuttgart, 1966.
[3] M. E. BOGOVSKII, *Solution of some vector analysis problems connected with operators* div *and* grad, Tr. Semin. S. L. Sobolev, 80 (1980), pp. 5–40 (in Russian).
[4] W. BORCHERS AND H. SOHR, *On the equation rotv = g and divu = f with zero boundary conditions,* Hokkaido Math. J., 19 (1990), pp. 67–87.
[5] C. BUET AND B. DESPRÉS, *Asymptotic Analysis of Fluid Models for the Coupling of Radiation and Hydrodynamics,* preprint, 2003.
[6] L. CAFFARELLI, R. V. KOHN, AND L. NIRENBERG, *On the regularity of the solutions of the Navier–Stokes equations,* Comm. Pure Appl. Math., 35 (1982), pp. 771–831.
[7] R. COIFMAN, P. L. LIONS, Y. MEYER, AND S. SEMMES, *Compensated compactness and Hardy spaces,* J. Math. Pures Appl., 72 (1993), pp. 247–286.
[8] R. J. DIPERNA, *Convergence of approximate solutions to conservation laws,* Arch. Ration. Mech. Anal., 82 (1983), pp. 27–70.
[9] R. J. DIPERNA, *Convergence of the viscosity method for isentropic gas dynamics,* Comm. Math. Phys., 91 (1983), pp. 1–30.
[10] R. J. DIPERNA AND P.-L. LIONS, *Ordinary differential equations, transport theory and Sobolev spaces,* Invent. Math., 98 (1989), pp. 511–547.
[11] J. DUCHON AND R. ROBERT, *Inertial energy dissipation for weak solutions of incompressible Euler and Navier–Stokes equations,* Nonlinearity, 13 (2000), pp. 249–255.
[12] B. DUCOMET AND E. FEIREISL, *A regularizing effect of radiation in the equations of fluid dynamics,* Math. Methods Appl. Sci., 28 (2005), pp. 661–685.

[13] B. Ducomet and E. Feireisl, *On the dynamics of gaseous stars,* Arch. Ration. Mech. Anal., 174 (2004), pp. 221–266.
[14] S. Eliezer, A. Ghatak, and H. Hora, *An Introduction to Equations of States, Theory and Applications,* Cambridge University Press, Cambridge, UK, 1986.
[15] G. L. Eyink, *Local* 4/5 *law and energy dissipation anomaly in turbulence,* Nonlinearity, 16 (2003), pp. 137–145.
[16] E. Feireisl, *Dynamics of Viscous Compressible Fluids,* Oxford University Press, Oxford, 2003.
[17] E. Feireisl, *On the motion of a viscous, compressible, and heat conducting fluid,* Indiana Univ. Math. J., 53 (2004), pp. 1705–1738.
[18] E. Feireisl and Novotný, *Large time behaviour of flows of compressible, viscous, heat conducting fluids,* Math. Methods Appl. Sci., submitted.
[19] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier–Stokes Equations,* I, Springer-Verlag, New York, 1994.
[20] G. Gallavotti, *Statistical Mechanics: A Short Treatise,* Springer-Verlag, Heidelberg, 1999.
[21] G. Gallavotti, *Foundations of Fluid Dynamics,* Springer-Verlag, New York, 2002.
[22] D. Hoff, *Strong convergence to global solutions for multidimensional flows of compressible, viscous fluids with polytropic equations of state and discontinuous initial data,* Arch. Ration. Mech. Anal., 132 (1995), pp. 1–14.
[23] D. Hoff, *Discontinuous solutions of the Navier–Stokes equations for multidimensional flows of heat conducting fluids,* Arch. Ration. Mech. Anal., 139 (1997), pp. 303–354.
[24] P.-L. Lions, *Mathematical Topics in Fluid Dynamics, Vol.* 2*, Compressible Models,* Oxford Science, Oxford, 1998.
[25] A. Matsumura and T. Nishida, *The initial value problem for the equations of motion of viscous and heat-conductive gases,* J. Math. Kyoto Univ., 20 (1980), pp. 67–104.
[26] A. Matsumura and T. Nishida, *The initial value problem for the equations of motion of compressible and heat conductive fluids,* Comm. Math. Phys., 89 (1983), pp. 445–464.
[27] B. Mihalas and B. Weibel-Mihalas, *Foundations of Radiation Hydrodynamics,* Dover Publication, Dover, 1984.
[28] T. Nagasawa, *A new energy inequality and partial regularity for weak solutions of Navier–Stokes equations,* J. Math. Fluid Mech., 3 (2001), pp. 40–56.
[29] A. Novotný and I. Straškraba, *Introduction to the Theory of Compressible Flow,* Oxford University Press, Oxford, 2004.
[30] K. R. Rajagopal and A. R. Shrinivasa, *On thermodynamical restrictions of continua,* Proc. R. Soc. Lond. Ser. A, 460 (2004), pp. 631–651.
[31] P. Secchi, *On the motion of gaseous stars in presence of radiation,* Comm. Partial Differential Equations, 15 (1990), pp. 185–204.
[32] P. Secchi, *On the uniqueness of motion of viscous gaseous stars,* Math. Methods Appl. Sci., 13 (1990), pp. 391–404.
[33] D. Serre, *Variation de grande amplitude pour la densité d'un fluid visqueux compressible,* Phys. D, 48 (1991), pp. 113–128.
[34] J. Smoller, *Shock Waves and Reaction-Diffusion Equations,* Springer-Verlag, New York, 1967.
[35] L. Tartar, *Compensated compactness and applications to partial differential equations,* in Nonlinear Anal. and Mech., Heriot-Watt Sympos., Res. Notes Math. 39, L. J. Knopps ed., Pitman, Boston, 1975, pp. 136–211.
[36] M. E. Taylor, *Tools for PDE: Pseudodifferential Operators, Paradifferential Operators, and Layer Potentials,* Math. Surveys Monogr. 81, AMS, Providence, RI, 2000.
[37] C. Truesdell and W. Noll, *The Non-Linear Field Theories of Mechanics,* Springer-Verlag, Heidelberg, 2004.
[38] V. A. Vaigant and A. V. Kazhikhov, *On the existence of global solutions to two-dimensional Navier–Stokes equations of a compressible viscous fluid,* Sibirsk. Mat. Zh., 36 (1995), pp. 1283–1316 (in Russian).
[39] G. J. Van Wylen and R. E. Sonntag, *Fundamentals of Classical Thermodynamics,* John Wiley, New York, 1985.
[40] Y. B. Zel'dovich and Y. P. Raizer, *Physics of Shock Waves and High-Temperature Hydrodynamic Phenomena,* Academic Press, New York, 1966.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 651–672

EXISTENCE, UNIQUENESS, AND REGULARITY RESULTS FOR PIEZOELECTRIC SYSTEMS*

D. MERCIER† AND S. NICAISE†

Abstract. We investigate the time-harmonic piezoelectric system (a system coupling the elasticity system with the full Maxwell's equations) in polyhedral domains of the space. Existence and uniqueness results of weak solutions are proved in different cases. We describe the corner and edge singularities of that system and deduce some regularity results.

Key words. elasticity system, Maxwell's system, singularities

AMS subject classifications. 35J25, 35Q60

DOI. 10.1137/040617728

1. Introduction. Smart structures made of piezoelectric and/or piezomagnetic materials are gaining attention in applications since they are able to transform the energy from one type to another (magnetic, electric, and mechanical), allowing them to be used as sensors and/or actuators. Commonly used piezoelectric materials are ceramics and quartz. The mathematical model of this system starts to be well established [2, 8, 14, 24, 26] and corresponds to a coupling between the elasticity system and Maxwell's equations (see below). A full mathematical analysis is not yet done, except in some particular cases [13, 19]. Namely, in these two works the electric field E is assumed to be curl free, i.e., $E = \nabla\varphi$, where φ is an electric potential and a two-dimensional reduction is made. In [13], existence and uniqueness results in smooth domains are obtained using integral equations, while in [19] a variational formulation in polygonal domains is given and two-dimensional singularities are briefly described.

On the other hand, there exists an extensive list of papers from mechanics literature describing singularities of some particular piezoelectric materials with a plane crack [25, 27, 30] or along wedges [29]. But to our knowledge, an exact description of corner/edge singularities of the general piezoelectric system in three-dimensional polyhedral domains is not yet obtained. Such a description is very important since piezoelectric ceramics are very brittle, and therefore their fracture behavior must be understood. The knowledge of such singularities also has numerical implications, such as convergence speed.

This paper has, therefore, the following goals: We present a general piezoelectric system, which includes standard models of ceramics like the PZT or the $BaTiO_3$. We further develop some variational formulations which are the natural ones because they lead to solutions in the energy spaces (here called weak solutions). We prove existence and uniqueness results of weak solutions of the time-harmonic system in two different cases: the case when the magnetic permeability matrix is positive definite ($BaTiO_3$) and the case when the magnetic permeability matrix is zero (PZT). In that second case, we even give two different formulations and show that generically they give rise to the same solutions. Moreover, we describe the corner and edge singularities of our

*Received by the editors October 27, 2004; accepted for publication (in revised form) February 22, 2005; published electronically November 9, 2005.

http://www.siam.org/journals/sima/37-2/61772.html

†Université de Valenciennes et du Hainaut Cambrésis, MACS, Institut des Sciences et Techniques de Valenciennes, F-59313 Valenciennes Cedex 9, France (Denis.Mercier@univ-valenciennes.fr, Serge.Nicaise@univ-valenciennes.fr).

general system and deduce some regularity results. Some edge singular exponents are briefly described; more examples will be given in a forthcoming paper [20], where a two-dimensional model and fracture criteria will be considered.

The analysis of more sophisticated models, like coupling between piezoelectric and magnetostrictive materials [26] or piezoelectric and purely elastic materials [29, 9], will be investigated in the future.

The paper is organized as follows. Section 2 introduces the nonstationary model problem and its time-harmonic version. Existence and uniqueness results are given in section 3 when the magnetic permeability matrix is positive definite. Section 4 is devoted to existence and uniqueness results in the case of a zero magnetic permeability matrix. There we consider two different formulations: for the first one (called the E-formulation) the magnetic field H is eliminated, while for the second one (the H-formulation) the electric field E is eliminated. For the latter formulation, like for the eddy current problem, Gauge conditions are necessary. We further show their generic equivalence. After a short description of corner and edge singularities of some useful elliptic systems in section 5, we obtain the corner and edge singularities of our system in section 6. Regularity results are deduced in section 7. Some edge singular exponents are finally presented in section 8.

2. Setting of the problem. Let Ω be a bounded domain of $\mathbb{R}^3$ with a Lipschitz boundary Γ. For the sake of simplicity, we suppose that Γ is piecewise plane and connected and that Ω is simply connected. In this domain, we consider the following nonstationary piezoelectric system of constitutive equations [2, 8, 14, 24]:

$$\sigma_{ij} = a_{ijkl}\gamma_{kl}(u) - e_{kij}E_k \ \forall i, j = 1, 2, 3, \tag{2.1}$$

$$D_i = \epsilon_{ij}E_j + e_{ikl}\gamma_{kl}(u) \ \forall i = 1, 2, 3, \tag{2.2}$$

$$B_i = \mu_{ij}H_j \ \forall i = 1, 2, 3. \tag{2.3}$$

The equations of equilibrium are

$$\partial_t^2 u_i = \partial_j \sigma_{ji} + f_i \ \forall i = 1, 2, 3 \tag{2.4}$$

for the elastic displacement and

$$\partial_t D + J = \mathbf{curl}\, H, \partial_t B = -\,\mathbf{curl}\, E$$

for the electric/magnetic fields (Maxwell's equations), where f is the body force density and J is the vector current density function. As usual $\mathbf{curl}\, H = (\partial_2 H_3 - \partial_3 H_2, \partial_3 H_1 - \partial_1 H_3, \partial_1 H_2 - \partial_2 H_1)^\top$, when $H = (H_1, H_2, H_3)^\top$.

This system models the coupling between Maxwell's system and the elastic one [2, 8, 14, 24], in which $E(x,t)$, $H(x,t)$ are the electric and magnetic fields at the point $x \in \Omega$ at time t, $u(x,t)$ is the displacement field at the point $x \in \Omega$ at time t, and $(\gamma_{ij}(u))_{i,j=1}^3$ is the strain tensor given by

$$\gamma_{ij}(u) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right).$$

Here $(\sigma_{ij})_{i,j=1}^3$, $D = (D_1, D_2, D_3)^\top$, and $B = (B_1, B_2, B_3)^\top$ are the stress tensor, electric displacement, and magnetic induction, respectively. ϵ, μ are the electric permittivity and magnetic permeability, respectively, and are supposed to be real, symmetric 3×3 matrices. In what follows the matrix ϵ is supposed to be positive definite, while μ is only supposed to be nonnegative. The elasticity tensor $(a_{ijkl})_{i,j,k,l=1,2,3}$ is made of constant entries such that

$$a_{ijkl} = a_{jikl} = a_{klij},$$

and satisfies the ellipticity condition

$$a_{ijkl}\gamma_{ij}\gamma_{kl} \geq \alpha\gamma_{ij}\gamma_{ij}, \tag{2.5}$$

for every symmetric tensor (γ_{ij}) and some $\alpha > 0$. The piezoelectric tensor e_{kij} is also made of constant entries such that

$$e_{kij} = e_{kji}.$$

The system is completed with the Dirichlet boundary conditions for the displacement field,

$$u = 0 \text{ on } \Gamma, \tag{2.6}$$

and those of a perfect conductor,

$$E \times n = 0 \text{ on } \Gamma. \tag{2.7}$$

As usual n is the exterior unit normal vector along Γ.

In order to write the above problem in a more compact form, the strain and stress tensors are expressed as 6×1 vectors, namely

$$\begin{aligned} \gamma(u) &= (\gamma_{11}(u), \gamma_{22}(u), \gamma_{33}(u), 2\gamma_{23}(u), 2\gamma_{31}(u), 2\gamma_{12}(u))^\top, \\ \sigma &= (\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{23}, \sigma_{31}, \sigma_{12})^\top. \end{aligned}$$

With this notation, the constitutive equations (2.1)–(2.3) may be equivalently written as

$$\begin{pmatrix} \sigma \\ D \\ B \end{pmatrix} = M \begin{pmatrix} \gamma(u) \\ E \\ H \end{pmatrix}, \tag{2.8}$$

where M is a 12×12 matrix given by

$$M = \begin{pmatrix} C & -e^\top & 0 \\ e & \epsilon & 0 \\ 0 & 0 & \mu \end{pmatrix},$$

where C is a 6×6 symmetric matrix depending on the elasticity tensor, e is a 3×6 matrix depending on the piezoelectric tensor, and ϵ, μ are as described above. Note that the ellipticity assumption (2.5) is equivalent to the fact that C is a positive definite matrix.

For a monoclinic material with poling direction in the x_3-axis, the material constant matrices are expressed by (see, for instance, [26])

$$C = \begin{pmatrix} c_{11} & c_{12} & c_{13} & 0 & 0 & c_{16} \\ c_{12} & c_{22} & c_{23} & 0 & 0 & c_{26} \\ c_{13} & c_{23} & c_{33} & 0 & 0 & c_{36} \\ 0 & 0 & 0 & c_{44} & c_{45} & 0 \\ 0 & 0 & 0 & c_{45} & c_{55} & 0 \\ c_{16} & c_{26} & c_{36} & 0 & 0 & c_{66} \end{pmatrix}, \quad e = \begin{pmatrix} 0 & 0 & 0 & e_{14} & e_{15} & 0 \\ 0 & 0 & 0 & e_{24} & e_{25} & 0 \\ e_{31} & e_{32} & e_{33} & 0 & 0 & e_{36} \end{pmatrix},$$

$$\epsilon = \begin{pmatrix} \epsilon_{11} & \epsilon_{12} & 0 \\ \epsilon_{12} & \epsilon_{22} & 0 \\ 0 & 0 & \epsilon_{33} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_{11} & \mu_{12} & 0 \\ \mu_{12} & \mu_{22} & 0 \\ 0 & 0 & \mu_{33} \end{pmatrix}.$$

For the orthotropic piezoelectric PZT-4, the material coefficients are given by (c_{ij} in $10^9 N/m^2$, e_{ij} in C/m^2, ϵ_{ij} in $10^{-9}C^2/Nm^2$):

$$\begin{array}{ccccc} c_{11}=c_{22}=23.8 & c_{33}=10.6 & c_{44}=2.15 & c_{55}=4.4 & c_{66}=\frac{c_{11}-c_{12}}{2}=6.43 \\ c_{12}=3.98 & c_{13}=2.19 & c_{23}=1.92 & \epsilon_{11}=\epsilon_{22}=0.110625 & \epsilon_{33}=0.106023 \\ e_{31}=-0.13 & e_{32}=-0.14 & e_{33}=-0.28 & e_{24}=e_{15}=-0.01. & \end{array}$$

Here and below, nongiven coefficients are equal to zero.

For the piezoelectric $BaTiO_3$, the material coefficients are given by (c_{ij} in $10^9 N/m^2$, e_{ij} in C/m^2, ϵ_{ij} in $10^{-9}C^2/Nm^2$, μ_{ij} in $10^{-6}Ns^2/C^2$):

$$\begin{array}{ccccc} c_{11}=c_{22}=166 & c_{33}=162 & c_{44}=c_{55}=43 & c_{66}=\frac{c_{11}-c_{12}}{2}=44.5 & \\ c_{12}=77 & c_{13}=c_{23}=78 & \epsilon_{11}=\epsilon_{22}=11.2 & \epsilon_{33}=12.6 & \\ e_{31}=e_{32}=-4.4 & e_{33}=-18.6 & e_{24}=e_{15}=11.6 & \mu_{11}=\mu_{22}=5 & \mu_{33}=10 \end{array}$$

With the above notation, the equation of equilibrium (2.4) is also equivalent to

$$\partial_t^2 u = \mathrm{Div}\sigma + f,$$

where Div is the operator-valued matrix defined by

$$\mathrm{Div} = \begin{pmatrix} \partial_1 & 0 & 0 & 0 & \partial_3 & \partial_2 \\ 0 & \partial_2 & 0 & \partial_3 & 0 & \partial_1 \\ 0 & 0 & \partial_3 & \partial_2 & \partial_1 & 0 \end{pmatrix}.$$

Assuming that u, E, H are of the form

$$u(x,t) = e^{-i\omega t}u(x), E(x,t) = e^{-i\omega t}E(x), H(x,t) = e^{-i\omega t}H(x),$$

for some real constant ω (the data being of the same form), the above system is reduced to the time-harmonic piezoelectric system in Ω consisting of the constitutive equation (2.8), the time-harmonic equilibrium equation

$$\mathrm{Div}\sigma + \omega^2 u = -f \text{ in } \Omega, \tag{2.9}$$

and the time-harmonic Maxwell's equations

$$\mathbf{curl}\, H + i\omega D = J \text{ in } \Omega, \tag{2.10}$$

$$\mathbf{curl}\, E - i\omega\mu H = 0 \text{ in } \Omega. \tag{2.11}$$

In the whole paper we consider the nonstationary case; namely, we assume that $\omega > 0$. In other words we assume that the variation in time of u, E, and H is periodic in time with a frequency equal to $2\pi/\omega$. The stationary case $\omega = 0$ requires the use of Gauss' law

$$\mathrm{div} D = \rho,$$

where ρ is the charge density function. This case is treated as in section 4.1 below and is then left to the reader.

3. Existence and uniqueness results when is positive definite. Replacing D by its expression (2.2) in the first Maxwell equation (2.10), we get

$$\mathbf{curl}\, H + i\omega(\epsilon E + e\gamma(u)) = J \text{ in } \Omega.$$

This identity is clearly equivalent to

$$i\omega\epsilon E = J - \mathbf{curl}\, H - i\omega e\gamma(u) \text{ in } \Omega. \tag{3.1}$$

With this identity, the second Maxwell equation (2.11) is then equivalent to

$$\mathbf{curl}(\epsilon^{-1}(\mathbf{curl}\, H + i\omega e\gamma(u))) - \omega^2\mu H = \mathbf{curl}(\epsilon^{-1}J) \text{ in } \Omega. \tag{3.2}$$

In the same way using the constitutive equation (2.8) in the equation of motion (2.9), we obtain

$$\operatorname{Div}\left(C\gamma(u) - e^\top E\right) + \omega^2 u = -f \text{ in } \Omega. \tag{3.3}$$

Using the identity (3.1), we arrive at

$$\begin{aligned}&\operatorname{Div}\left(C\gamma(u) + e^\top\epsilon^{-1}e\gamma(u) + \frac{1}{i\omega}e^\top\epsilon^{-1}(\mathbf{curl}\, H)\right)\\ &+\omega^2 u = -f + \frac{1}{i\omega}\operatorname{Div}\left(e^T\epsilon^{-1}J\right) \text{ in } \Omega.\end{aligned} \tag{3.4}$$

The two equations (3.2) and (3.4) constitute the system of partial differential equations that we will study and whose unknowns are u and H. Due to the boundary conditions (2.6) and (2.7), this system is completed with (2.6) and

$$\epsilon^{-1}(\mathbf{curl}\, H + i\omega e\gamma(u)) \times n = J \times n, (\mu H)\cdot n = 0 \text{ on } \Gamma. \tag{3.5}$$

The weak formulation of the above problem is obtained in the following way: We introduce the space (see [4, 5])

$$X_T(\Omega,\mu) = \{v \in L^2(\Omega)^3 : \operatorname{div}(\mu v) \in L^2(\Omega), \mathbf{curl}\, v \in L^2(\Omega)^3 \text{ and } (\mu v)\cdot n = 0 \text{ on } \Gamma\}$$

equipped with its natural norm. Then we multiply the system (3.2) by a test function $\bar{H}' \in X_T(\Omega,\mu)$, integrate the result in Ω, and integrate by parts to get (assuming that u and H are regular enough)

$$\begin{aligned}&\int_\Omega \left\{\epsilon^{-1}(\mathbf{curl}\, H + i\omega e\gamma(u))\cdot \mathbf{curl}\, \bar{H}' - \omega^2\mu H\cdot \bar{H}'\right\}\, dx\\ &= \int_\Omega \epsilon^{-1}J\cdot \mathbf{curl}\, \bar{H}'\, dx + \int_\Gamma (J\times n)\cdot \bar{H}'\, ds\ \forall H' \in X_T(\Omega,\mu).\end{aligned}$$

Since from the second Maxwell equation (2.11) $\operatorname{div}(\mu H) = 0$ in Ω, the above identity is equivalent to

$$\begin{aligned}&\int_\Omega \left\{\epsilon^{-1}(\mathbf{curl}\, H + i\omega e\cdot\gamma(u))\cdot \mathbf{curl}\, \bar{H}' + \operatorname{div}(\mu H)\operatorname{div}(\mu\bar{H}') - \omega^2\mu H\cdot \bar{H}'\right\}\, dx\\ &= \int_\Omega \epsilon^{-1}J\cdot \mathbf{curl}\, \bar{H}'\, dx + \int_\Gamma (J\times n)\cdot \bar{H}'\, ds\ \forall H' \in X_T(\Omega,\mu).\end{aligned} \tag{3.6}$$

This last equation may be called the regularized formulation of the system (3.2).

Similarly multiplying (3.4) by $\bar{v} \in H_0^1(\Omega)^3$, integrating the result in Ω, and using Green's formula

$$\int_\Omega (\mathrm{Div}\gamma) \cdot \bar{v}\, dx = -\int_\Omega \gamma \cdot \gamma(\bar{v})\, dx,$$

obtained by componentwise integration by parts in Ω, we arrive at

$$\begin{aligned}(3.7)\qquad &\int_\Omega \left\{ \left(C\gamma(u) + e^\top \epsilon^{-1} e\gamma(u) + \frac{1}{i\omega} e^\top \epsilon^{-1}\, \mathbf{curl}\, H \right) \cdot \gamma(\bar{v}) - \omega^2 u \cdot \bar{v} \right\} dx \\ &= \int_\Omega \left(f \cdot \bar{v} - \frac{1}{i\omega} e^T \epsilon^{-1} J \cdot \gamma(\bar{v}) \right) dx \ \forall v \in H_0^1(\Omega)^3.\end{aligned}$$

As these two identities are coupled, multiplying the second one by $i\omega$ and summing the result we arrive at the following problem:

$$\begin{aligned}&\int_\Omega \Big\{ \epsilon^{-1}(\mathbf{curl}\, H + i\omega e\gamma(u)) \cdot \mathbf{curl}\, \bar{H}' + \mathrm{div}(\mu H)\mathrm{div}(\mu \bar{H}') - \omega^2 \mu H \cdot \bar{H}' \\ &+ i\omega \left(C\gamma(u) + e^\top \epsilon^{-1} e\gamma(u) + \frac{1}{i\omega} e^\top \epsilon^{-1}\, \mathbf{curl}\, H \right) \cdot \gamma(\bar{v}) - i\omega^3 u \cdot \bar{v} \Big\} dx = F(v, H'),\end{aligned}$$

where we have set

$$(3.8)\quad F(v, H') = \int_\Omega \left(\epsilon^{-1} J \cdot \mathbf{curl}\, \bar{H}' - e^T \epsilon^{-1} J \cdot \gamma(\bar{v}) + i\omega f \cdot \bar{v} \right) dx + \int_\Gamma (J \times n) \cdot \bar{H}'\, ds.$$

In order to get a well-posed problem we set $u_\omega = i\omega u$. Then we see that the above problem is equivalent to finding a solution $(u_\omega, H) \in V$ of

$$(3.9)\qquad a((u_\omega, H), (v, H')) = F(v, H')\ \forall (v, H') \in V,$$

where we set

$$\begin{aligned} V &= H_0^1(\Omega)^3 \times X_T(\Omega, \mu), \\ a((u, H), (v, H')) &= \int_\Omega \big\{ \epsilon^{-1}(\mathbf{curl}\, H + e\gamma(u)) \cdot \mathbf{curl}\, \bar{H}' + \mathrm{div}(\mu H)\mathrm{div}(\mu \bar{H}') - \omega^2 \mu H \cdot \bar{H}' \\ &\quad + \left(C\gamma(u) + e^\top \epsilon^{-1} e\gamma(u) + e^\top \epsilon^{-1}\, \mathbf{curl}\, H \right) \cdot \gamma(\bar{v}) - \omega^2 u \cdot \bar{v} \big\}\, dx.\end{aligned}$$

In summary we have shown the following lemma.

LEMMA 3.1. *If u, E, H are solutions of* (2.8), (2.9), (2.10), *and* (2.11) *with the boundary conditions* (2.6) *and* (2.7), *then $(i\omega u, H)$ is a solution of* (3.9).

We now remark that the bilinear form a may be equivalently written

$$\begin{aligned}& a\ ((u, H), (v, H')) \\ &= \int_\Omega \left\{ \mathcal{A} \begin{pmatrix} \mathbf{curl}\, H \\ \gamma(u) \end{pmatrix} \cdot \begin{pmatrix} \mathbf{curl}\, \bar{H}' \\ \gamma(\bar{v}) \end{pmatrix} + \mathrm{div}(\mu H)\mathrm{div}(\mu \bar{H}') - \omega^2 \mu H \cdot \bar{H}' - \omega^2 u \cdot \bar{v} \right\} dx,\end{aligned}$$

where $\mathcal{A}$ is the 9×9 symmetric matrix defined by

$$\mathcal{A} = \begin{pmatrix} \epsilon^{-1} & \epsilon^{-1} e \\ e^\top \epsilon^{-1} & C + e^\top \epsilon^{-1} e \end{pmatrix}.$$

One easily checks that this matrix is positive definite for the material coefficients of the PZT-4 and $BaTiO_3$, for instance. In fact this is always true if ϵ and C are positive definite (always true in our setting), independently of the piezoelectric coefficients.

LEMMA 3.2. *If ϵ and C are positive definite matrices, then the matrix $\mathcal{A}$ defined above is also positive definite.*

Proof. We only need to show that

$$\mathcal{A}\begin{pmatrix} X \\ \gamma \end{pmatrix} \cdot \begin{pmatrix} X \\ \gamma \end{pmatrix} \geq 0, \forall \begin{pmatrix} X \\ \gamma \end{pmatrix} \in \mathbb{R}^9,$$
$$\mathcal{A}\begin{pmatrix} X \\ \gamma \end{pmatrix} \cdot \begin{pmatrix} X \\ \gamma \end{pmatrix} = 0 \Rightarrow \begin{pmatrix} X \\ \gamma \end{pmatrix} = 0.$$

By the definition of $\mathcal{A}$ we see that

$$\mathcal{A}\begin{pmatrix} X \\ \gamma \end{pmatrix} \cdot \begin{pmatrix} X \\ \gamma \end{pmatrix} = X^\top \epsilon^{-1} X + 2X^\top \epsilon^{-1} e\gamma + \gamma^\top C\gamma + \gamma^\top e^\top \epsilon^{-1} e\gamma.$$

Setting $\tilde{X} = \epsilon^{-1/2}X$ and $\tilde{Y} = \epsilon^{-1/2}e\gamma$ (note that both vectors are in $\mathbb{R}^3$), we see that

$$\begin{aligned}\mathcal{A}\begin{pmatrix} X \\ \gamma \end{pmatrix} \cdot \begin{pmatrix} X \\ \gamma \end{pmatrix} &= \tilde{X}^\top \tilde{X} + 2\tilde{X}^\top \tilde{Y} + \gamma^\top C\gamma + \tilde{Y}^\top \tilde{Y} \\ &= \|\tilde{X} + \tilde{Y}\|_2^2 + \gamma^\top C\gamma,\end{aligned}$$

where $\|\cdot\|_2$ clearly means the Euclidean norm of $\mathbb{R}^3$. This identity directly implies the first assertion by the positive definitiveness of C.

For the second assertion, if

$$\mathcal{A}\begin{pmatrix} X \\ \gamma \end{pmatrix} \cdot \begin{pmatrix} X \\ \gamma \end{pmatrix} = 0,$$

then the above identity and again the positive definitiveness of C imply that

$$\|\tilde{X} + \tilde{Y}\|_2^2 = \gamma^\top C\gamma = 0.$$

Therefore $\gamma = 0$ and, consequently, $\tilde{Y} = 0$ in view of its definition (independently of e). We then obtain that $\tilde{X} = 0$ and, by the positive definitiveness of ϵ, we conclude that $X = 0$. $\square$

This lemma allows us to show that problem (3.9) enters within the framework of the Fredholm alternative. Indeed, we shall prove the following lemma.

LEMMA 3.3. *There exists a discrete set S such that for $1 + \omega^2 \notin S$, the problem* (3.9) *has a unique solution for any $F \in V'$.*

Proof. Introduce the sesquilinear form

$$b((u,H),(v,H')) := a((u,H),(v,H')) + (1+\omega^2)\left(\int_\Omega \mu H \cdot \bar{H}'\,dx + \int_\Omega u \cdot \bar{v}\,dx\right).$$

Then the above considerations yield

$$\begin{aligned}&b\ ((u,H),(v,H')) \\ &= \int_\Omega \left\{\mathcal{A}\begin{pmatrix} \mathbf{curl}\,H \\ \gamma(u) \end{pmatrix} \cdot \begin{pmatrix} \mathbf{curl}\,\bar{H}' \\ \gamma(\bar{v}) \end{pmatrix} + \operatorname{div}(\mu H)\operatorname{div}(\mu\bar{H}') + \mu H \cdot \bar{H}' + u \cdot \bar{v}\right\} dx.\end{aligned}$$

It is coercive on V, since by Lemma 3.2 we have

$$b((u,H),(u,H)) \gtrsim \int_\Omega \left\{|\,\mathbf{curl}\, H|^2 + |\gamma(u)|^2 + |\mathrm{div}(\mu H)|^2 + |H|^2 + |u|^2\right\} dx.$$

Therefore by Korn's inequality we get

$$b((u,H),(u,H)) \gtrsim \|u\|^2_{1,\Omega} + \|H\|^2_{X_T(\Omega,\mu)}.$$

Since the space $X_T(\Omega,\mu)$ is compactly embedded into $L^2(\Omega)^3$ [28], and since by the Rellich–Kondrasov theorem $H^1_0(\Omega)$ is compactly embedded into $L^2(\Omega)$, we deduce that V is compactly embedded into $\mathcal{H} := L^2(\Omega)^3 \times L^2(\Omega)^3$. These facts imply that the Friedrichs extension B from $\mathcal{H}$ into $\mathcal{H}$ induced by the triple $(V,\mathcal{H},b)$ is invertible with a compact inverse. Now denote by A the Friedrichs extension of the triple $(V,\mathcal{H},a)$. The relation between a and b implies that $A = B - (1+\omega^2)I_\mu$, where the operator I_μ is defined by

$$I_\mu(u,H)^\top = (u,\mu H)^\top$$

and is clearly continuous from $\mathcal{H}$ into $\mathcal{H}$. Using this operator we may write

$$(1+\omega^2)^{-1}B^{-1}A = (1+\omega^2)^{-1}I - B^{-1}I_\mu,$$

where I is the identity operator from $\mathcal{H}$ into $\mathcal{H}$. Since $B^{-1}I_\mu$ is a compact operator, it has a discrete spectrum with positive eigenvalues. If we denote by S the set of the inverses of these eigenvalues, then for $(1+\omega^2) \notin S$, the operator $(1+\omega^2)^{-1}I - B^{-1}I_\mu$ is invertible and consequently A is as well. $\square$

Remark 3.4. Note that in the case $\mu = 0$, problem (3.9) does not enter in the Fredholm framework since $X_T(\Omega,0)$ is reduced to $H(\mathbf{curl},\Omega)$ and is no more compactly embedded into $L^2(\Omega)^3$. This drawback will be avoided by introducing Gauge conditions; see section 4.2.

Let us now show that the converse of Lemma 3.1 holds under some conditions on ω (compare with Theorem 0.1 of [4]).

THEOREM 3.5. *Assume that ω^2 is not an eigenvalue of $-\Delta^{Neu}_\mu := -\mathrm{div}(\mu\nabla\cdot)$, the (nonnegative) Laplace operator with Neumann boundary conditions in Ω. If $(u_\omega,H) \in V$ is a solution of* (3.9) *with F defined by* (3.8)*, then* $\mathrm{div}(\mu H) = 0$ *and u,E,H are solutions of* (2.9), (2.10)*, and* (2.11) *with the boundary conditions* (2.6) *and* (2.7) *when $u = \frac{u_\omega}{i\omega}$, E is given by* (3.1)*, and σ, D, B are defined by* (2.8)*.*

Proof. In (3.9) we take $v = 0$ and $H' = \nabla\bar{\Phi}$, with $\Phi \in D(\Delta^{Neu}_\mu)$. This yields

$$\int_\Omega \left\{\mathrm{div}(\mu H)\mathrm{div}(\mu\nabla\Phi) - \omega^2\mu H\cdot\nabla\Phi\right\} dx = 0.$$

By Green's formula we obtain

$$\int_\Omega \mathrm{div}(\mu H)(\Delta^{Dir}_\mu\Phi + \omega^2\Phi)\, dx = 0 \;\forall\Phi \in D(\Delta^{Neu}_\mu).$$

Since $D(\Delta^{Neu}_\mu)$ is dense in $L^2(\Omega)$, we conclude that $\mathrm{div}(\mu H) = 0$ in Ω.

The equations (2.9), (2.10), and (2.11) are obtained using Green's formula with test functions $H' \in \mathcal{D}(\Omega)^3$ or $v \in \mathcal{D}(\Omega)^3$. It then remains to show the boundary condition

$$\epsilon^{-1}(\mathbf{curl}\, H + i\omega e\gamma(u)) \times n = J \times n \text{ on } \Gamma,$$

which then implies (2.7), by the definition of E.

We first remark that the above arguments show that

$$w = \epsilon^{-1}(\mathbf{curl}\, H + i\omega e\gamma(u)) - J$$

belongs to $H(\mathbf{curl}, \Omega)$.

We now state the next Green's formula, which is a variant of the "standard" one (see Theorem I.2.11 of [10] or the identity (2.5) from [1]). A face F of Ω being fixed we denote by

$$V_F = \{\varphi \in H^1(\Omega) : \varphi = 0 \text{ on } \Gamma \setminus F\}.$$

We recall that for $\varphi \in V_F$, Theorem 2.2 of [21] implies that the trace of φ on F belongs to $H^{1/2}_{00}(F)$. By standard Green's formula

$$\int_\Omega (w \cdot \mathbf{curl}\, \varphi - \varphi \cdot \mathbf{curl}\, w)\, dx = \int_\Gamma (w \times n) \cdot \varphi,$$

valid for any $w, \varphi \in H^1(\Omega)^3$; from the fact that the above left-hand side is continuous on $H(\mathbf{curl}, \Omega) \times (V_F)^3$, we deduce that if $w \in H(\mathbf{curl}, \Omega)$, then $w \times n$ belongs to $(H^{1/2}_{00}(F)')^3$, and that Green's identity

$$\int_\Omega (w \cdot \mathbf{curl}\, \varphi - \varphi \cdot \mathbf{curl}\, w)\, dx = \langle w \times n, \varphi\rangle_F \;\; \forall \varphi \in (V_F)^3 \tag{3.10}$$

holds (where $\langle ., .\rangle_F$ means the duality pairing between $(H^{1/2}_{00}(F)')^3$ and $H^{1/2}_{00}(F)^3$).

Now for a fixed face F, we temporarily assume that the z-axis is perpendicular to F and that F is included in the plane $z = 0$. For $\varphi_1, \varphi_2 \in \mathcal{D}(F)$, we take

$$H'(x, y, z) = \eta(z)\mu^{-1} \cdot \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ 0 \end{pmatrix}$$

with a cut-off function $\eta \in \mathcal{D}(\mathbb{R})$ such that $\eta(0) = 1$ and a sufficiently small support so that H' is zero on $\Gamma \setminus F$. By construction the function H' belongs to $X_T(\Omega, \mu) \cap (V_F)^3$. Therefore in (3.9), taking as a test function $v = 0$ and this function H', by Green's formula (3.10) we get

$$\langle w \times n, H'\rangle_F = 0.$$

As the third component of $w \times n$ is zero and the two first components of

$$\mu^{-1} \cdot \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ 0 \end{pmatrix}$$

are arbitrary in $H^{1/2}_{00}(F)$, we deduce that

$$w \times n = 0 \text{ in } (H^{1/2}_{00}(F)')^3.$$

By the definition of w, the requested boundary condition is proved. $\square$

4. Existence and uniqueness results when = 0.

4.1. The -formulation. In the case $\mu = 0$, the second Maxwell equation (2.11) and the boundary condition (2.7) imply that

$$i\omega E = \nabla\varphi, \tag{4.1}$$

for some $\varphi \in H_0^1(\Omega)$. Therefore, in order to eliminate the vector field H, we take the divergence of the first Maxwell equation (2.10) to get

$$\operatorname{div}(i\omega D) = \operatorname{div} J,$$

and by the constitutive equation (2.8) we obtain

$$\operatorname{div}(i\omega e\gamma(u) + \epsilon i\omega E) = \operatorname{div} J.$$

As before setting $u_\omega = i\omega u$, the above identity is equivalent to

$$\operatorname{div}(e\gamma(u_\omega) + \epsilon\nabla\varphi) = \operatorname{div} J. \tag{4.2}$$

This equation is now coupled with (3.3) or, equivalently, with

$$\operatorname{Div}\left(C\gamma(u_\omega) - e^\top\nabla\varphi\right) + \omega^2 u_\omega = -i\omega f \text{ in } \Omega \tag{4.3}$$

to form our system of partial differential equations. Clearly, its weak formulation is in finding a solution $(u_\omega, \varphi) \in H_0^1(\Omega)^4$ of

$$a((u_\omega,\varphi),(v,\psi)) = F(v,\psi), \forall (v,\psi) \in H_0^1(\Omega)^4, \tag{4.4}$$

where the bilinear form a is defined by

$$a((u,\varphi),(v,\psi)) = \int_\Omega \left\{(e\gamma(u) + \epsilon\nabla\varphi)\cdot\nabla\psi + (C\gamma(u) - e^\top\nabla\varphi)\cdot\gamma(v) - \omega^2 u\cdot v\right\}\, dx,$$

and the linear form F is given by

$$F(v,\psi) = \int_\Omega (i\omega f\cdot v - J\cdot\nabla\psi)\, dx. \tag{4.5}$$

The above considerations show the following lemma.

LEMMA 4.1. *If u, E, H are solutions of* (2.8), (2.9), (2.10), *and* (2.11) *with the boundary conditions* (2.6) *and* (2.7), *then* $(i\omega u, \varphi)$, *with* φ *given by* (4.1), *is a solution of* (4.4).

As in the previous section, problem (4.4) satisfies the Fredholm alternative; namely, we shall prove the following lemma.

LEMMA 4.2. *There exists a discrete set S_0 such that for $\omega^2 \notin S_0$, the problem* (4.4) *has a unique solution for any $F \in H^{-1}(\Omega)^4$.*

Proof. Introduce the bilinear form

$$b((u,\varphi),(v,\psi)) = \int_\Omega \left\{(e\gamma(u) + \epsilon\nabla\varphi)\cdot\nabla\psi + (C\gamma(u) - e^\top\nabla\varphi)\cdot\gamma(v)\right\}\, dx.$$

This form is coercive on $H_0^1(\Omega)^4$ since

$$b((u,\varphi),(u,\varphi)) = \int_\Omega \{\epsilon\nabla\varphi\cdot\nabla\varphi + C\gamma(u)\cdot\gamma(u)\}\, dx \gtrsim \|\varphi\|_{1,\Omega}^2 + \|u\|_{1,\Omega}^2,$$

thanks to the positive definiteness of ϵ and C and Korn's inequality.

We conclude as in Lemma 3.3. $\square$

We end this subsection by the converse result of Lemma 4.1.

THEOREM 4.3. *If* $(u_\omega, \varphi) \in H_0^1(\Omega)^4$ *is the solution of* (4.4) *with* F *defined by* (4.5)*, then* u, E *are solutions of* (2.9) *and* (2.11) *with the boundary conditions* (2.6) *and* (2.7) *when* $u = \frac{u_\omega}{i\omega}$*,* E *is given by* (4.1)*, and* σ, D, B *are defined by* (2.8) *(implying* $B = 0$ *since* $\mu = 0$*). Moreover, there exists* $H \in H^1(\Omega)^3$ *such that* (2.10) *holds.*

Proof. The first part of the lemma follows from Green's formula. For the second part we simply remark that D defined by (2.8) satisfies

$$\operatorname{div}(i\omega D) = \operatorname{div} J \text{ in } \Omega,$$

or, equivalently, $J - i\omega D$ is divergence free. Therefore the existence of H such that $\mathbf{curl}\, H = J - i\omega D$ follows from Theorem I.3.4 of [10]. □

In the above lemma, we may notice that the magnetic field H is not unique, but this is not important since, in the case $\mu = 0$, only u and E are of practical interest.

4.2. The -formulation. Following the arguments of section 3, we may eliminate the electric field E and keep as unknowns u and H. Unfortunately H is no more uniquely determined, since the condition $\operatorname{div}(\mu H) = 0$ and the boundary condition $\mu H \cdot n = 0$ on Γ are trivially satisfied. Therefore as for eddy current problems [3] we impose the following Gauge conditions:

$$\operatorname{div} H = 0 \text{ in } \Omega, \tag{4.6}$$

$$H \cdot n = 0 \text{ on } \Gamma. \tag{4.7}$$

These conditions may be justified by an asymptotic argument, namely by taking $\mu = \eta I$, with $\eta > 0$ and letting η tends to 0 (cf. [6]).

These arguments imply the following lemma.

LEMMA 4.4. *If* u, E, H *are solutions of* (2.8), (2.9), (2.10), *and* (2.11) *with the boundary conditions* (2.6) *and* (2.7)*, assume that* H *satisfies the Gauge conditions* (4.6)–(4.7)*. Then* $(i\omega u, H)$ *belongs to* $V = H_0^1(\Omega)^3 \times X_T(\Omega, I)$ *and is solution of* (3.9)*, with the sesquilinear form here defined by*

$$\begin{aligned} a((u,H),(v,H')) = \int_\Omega \big\{ &\epsilon^{-1}(\mathbf{curl}\, H + e\gamma(u)) \cdot \mathbf{curl}\, \bar{H}' + \operatorname{div}(H)\operatorname{div}(\bar{H}') \\ &+ \big(C\gamma(u) + e^\top \epsilon^{-1} e\gamma(u) + e^\top \epsilon^{-1}\, \mathbf{curl}\, H\big) \cdot \gamma(\bar{v}) - \omega^2 u \cdot \bar{v} \big\}\, dx. \end{aligned}$$

Again, the above problem enters in the Fredholm alternative.

LEMMA 4.5. *There exists a discrete set* S_1 *such that for* $1 + \omega^2 \notin S_1$*, the problem* (3.9)*, with* a *and* V *defined in Lemma 4.4, has a unique solution for any* $F \in V'$*.*

Proof. Introduce the sesquilinear form

$$b((u,H),(v,H')) := a((u,H),(v,H')) + (1+\omega^2)\int_\Omega u \cdot \bar{v}\, dx.$$

Using Lemma 3.2, Korn's inequality, and the compact embedding of $X_T(\Omega, I)$ into $L^2(\Omega)^3$ [28], the bilinear form b is coercive on V. Consequently, the Friedrichs extension B associated with the triple $(V, \mathcal{H}, b)$ (with $\mathcal{H}$ defined as in the proof of Lemma 3.3) is invertible with a compact inverse. Again denote by A the Friedrichs extension of the triple $(V, \mathcal{H}, a)$. Clearly $A = B - (1+\omega^2)K$, where the continuous operator K is defined by

$$K(u,H)^\top = (u,0)^\top.$$

Writing

$$(1+\omega^2)^{-1}B^{-1}A = (1+\omega^2)^{-1}I - B^{-1}K,$$

we conclude as in Lemma 3.3, since $B^{-1}K$ is a compact operator. ∎

Similarly to Theorem 3.5, the converse of Lemma 4.4 holds under some conditions on ω.

THEOREM 4.6. *Assume that ω^2 is not an eigenvalue of $-\Delta_I^{Neu}$. If $(u_\omega, H) \in V$ is a solution of* (3.9) *with F defined by* (3.8) *and a, V defined in Lemma* 4.4, *then* $\mathrm{div} H = 0$ *and u, E, H are solutions of* (2.9), (2.10), *and* (2.11) *with the boundary conditions* (2.6) *and* (2.7) *when $u = \frac{u_\omega}{i\omega}$, E is given by* (3.1), *and $\sigma, D, B = 0$ are defined by* (2.8).

4.3. Equivalence between the -formulation and the -formulation. The goal of this section is to show that the displacement and electric fields obtained by the E-formulation and the H-formulation are identical.

THEOREM 4.7. *Let u_1, E_1, H_1 (resp., u_2, E_2, H_2) be the solutions of* (2.9), (2.10), *and* (2.11) *with the boundary conditions* (2.6) *and* (2.7) *obtained by the E-formulation (resp., H-formulation). If $\omega^2 \notin S_0$ (cf. Lemma 4.2), then $u_1 = u_2$ and $E_1 = E_2$.*

Proof. Denote $u = u_1 - u_2$ and $E = E_1 - E_2$. Since u_1, E_1 and u_2, E_2 satisfy (2.9) and (2.11), u, E satisfy

$$\begin{aligned}
&\mathrm{Div}\sigma + \omega^2 u = 0 \text{ in } \Omega,\\
&\mathbf{curl}\, H + i\omega D = 0 \text{ in } \Omega,\\
&\mathbf{curl}\, E = 0 \text{ in } \Omega,
\end{aligned}$$

where $\sigma = \sigma_1 - \sigma_2, and D = D_1 - D_2$ are given by (2.8). Therefore $E = \nabla\chi$, with $\chi \in H_0^1(\Omega)$ and the pair $(u, \chi) \in H_0^1(\Omega)^4$ is a solution of (4.4) with $F = 0$. By Lemma 4.2, we conclude that $u = 0$ and $\chi = 0$. ∎

Remark 4.8. The magnetic fields obtained by the E-formulation and the H-formulation cannot be identical, since the one obtained by the E-formulation does not necessarily satisfy the Gauge conditions.

5. Singularities of some elliptic systems. It is well known that the singularities of elliptic systems in Ω are produced by the corners and edges of Ω. Here we briefly describe the corner and edge singularities of the operator Δ_μ^{Neu} and of the 4×4 system (4.2)–(4.3) with Dirichlet boundary conditions. For the sake of brevity we restrict ourselves to a minimal description and refer the reader to the pioneer work [15] or standard books [11, 7, 18] for more details.

5.1. Corner singularities. Fix a corner $c \in \mathcal{C}$ of Ω and denote by (ρ_c, ϑ_c) the spherical coordinates centered at c. Denote by Γ_c the infinite polyhedral cone which coincides with Ω near c. Let G_c be the intersection of Γ_c with the unit sphere centered at c.

For any $\lambda \in \mathbb{C}$, let us set

$$S^\lambda(\Gamma_c) = \left\{ \psi(x) = \rho_c^\lambda \sum_{q=0}^{Q} (\log\rho_c)^q \psi_q(\vartheta_c) : \psi_q \in H^1(G_c) \right\}.$$

The set $\Lambda_\mu^{Neu}(\Gamma_c)$ of corner singular exponents of the operator Δ_μ^{Neu} in Γ_c is the set of $\lambda \in \mathbb{C}$ such that there exists a nonpolynomial solution $\Psi \in S^\lambda(\Gamma_c)$ of

$$\Delta_\mu \Psi = \mathrm{div}(\mu\nabla\Psi) = 0 \text{ in } \Gamma_c, \tag{5.1}$$

$$\mu\nabla\Psi \cdot n = 0 \text{ on } \partial\Gamma_c\,. \tag{5.2}$$

We denote by $Z^{\lambda}_{Neu}(\Gamma_c, \mu)$ the space of these solutions.

Similarly for the 4×4 system (4.2)–(4.3) with Dirichlet boundary conditions, its set $\Lambda^{Dir}_{C,\epsilon,e}(\Gamma_c)$ of corner singular exponents is the set of $\lambda \in \mathbb{C}$ such that there exists a nonpolynomial solution $(u, \chi) \in S^{\lambda}(\Gamma_c)^4$ of

$$\left\{ \begin{array}{ll} \operatorname{div}(\epsilon \nabla \chi + e\gamma(u)) = 0 & \text{in } \Gamma_c, \\ \operatorname{Div}\left(C\gamma(u) - e^{\top}\nabla\chi\right) = 0 & \text{in } \Gamma_c, \\ \chi = 0, u = 0 & \text{on } \partial\Gamma_c. \end{array} \right. \tag{5.3}$$

The space of these solutions is denoted by $Z^{\lambda}_{Dir}(\Gamma_c, C, \epsilon, e)$.

5.2. Edge singularities. Fix an edge $a \in \mathcal{A}$ of Ω and denote by $\Gamma_a \times \mathbb{R}$ the infinite polyhedral cone which coincides with Ω near a (Γ_a is then a two-dimensional sector). Denote by (r_a, θ_a, z_a) the cylindrical coordinates along a. Let G_a be the intersection of Γ_a with the unit sphere $r_a = 1$.

As before, for any $\lambda \in \mathbb{C}$ let us set

$$S^{\lambda}(\Gamma_a) = \left\{ \psi(x) = r_a^{\lambda} \sum_{q=0}^{Q} (\log r_a)^q \psi_q(\theta_a) : \psi_q \in H^1(G_a) \right\}.$$

The set $\Lambda^{Neu}_{\mu}(\Gamma_a)$ of corner singular exponents of the operator Δ^{Neu}_{μ} in Γ_a is the set of $\lambda \in \mathbb{C}$ such that there exists a nonpolynomial solution $\Psi \in S^{\lambda}(\Gamma_a)$ of

$$\begin{array}{l} \operatorname{div}_2(\mu_2 \nabla_2 \Psi) = 0 \text{ in } \Gamma_a, \\ \mu_2 \nabla_2 \Psi \cdot n = 0 \text{ on } \partial\Gamma_a, \end{array}$$

where div_2 (resp., ∇_2) means the two-dimensional divergence (resp., gradient) and μ_2 is the 2×2 matrix obtained from μ by dropping the third line and the third column of μ.

We denote by $Z^{\lambda}_{Neu}(\Gamma_a, \mu)$ the space of these solutions.

In the same way, the set $\Lambda^{Dir}_{C,\epsilon,e}(\Gamma_a)$ of edge singular exponents of the 4×4 system (4.2)–(4.3) with Dirichlet boundary conditions is the set of $\lambda \in \mathbb{C}$ such that there exists a nonpolynomial solution $(u, \chi) \in S^{\lambda}(\Gamma_a)^4$ of

$$\left\{ \begin{array}{ll} \tilde{\operatorname{div}}(\epsilon \tilde{\nabla} \chi + e\tilde{\gamma}(u)) = 0 & \text{in } \Gamma_a, \\ \tilde{\operatorname{Div}}\left(C\tilde{\gamma}(u) - e^{\top}\tilde{\nabla}\chi\right) = 0 & \text{in } \Gamma_a, \\ \chi = 0, u = 0 & \text{on } \partial\Gamma_a, \end{array} \right. \tag{5.4}$$

where the sign $\tilde{}$ means that all derivatives in the z_a-variable are replaced by zero. The space of these solutions is denoted by $Z^{\lambda}_{Dir}(\Gamma_a, C, \epsilon, e)$.

6. Corner and edge singularities when is positive definite. Our goal is to describe the corner and edge singularities of the (regularized) problem (3.9). These singularities are obtained using some ideas from [4, 5, 6].

Let us start with the corner singularities.

6.1. Corner singularities. Fix a corner $c \in \mathcal{C}$ of Ω. We drop the index c in the above considerations. As usual we are looking for solutions of the homogeneous problem in the space

$$\mathbf{S}_T^\lambda(\Gamma,\mu) = \Big\{(u,H) \in H^1_{0,loc}(\Gamma)^3 \times X_{T,loc}(\Gamma,\mu) : \operatorname{div}(\mu H) \in H^1_{loc}(\Gamma),\\ u(x) = \rho^\lambda \sum_{q=0}^{Q} (\log\rho)^q U_q(\vartheta),\\ H(x) = \rho^\lambda \sum_{q=0}^{Q} (\log\rho)^q H_q(\vartheta)\Big\},$$

the index *loc* meaning that the properties hold in all bounded domains far from c. This means that we look for a nonpolynomial solution $(u,H) \in \mathbf{S}_T^\lambda(\Gamma,\mu)$ of

$$\left\{\begin{array}{ll} \mathbf{curl}\left(\epsilon^{-1}\left(\mathbf{curl}\, H + e\gamma(u)\right)\right) - \mu\nabla\operatorname{div}(\mu H) = 0 & \text{in } \Gamma,\\ \operatorname{Div}\left(C\gamma(u) + e^\top\epsilon^{-1}e\gamma(u) + e^\top\epsilon^{-1}\,\mathbf{curl}\, H\right) = 0 & \text{in } \Gamma,\\ u = 0 & \text{on } \partial\Gamma,\\ \mu H\cdot n = 0 & \text{on } \partial\Gamma,\\ \mu\nabla(\operatorname{div}(\mu H))\cdot n = 0 & \text{on } \partial\Gamma,\\ \epsilon^{-1}\left(\mathbf{curl}\, H + e\gamma(u)\right)\times n = 0 & \text{on } \partial\Gamma. \end{array}\right. \tag{6.1}$$

If a nontrivial solution exists, then we say that λ is a corner exponent.

Inspired by [4, 5, 6], this problem is split up into three subproblems by introducing the auxiliary unknowns

$$\psi = -\epsilon^{-1}\left(\mathbf{curl}\, H + e\gamma(u)\right)$$

and

$$q = \operatorname{div}(\mu H).$$

With this notation, problem (6.1) is equivalent to looking for $q, (\psi,u), H$ successive solutions of

$$\left\{\begin{array}{ll} \operatorname{div}(\mu\nabla q) = 0 & \text{in } \Gamma,\\ \mu\nabla q\cdot n = 0 & \text{on } \partial\Gamma, \end{array}\right. \tag{6.2}$$

$$\left\{\begin{array}{ll} \mathbf{curl}\,\psi = -\mu\nabla q & \text{in } \Gamma,\\ \operatorname{div}(\epsilon\psi + e\gamma(u)) = 0 & \text{in } \Gamma,\\ \operatorname{Div}\left(C\gamma(u) - e^\top\psi\right) = 0 & \text{in } \Gamma,\\ \psi\times n = 0, u = 0 & \text{on } \partial\Gamma, \end{array}\right. \tag{6.3}$$

$$\left\{\begin{array}{ll} \mathbf{curl}\, H = -\epsilon\psi - e\gamma(u) & \text{in } \Gamma,\\ \operatorname{div}(\mu H) = q & \text{in } \Gamma,\\ \mu H\cdot n = 0 & \text{on } \partial\Gamma. \end{array}\right. \tag{6.4}$$

This means that three types of singularities may appear:

Type 1: $q = 0$, $\psi = u = 0$, and H is a general solution of (6.4).

Type 2: $q = 0$, (ψ,u) is a general solution of (6.3), and H is a particular solution of (6.4).

Type 3: q is a general solution of (6.2), (ψ, u) is a particular solution of (6.3), and H is a particular solution of (6.4).

These three types of singularities may be described with the help of the corner singularities of the operator Δ_μ^{Neu} and of the 4×4 elliptic system (4.2)–(4.3) with Dirichlet boundary conditions described in subsection 4.1.

Since for our problem (3.9) $\operatorname{div}(\mu H) = 0$ and is then regular, we do not describe the singularities of Type 3 because they cannot occur for any solution of (3.9).

The singularities of Type 1 are obtained exactly as in Lemma 5.1 of [5] since only the magnetic field H is involved.

LEMMA 6.1. *Assume that $\lambda \neq -1$. Then $(u, H) \in \mathbf{S}_T^\lambda(\Gamma, \mu)$ is a singularity of type 1 if and only if $\lambda + 1$ belongs to $\Lambda_\mu^{Neu}(\Gamma)$ and $H = \nabla\Phi$, with $\Phi \in Z_{Neu}^{\lambda+1}(\Gamma, \mu)$.*

The situation is different for singularity of Type 2 since the coupling between the elasticity system and Maxwell's equations appear via problem (6.3).

LEMMA 6.2. *Assume that $\lambda \neq -1$. Then $(u, H) \in \mathbf{S}_T^\lambda(\Gamma, \mu)$ is a singularity of type 2 if and only if λ belongs to $\Lambda_{C,\epsilon,e}^{Dir}(\Gamma)$, $\psi = \nabla\chi$, with $(u, \chi) \in Z_{Dir}^\lambda(\Gamma, C, \epsilon, e)$, and H is given by*

$$H = \frac{1}{\lambda + 1} \left(a \times \mathbf{x} + \nabla r \right), \tag{6.5}$$

where $a = -(\epsilon\psi + e\gamma(u))$ and $r \in S^{\lambda+1}(\Gamma)$ is solution of

$$\left\{ \begin{array}{ll} \operatorname{div}(\mu\nabla r) = -\operatorname{div}(\mu(a \times \mathbf{x})) & \textit{in } \Gamma, \\ \mu\nabla r \cdot n = -\mu(a \times \mathbf{x}) \cdot n & \textit{on } \partial\Gamma. \end{array} \right. \tag{6.6}$$

Proof. As

$$\mathbf{curl}\, \psi = 0 \text{ in } \Gamma,$$

there exists $\chi \in S^\lambda(\Gamma)$ such that

$$\psi = \nabla\chi \text{ in } \Gamma.$$

From (6.3) we deduce that

$$\left\{ \begin{array}{ll} \operatorname{div}(\epsilon\nabla\chi + e\gamma(u)) = 0 & \text{in } \Gamma, \\ \operatorname{Div}\left(C\gamma(u) - e^\top \nabla\chi\right) = 0 & \text{in } \Gamma, \\ \chi = 0, u = 0 & \text{on } \partial\Gamma. \end{array} \right.$$

In view of section 5, we deduce that $(u, \chi) \in Z_{Dir}^\lambda(\Gamma, C, \epsilon, e)$.

Now we readily check that H in the form (6.5) is a solution of (6.4) if and only if r is a solution of (6.6), whose solution exists by Theorem 4.14 of [23] and because $\mathbf{curl}(a \times \mathbf{x}) = (\lambda + 1)a$. ∎

LEMMA 6.3. *There is no corner singularity of Type 2 in the strip $\Re\lambda \in \left]-1, 0\right]$.*

Proof. By Theorem 1 of [16] (see Remark 2 of [16], or Theorem 2 of [17]), the set $\Lambda_{C,\epsilon,e}^{Dir}(\Gamma) \cap [-1, 0]$ is empty. We conclude by Lemma 6.2. ∎

Since the singularities of our problem (3.9) have to be locally in $X_T(\Gamma, \mu)$, among the singular exponents described above we select the subset Λ_c of $\lambda > -\frac{3}{2}$ such that there exists $(u, H) \in \mathbf{S}_T^\lambda(\Gamma, \mu)$ solution of (6.1) such that

$$\chi(u, H) \in X_T(\Gamma, \mu),$$

where χ is a cut-off function equal to 1 near c. This last condition implies the following constraints for our first two types of singularities (see [5, 6]):

Type 1: $\lambda+1 \in \Lambda_{\mu}^{Neu}(\Gamma)$ and since $\Lambda_{\mu}^{Neu}(\Gamma) \cap [-1,0]$ is empty, by Lemma 5.4 of [5] we get the condition $\lambda > -1$.

Type 2: $\lambda \in \Lambda_{C,\epsilon,e}^{Dir}(\Gamma)$ and by the condition $\chi u \in H^1(\Gamma)^3$, we get $\lambda > -\frac{1}{2}$. By Lemma 6.3, we arrive at the condition $\lambda > 0$.

In conclusion, the set of corner singular exponents is

$$\Lambda_c = \{\lambda > -1 : \lambda+1 \in \Lambda_{\mu}^{Neu}(\Gamma_c)\} \cup \{\lambda > 0 : \lambda \in \Lambda_{C,\epsilon,e}^{Dir}(\Gamma_c)\}.$$

6.2. Edge singularities. Fix an edge $a \in \mathcal{A}$ of Ω and drop the index a. As before, we are looking for solutions of the homogeneous piezoelectric system (6.1) in $\Gamma \times \mathbb{R}$. For the elasticity system or Maxwell's equations, the system obtained in the Cartesian coordinates (x, y, z) (according to the notation introduced in section 5, the z-axis contains the edge a) is split up into two independent problems in Γ_a. Here the coupling between these systems prevents this splitting. Therefore we simply follow the approach of the previous subsection. Namely, we search nonpolynomial solutions (u, H) of (6.1) independent of the z-variable and in the space

$$\begin{aligned}\mathbf{S}_T^{\lambda}(\Gamma,\mu) = \Big\{ &(u,H) \in H^1_{0,loc}(\Gamma\times\mathbb{R})^3 \times X_{T,loc}(\Gamma\times\mathbb{R},\mu) : \operatorname{div}(\mu H) \in H^1_{loc}(\Gamma\times\mathbb{R}),\\ &u(x,y,z) = r^{\lambda}\sum_{q=0}^{Q}(\log r)^q U_q(\theta),\\ &H(x,y,z) = r^{\lambda}\sum_{q=0}^{Q}(\log r)^q H_q(\theta)\Big\};\end{aligned}$$

the index *loc* here means that the properties hold in all bounded domains far from a.

Then we introduce the auxiliary unknowns ψ and q defined as before but here independent of the variable z. This leads to the successive problems (6.2), (6.3), and (6.4) (where all derivatives in z are equal to zero). As before, singularities of Types 1, 2, and 3 appear. The singularities of Type 3 are not studied for the same reason as before. We now describe the singularities of Type 1 (compare with Lemma 6.1).

LEMMA 6.4. *Assume that $\lambda \neq -1$. Then $(u, H) \in \mathbf{S}_T^{\lambda}(\Gamma,\mu)$ is an edge singularity of type 1 if and only if $\lambda+1$ belongs to $\Lambda_{\mu}^{Neu}(\Gamma)$, $u = 0$ and*

$$H = (\nabla_2\Phi, 0)^{\top},$$

with $\Phi \in Z_{Neu}^{\lambda+1}(\Gamma,\mu)$.

Proof. As

$$\mathbf{curl}\, H = \begin{pmatrix} \partial_2 H_3 \\ -\partial_1 H_3 \\ \partial_1 H_2 - \partial_2 H_1 \end{pmatrix} = 0,$$

the third component H_3 is constant and the field

$$h = \begin{pmatrix} H_1 \\ H_2 \end{pmatrix}$$

has a two-dimensional zero curl. Therefore there exists Φ such that $h = \nabla_2\Phi$, and we conclude as in Lemma 6.1. $\square$

Let us continue with singularities of Type 2.

LEMMA 6.5. *Assume that $\lambda \neq -1$ and $\lambda \neq 0$. Then $(u, H) \in \mathbf{S}_T^\lambda(\Gamma, \mu)$ is an edge singularity of Type 2 if and only if λ belongs to $\Lambda^{Dir}_{C,\epsilon,e}(\Gamma)$, $\psi = (\nabla_2\chi, 0)^\top$, with $(u, \chi) \in Z^\lambda_{Dir}(\Gamma, C, \epsilon, e)$ and H given by*

$$H = \begin{pmatrix} \frac{a_3 x_2}{\lambda} \\ \frac{a_3 x_1}{\lambda} \\ \frac{a_1 x_2 - a_2 x_1}{\lambda+1} \end{pmatrix} + (\nabla_2 r, 0)^\top, \tag{6.7}$$

where $a = -(\epsilon\psi + e\gamma(u))$ and $r \in S^{\lambda+1}(\Gamma)$ is the solution of

$$\left\{ \begin{array}{ll} \operatorname{div}_2(\mu\nabla_2 r) = -\operatorname{div}_2\left(\mu\begin{pmatrix} \frac{a_3 x_2}{\lambda} \\ \frac{a_3 x_1}{\lambda} \end{pmatrix}\right) & \text{in } \Gamma, \\ \mu\nabla r \cdot n = -\mu\begin{pmatrix} \frac{a_3 x_2}{\lambda} \\ \frac{a_3 x_1}{\lambda} \end{pmatrix} \cdot n & \text{on } \partial\Gamma. \end{array} \right. \tag{6.8}$$

Proof. As in the previous lemma, since $\operatorname{\mathbf{curl}}\psi = 0$ in Γ, there exists $\chi \in S^\lambda(\Gamma)$ such that

$$\psi = (\nabla_2\chi, 0)^\top \text{ in } \Gamma.$$

From (6.3) we deduce that

$$\left\{ \begin{array}{ll} \operatorname{div}(\epsilon\nabla\chi + e\gamma(u)) = 0 & \text{in } \Gamma, \\ \operatorname{Div}\left(C\gamma(u) - e^\top\nabla\chi\right) = 0 & \text{in } \Gamma, \\ \chi = 0, u = 0 & \text{on } \partial\Gamma. \end{array} \right.$$

This means that $(u, \chi) \in Z^\lambda_{Dir}(\Gamma, C, \epsilon, e)$ owing to section 5 (recalling that u and χ are independent of the z-variable).

As in Lemma 6.2, we easily check that H in the form (6.7) is a solution of (6.4) if and only if r is a solution of (6.8), whose existence follows from Theorem 4.14 of [23]. $\square$

In summary we may formulate the following corollary.

COROLLARY 6.6. *The set Λ_a of edge exponents associated with a is given by*

$$\Lambda_a = \{\lambda > -1 : \lambda + 1 \in \Lambda^{Neu}_\mu(\Gamma_a)\} \cup \{\lambda > 0 : \lambda \in \Lambda^{Dir}_{C,\epsilon,e}(\Gamma_a)\}.$$

7. Regularity results. In this section, we describe regularity results of a solution u, H, E of our time-harmonic piezoelectric system. These results use its weak formulation obtained above, are based on the knowledge of corner and edge singularities of these formulation described in the previous section, and rely on the application of Mellin's techniques as in [4, 5].

For the sake of brevity, we do not describe singular decomposition of such a solution. Using the techniques from [4, 5], for sufficiently smooth data we may obtain a decomposition of u, H, E into a regular part and a singular one.

7.1. The case positive definite. Before stating our regularity results, let us introduce the following notation: For any corner $c \in \mathcal{C}$ introduce

$$\lambda_{c,u} = \min\{\lambda > 0 : \lambda \in \Lambda^{Dir}_{C,\epsilon,e}(\Gamma_c)\},$$
$$\lambda_{c,H} = \min\{\lambda : \lambda \in \Lambda_c\}.$$

Similarly, for any edge $a \in \mathcal{A}$ define

$$\lambda_{a,u} = \min\{\lambda > 0 : \lambda \in \Lambda^{Dir}_{C,\epsilon,e}(\Gamma_a)\},$$
$$\lambda_{a,H} = \min\{\lambda : \lambda \in \Lambda_a\}.$$

Finally, we set

$$\tau_u := \min\left(\min_{a\in\mathcal{A}} \lambda_{a,u}, \frac{1}{2} + \min_{c\in\mathcal{C}} \lambda_{c,u}\right),$$
$$\tau_H := \min\left(\min_{a\in\mathcal{A}} \lambda_{a,H}, \frac{1}{2} + \min_{c\in\mathcal{C}} \lambda_{c,H}\right).$$

We remark that τ_u does not depend on μ.

THEOREM 7.1. *Let* $(u_\omega, H) \in H^1_0(\Omega)^3 \times X_T(\Omega,\mu)$ *be a solution of problem* (3.9), *with* J *and* f *smooth enough. Then we have*

$$u_\omega \in H^{1+\tau}(\Omega)^3 \ \forall \tau < \tau_u, \tag{7.1}$$
$$H \in H^{\tau}(\Omega)^3 \ \forall \tau < \tau_H. \tag{7.2}$$

Proof. Mellin's techniques directly imply that

$$(u_\omega, H) \in H^{\tau}(\Omega)^6 \ \forall \tau < \tau_H.$$

The extra regularity for u_ω follows from the description of the singularities of problem (3.9) made in section 6, where we see that the regularity of u_ω is only determined by the singularity of type 2. ∎

COROLLARY 7.2. *Assume that* J *and* f *are smooth. Let* u, E, H *be solutions of* (2.8), (2.9), (2.10), *and* (2.11) *with the boundary conditions* (2.6) *and* (2.7) *obtained by Theorem* 3.5. *Then* u *has the regularity* (7.1), H *has the regularity* (7.2), *and* E *satisfies*

$$E \in H^{\tau}(\Omega)^3 \ \forall \tau < \tau_u. \tag{7.3}$$

Proof. The previous lemma directly yields the regularity for u and H. By (3.1), E has the regularity of $\mathbf{curl}\, H + i\omega e\gamma(u)$. Since the singularities of type 1 for H are gradient, the regularity of E is only determined by the singularities of type 2. ∎

7.2. The case = 0. Here, to solve our piezoelectric system we may use either the E-formulation or the H-formulation. Both formulations give the same regularity for u and E, which is quite natural since generically they are identical (see Theorem 4.7).

THEOREM 7.3. *Let* $(u_\omega, \varphi) \in H^1_0(\Omega)^4$ *be a solution of problem* (4.4) *with* J *and* f *smooth enough; then we have*

$$(u_\omega, \varphi) \in H^{1+\tau}(\Omega)^4 \ \forall \tau < \tau_u, \tag{7.4}$$

where τ_u *is defined as before.*

Proof. Since the system associated with (4.4) is the strongly elliptic system (4.2)–(4.3) with Dirichlet boundary conditions, the mentioned regularity result follows from Theorem 5.11 of [7]. ∎

COROLLARY 7.4. *Assume that* J *and* f *are smooth. Let* u, E, H *be solutions of* (2.8), (2.9), (2.10), *and* (2.11) *with the boundary conditions* (2.6) *and* (2.7) *obtained by Theorem* 4.3. *Then* u *has the regularity* (7.1) *and* E *satisfies* (7.3).

Proof. This is a direct consequence of the previous theorem since $E = \frac{1}{i\omega}\nabla\varphi$. ◻

For the H-formulation, the arguments of the previous subsection directly give the next results.

THEOREM 7.5. *Let* $(u_\omega, H) \in H_0^1(\Omega)^3 \times X_T(\Omega, I)$ *be a solution of problem* (3.9) *in the sense of Lemma* 4.4 *with* J *and* f *smooth enough. Then the regularity results* (7.1)–(7.2) *hold,* τ_H *being defined as before but with* $\mu = I$.

COROLLARY 7.6. *Assume that* J *and* f *are smooth. Let* u, E, H *be solutions of* (2.8)*,* (2.9)*,* (2.10)*, and* (2.11) *with the boundary conditions* (2.6) *and* (2.7) *obtained by Theorem* 4.6*. Then* u *has the regularity* (7.1)*,* H *has the regularity* (7.2)*, and* E *satisfies* (7.3).

8. Numerical examples of edge singular exponents. To illustrate our theoretical results, we give some edge singular exponents of our piezoelectric system corresponding to some illustrative materials when $c_{45} = c_{16} = c_{26} = c_{36} = 0$, $c_{11} = c_{22}$, $c_{44} = c_{55}$, $c_{66} = \frac{c_{11}-c_{12}}{2}$, $\epsilon_{11} = \epsilon_{22}$, and $\epsilon_{12} = 0$. We further consider a dihedral cone $\Gamma_a \times \mathbb{R}$ such that the edge a of this cone (corresponding to the axis $z = 0$) is parallel to the poling direction x_3. In that case, the system (5.4) reduces to the following 4×4 system (see, for instance, [19]):

$$\begin{aligned}
&\begin{pmatrix} c_{11}\partial_1^2 + \frac{c_{11}-c_{12}}{2}\partial_2^2 & \frac{c_{11}+c_{12}}{2}\partial_1\partial_2 \\ \frac{c_{11}+c_{12}}{2}\partial_1\partial_2 & \frac{c_{11}-c_{12}}{2}\partial_1^2 + c_{11}\partial_2^2 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ in } \Gamma_a,\\
&\begin{pmatrix} c_{44}\Delta & e_{15}\Delta \\ -e_{15}\Delta & \epsilon_{11}\Delta \end{pmatrix}\begin{pmatrix} u_3 \\ \chi \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{in } \Gamma_a,\\
&\chi = 0, u = 0 \quad \text{on } \partial\Gamma_a,
\end{aligned}$$

where, as usual, $\Delta = \partial_1^2 + \partial_2^2$. Setting $v = c_{44}u_3 + e_{15}\chi$ and $w = -e_{15}u_3 + \epsilon_{11}\chi$, this system is equivalent to

$$\begin{aligned}
&\begin{pmatrix} c_{11}\partial_1^2 + \frac{c_{11}-c_{12}}{2}\partial_2^2 & \frac{c_{11}+c_{12}}{2}\partial_1\partial_2 \\ \frac{c_{11}+c_{12}}{2}\partial_1\partial_2 & \frac{c_{11}-c_{12}}{2}\partial_1^2 + c_{11}\partial_2^2 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ in } \Gamma_a,\\
&\Delta v = 0 \text{ in } \Gamma_a,\\
&\Delta w = 0 \text{ in } \Gamma_a,\\
&u_1 = u_2 = v = w = 0 \text{ on } \partial\Gamma_a.
\end{aligned}$$

This last system is decoupled into a 2×2 system of the isotropic elasticity for the pair (u_1, u_2) (with Lamé coefficients given by $\lambda = c_{12}$ and $\mu = \frac{c_{11}-c_{12}}{2}$) with Dirichlet boundary conditions and two Dirichlet problems for v and w. For this last Dirichlet problem, it is well known that the singular exponents are given by $\frac{l\pi}{\omega}$, with $l \in \mathbb{N}$, when ω is the interior opening of Γ_a. On the other hand, the singular exponents of the isotropic elasticity are also well known and are the set of $\lambda \in \mathbb{C}$ such that (see, for instance, [12])

$$k^2 \sin(\lambda\omega)^2 = \lambda^2 \sin(\omega)^2,$$

with $k = \frac{3c_{11}-c_{12}}{c_{11}+c_{12}}$.

The zeros of this equation can be easily computed using a Newton method. Figures 8.1 and 8.2 show the real part of the singular exponents in the strip $\Re\lambda \in (0, 2)$ for all values of $\omega \in (0, 2\pi]$ for the piezoelectric materials PZT-4 and $BaTiO_3$ in black lines. In these figures, the gray lines correspond to the curves $\frac{l\pi}{\omega}$, for $l = 1, 2, 3$. From these figures, we may conclude that for these materials the strongest singular

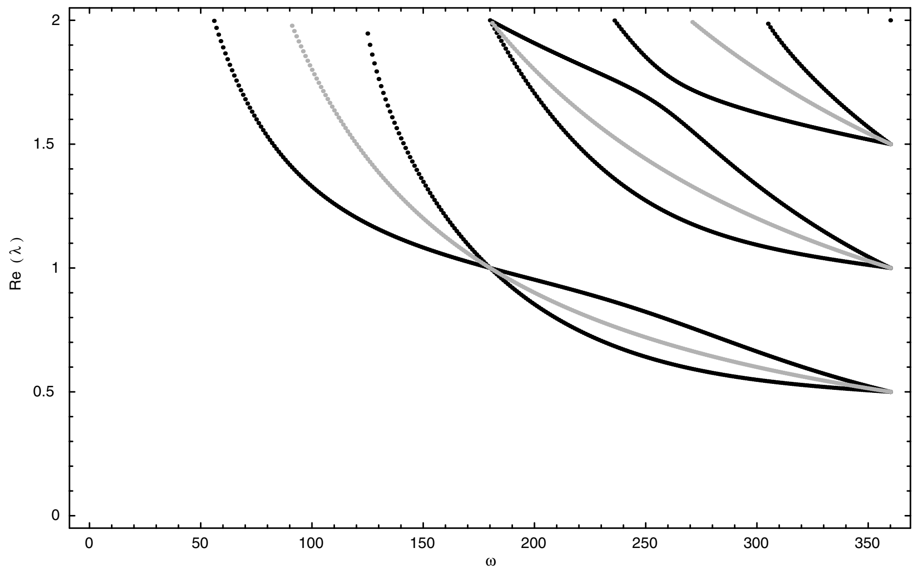


FIG. 8.1. *The edge exponents for the* PZT*-4.*

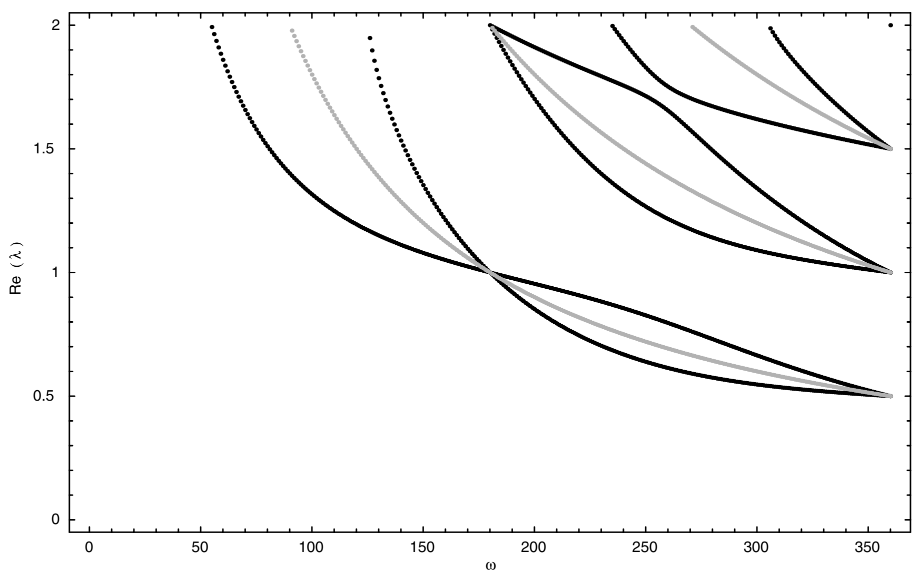


FIG. 8.2. *The edge exponents for the* $BaTiO_3$*.*

exponent λ_0 (strongest in the sense that $\Re\lambda_0$ is minimal) is always the one coming from the 2×2 system of elasticity and that $\Re\lambda_0 > 1/2$ (see Theorem 2.2 of [22]). This last property implies the $H^{3/2}$-regularity (resp., $H^{1/2}$-regularity) for u (resp., for E) along the edge. Note further that the curves obtained for the PZT-4 are similar to the ones shown in Figure 4b of [29] (in that paper, the singular exponents are obtained using an extension of Lekhnitskii's representation for elastic solids).

Remark 8.1. If the edge is not parallel to the poling direction, then the above decoupling phenomenon does not appear. The calculation of the edge singular exponents is then more complicated and will be done using a finite element method. This will be done in a forthcoming paper [20].

9. Conclusions. We have investigated a general time-harmonic piezoelectric system, which contains as special cases standard models like ceramics. We developed the appropriate formalism in order to get existence and uniqueness results of weak solutions in the case when the magnetic permeability matrix is positive definite (case of the $BaTiO_3$) and the case when the magnetic permeability matrix is equal to zero (case of the PZT). In the latter case, we gave two different formulations: the E-formulation and the H-formulation. For this second one, Gauge conditions are introduced as for eddy current problem. We further show that generically these two formulations yield the same solutions. We described the corner and edge singularities of our general system and deduced some regularity results. Some edge singular exponents were given in order to illustrate our theoretical results.

Acknowledgments. We thank C. Courtois and A. Leriche from the laboratory LMP (UVHC) for valuable discussions about these topics.

REFERENCES

[1] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional nonsmooth domains*, Math. Methods Applied Sci., 21 (1998), pp. 823–864.

[2] D. A. Berlincourt, D. R. Curran, and H. Jaffe, *Piezoelectric and piezomagnetic materials and their function in transducers*, Physical Acoustics, 1 (1964), pp. 169–270.

[3] A. Bossavit, *Electromagnétisme en vue de la modélisation*, Springer-Verlag, Paris, 1993.

[4] M. Costabel and M. Dauge, *Singularities of electromagnetic fields in polyhedral domains*, Arch. Ration. Mech. Anal., 151 (2000), pp. 221–276.

[5] M. Costabel, M. Dauge, and S. Nicaise, *Singularities of Maxwell interface problems*, RAIRO Modél. Math. Anal. Numér., 33 (1999), pp. 627–649.

[6] M. Costabel, M. Dauge, and S. Nicaise, *Singularities of eddy current problems*, RAIRO Modél. Math. Anal. Numér., 37 (2003), pp. 807–831.

[7] M. Dauge, *Elliptic Boundary Value Problems in Corner Domains. Smoothness and Asymptotics of Solutions*, L. N. in Math. 1341, Springer-Verlag, Berlin, 1988.

[8] A. C. Eringen and G. A. Maugin, *Electrodynamics of Continua*, Vols. 1, 2, Springer-Verlag, New York, 1990.

[9] W. Geis, A.-M. Sändig, and G. Mishuris, *Piezoelectricity in Multi-layer Actuators: Modelling and Analysis in Two and Three-Dimensions*, Preprint IANS, 2003/23, Univ. Stuttgart, Germany, 2003.

[10] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer Ser. Comput. Math. 5, Springer-Verlag, New York, 1986.

[11] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monogr. Stud. Math. 24, Pitman, Boston, 1985.

[12] P. Grisvard, *Singularité en elasticité*, Archiv. Ration. Mech. Anal., 107 (1989), pp. 157–180.

[13] D. Iesan, *Plane strain problems in piezoelectricity*, Internat. J. Engrg. Sci., 25 (1987), pp. 1511–1523.

[14] T. Ikeda, *Fundamentals of Piezoelectricity*, Oxford University Press, Oxford, 1996.

[15] V. A. Kondrat'ev, *Boundary-value problems for elliptic equations in domains with conical or angular points*, Trans. Moscow Math. Soc., 16 (1967), pp. 227–313.

[16] V. A. Kozlov and V. G. Maz'ya, *Spectral properties of the operator bundles generated by elliptic boundary value problems in a cone*, Funct. Anal. Appl., 22 (1988), pp. 38–46.

[17] V. A. Kozlov and V. G. Maz'ya, *On the behaviour of the spectrum of parameter-depending operators under small variation of the domain and an application to operator pencils generated by boundary value problems in a cone*, Math. Nachr., 153 (1991), pp. 123–129.

[18] V. A. Kozlov, V. G. Maz'ya, and J. Rossman, *Elliptic Boundary Value Problems in Domains with Point Singularities*, Math. Surveys Monogr. 52, AMS, Providence, RI, 1997.

[19] L. Marin, *The Behaviour of Piezoelectric Materials under Mechanical and Electrical Loadings*, Master Thesis, Univ. Kaiserslautern, Germany, 1998.

[20] D. Mercier and S. Nicaise, *Regularity Results for a Two-Dimensional Piezoelectric System and Crack Propagation Criteria*, manuscript.
[21] S. Nicaise, *Exact controllability of a pluridimensional coupled problem*, Rev. Mat. Complut., 5 (1992), pp. 91–135.
[22] S. Nicaise, *About the Lamé system in a polygonal or a polyhedral domain and a coupled problem between the Lamé system and the plate equation* I: *Regularity of the solutions*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 19 (1992), pp. 327–361.
[23] S. Nicaise, *Polygonal Interface Problems*, Peter Lang, Berlin, 1993.
[24] W. Novacki, *Mathematical models of phenomenological piezoelectricity*, in Mathematical Models and Methods in Mechanics, Banach Center Publ. 15, Polish Academy of Sciences, Warsaw, Poland, 1985, pp. 593–607.
[25] Y. E. Pak, *Linear electro-elastic fracture mechanics of piezoelectric materials*, Internat. J. Fracture, 54 (1992), pp. 79–100.
[26] E. Pan and P. R. Heyliger, *Exact solutions for magneto-electro-elastic laminates in cylindrical bending*, Internat. J. Solids Structures, 40 (2003), pp. 6859–6876.
[27] H. A. Sosa and Y. E. Pak, *Three-dimensional eigenfunction analysis of a crack in a piezoelectric material*, Internat. J. Solids Structures, 26 (1990), pp. 1–15.
[28] Ch. Weber, *A local compactness theorem for Maxwell's equations*, Math. Methods Appl. Sci., 2 (1980), pp. 12–25.
[29] X.-L. Xu and R. Rajapakse, *On singularities in composite piezoelectric wedges and junctions*, Internat. J. Solids Structures, 37 (2000), pp. 3253–3275.
[30] X.-L. Xu and R. Rajapakse, *On a plane crack in piezoelectric solids*, Internat. J. Solids Structures, 38 (2001), pp. 7643–7658.

SIAM J. MATH. ANAL.
Vol. 37, No. 2, pp. 673–683

APPLICATION OF THE REALIZATION OF HOMOGENEOUS SOBOLEV SPACES TO NAVIER–STOKES*

LORENZO BRANDOLESE†

Abstract. Molecule spaces have been introduced by Furioli and Terraneo [*Funkcial. Ekvac.*, 45 (2002), pp. 141–160] to study some local behavior of solutions to the Navier–Stokes equations. In this paper we give a new characterization of these spaces and simplify Furioli and Terraneo's result. Our analysis also provides a persistence result for Navier–Stokes in a subspace of $L^2(\mathbb{R}^3, (1+|x|^2)^\alpha dx)$, $\alpha < 5/2$, which fills a gap between previously known results in the weighted-L^2 setting and those on the pointwise decay of the velocity field at infinity. Our main tool is the realization of homogeneous Sobolev spaces introduced by Bourdaud.

Key words. molecules, Hardy space, pointwise multipliers

AMS subject classifications. 36B40, 76D05, 35Q30

DOI. 10.1137/S0036141004444408

1. Introduction. Consider the Navier–Stokes equations for a viscous incompressible fluid in the three-dimensional space and not submitted to external forces:

$$\text{(NS)} \qquad \begin{cases} \partial_t u + (u\cdot\nabla)u = \Delta u - \nabla p, \\ \nabla\cdot u = 0, \\ u(x,0) = a(x). \end{cases}$$

Here $u = (u_1, u_2, u_3)$ is the velocity field and p is the pressure, both defined in $\mathbb{R}^3 \times [0,\infty[$. Moreover, $\nabla\cdot u = \sum_{j=1}^3 \partial_j u_j$ and $(u\cdot\nabla)u = \sum_{j=1}^3 u_j\partial_j u$.

If $a \in L^2(\mathbb{R}^3)$, then we have known for a very long time that a *weak* solution to (NS) exists such that $u \in L^\infty(]0,\infty[, L^2(\mathbb{R}^3))$ and $\nabla u \in L^2(]0,\infty[, L^2(\mathbb{R}^3))$. If we know, in addition, that the initial datum is well localized in $\mathbb{R}^3$, then these conditions, of course, do not give us so much information on the spatial localization of $u(t)$ during the evolution. Then the natural problem arises of finding the functional spaces that would provide the good setting for obtaining such information. Several papers have been written on this topic; see, e.g., [14], [9], [10], [13], [1], [16] and the references therein. In particular, it was shown in [14] that the condition $a \in L^2(\mathbb{R}^3, (1+|x|^2)^\delta dx)$ $(0 \le \delta \le \frac{3}{2})$ is conserved during the evolution, for a suitable class of weak solutions. Here and below, this weighted-L^2 space is equipped with the natural norm $\left(\int |a(x)|^2(1+|x|^2)^\delta\, dx\right)^{1/2}$. As far as we deal with data belonging to general weighted-L^2 spaces, it seems difficult to improve the upper bound on δ.

When dealing with *strong* solutions to (NS) one can obtain sharper conclusions on the localization of u. For example, assuming that $a \in L^1 \cap L^2(\mathbb{R}^3)$, He [13] proves, among other things, that $u(t)$ belongs to $L^2(\mathbb{R}^3, (1+|x|^2)^2 dx)$ at least in some time interval $[0,T]$, $T > 0$ (and uniformly in $[0,+\infty[$, under a supplementary smallness assumption). In a slightly different context, we would also like to mention the work of Miyakawa [17], in which it is shown that $u(x,t) \sim |x|^{-\alpha} t^{-\beta/2}$ as $|x| \to \infty$ or $t \to \infty$,

*Received by the editors June 15, 2004; accepted for publication (in revised form) April 6, 2005; published electronically November 9, 2005.

http://www.siam.org/journals/sima/37-2/44440.html

†Institut Camille Jordan Université Claude Bernard Lyon 1, 21 avenue Claude Bernard 69622 Villeurbanne Cedex, France (brandolese@math.univ-lyon1.fr).

for all $\alpha, \beta \geq 0$ and $1 \leq \alpha + \beta \leq 4$, under suitable assumptions on a. The main tool here is the application of the contraction mapping theorem to the integral equation

$$u(t) = e^{t\Delta}a - \int_0^t \nabla \cdot e^{(t-s)\Delta}\mathbb{P}(u \otimes u)(s)\, ds, \tag{IE}$$

where $e^{t\Delta}$ is the heat semigroup and $\mathbb{P}$ is the Leray–Hopf projector onto the solenoidal vector fields, defined by $\mathbb{P}f = f - \nabla\Delta^{-1}(\nabla \cdot f)$, where $f = (f_1, f_2, f_3)$. Note that (IE), together with the divergence-free condition $\nabla \cdot a = 0$, is equivalent to (NS) under very general assumptions (see [11]).

If we compare the results on the spatial localization contained in [13] and [17], we see that Miyakawa's results seem to give a slightly better conclusion. Indeed, [17] tells us that the condition $a(x) \sim |x|^{-4}$ at infinity is conserved during the evolution (furthermore, $|x|^{-4}$ is known to be the optimal decay in the generic case), whereas, according to [13], the condition $u(t) \in L^2(\mathbb{R}^3, (1+|x|^2)^2 dx)$ only tells us, formally, that $u(t) \sim |x|^{-7/2}$ at infinity. Then there is a small gap between the results on the pointwise decay and those in the weighted-L^2 setting.

The first purpose of this paper is to obtain a persistence result in suitable subspaces of $L^2(\mathbb{R}^3, (1+|x|^2)^\alpha dx)$, for all $0 \leq \alpha < \frac{5}{2}$, which, at least formally, will allow us to recover the optimal decay of the velocity field. To do this, rather than establishing a new theorem we shall give a new interpretation of a known result by Furioli and Terraneo on the molecules of the Hardy space [12]. More precisely, let us introduce the space Z_δ of functions (or vector fields) f such that

$$f \in L^2(\mathbb{R}^3, (1+|x|^2)^{\delta-2} dx), \tag{1.1}$$

$$\nabla f \in L^2(\mathbb{R}^3, (1+|x|^2)^{\delta-1} dx), \tag{1.2}$$

$$\Delta f \in L^2(\mathbb{R}^3, (1+|x|^2)^{\delta} dx). \tag{1.3}$$

We provide such space with its natural norm. We will prove the following theorem (announced, in a weaker form, in [5]).

THEOREM 1.1. *Let $\frac{1}{2} \leq \delta < \frac{9}{2}$ and let $a \in Z_\delta$ be a solenoidal vector field. Then there exists $T > 0$ such that* (IE) *possesses a unique strong solution $u \in C([0,T], Z_\delta)$.*

The restriction $\delta < \frac{9}{2}$ is consistent with the spatial spreading effect of the velocity field described, e.g., in [6]: we cannot have $u \in C([0,T], Z_{9/2})$ unless the initial data have some symmetry properties. As we shall see, the elements of Z_δ are $o(|x|^{-\delta+1/2})$ at infinity. Hence the correspondence between this result and those on the pointwise decay is not merely formal. The condition $\delta \geq \frac{1}{2}$ is physically reasonable since it prevents $u \to \infty$ as $|x| \to \infty$.

In section 3 we show that Theorem 1.1 is essentially equivalent to (but slightly improves) the result by Furioli and Terraneo [12]. Their motivation was different and this is probably the reason why the relation between their *space of molecules* X_δ (defined below) and the more natural space Z_δ is not found in [12]; motivated by the problem of the unicity of mild solutions in *critical spaces* (i.e., homogeneous spaces of degree -1), they studied the Navier–Stokes equations in $\Delta^{-1}\mathcal{H}^1$, which is the space made of all distributions vanishing at infinity and such that their Laplacian belongs to the Hardy space $\mathcal{H}^1(\mathbb{R}^3)$. As discussed also in [15], such space gives a useful insight of solutions to the Navier–Stokes equations. Indeed, if a solution u satisfies $\Delta u \in \mathcal{H}^1(\mathbb{R}^3)$, then $u \in L^3(\mathbb{R}^3)$, $\nabla \otimes u \in L^{3/2}$, and $(u \cdot \nabla)u \in \mathcal{H}^1(\mathbb{R}^3)$ (this is a consequence of the so-called div-curl lemma as stated in [7]). Furthermore, $\nabla p \in \mathcal{H}^1(\mathbb{R}^3)$ (this follows from the classical relation $\Delta p = \sum_{h,k=1}^3 \partial_h\partial_k(u_h u_k)$ and the boundedness of the Riesz transforms in the Hardy space). Thus, the three terms which contribute to $\partial_t u$ in (NS) have the same regularity.

Moreover, the Hardy space has a very simple structure, due to its well-known atomic decomposition. Hence, solving the equations in $\Delta^{-1}\mathcal{H}^1$ yields a natural decomposition of the flow into simple "building blocks." Furioli and Terraneo considered the converse problem of studying the evolution of each building block. The result of [12] essentially states that if Δu is a *molecule* of the Hardy space (in a sense close to that of Coifman and Weiss [8]) at the beginning of the evolution, then this property remains true for a certain time. To do this they introduced, for $\delta > \frac{3}{2}$, the space X_δ defined as the set of all tempered distributions f vanishing at infinity such that $\Delta f \in L^2(\mathbb{R}^3, (1+|x|^2)^\delta dx)$ and $\int x^\alpha \Delta f(x)\, dx = 0$ for all $\alpha \in \mathbb{N}^3$ such that $\alpha_1 + \alpha_2 + \alpha_3 < \delta - \frac{3}{2}$. The norm of X_δ is defined by

$$||f||^2_{X_\delta} \equiv \int |\Delta f(x)|^2 (1+|x|^2)^\delta \, dx.$$

Furioli and Terraneo's theorem then is stated as Theorem 1.1, with X_δ instead of Z_δ, and with the additional restrictions $\frac{3}{2} < \delta < \frac{9}{2}$ and $\delta \neq \frac{5}{2}, \frac{7}{2}$. The condition $\delta > \frac{3}{2}$ is important for the embedding $X_\delta \subset \Delta^{-1}\mathcal{H}^1$. However, P.-G. Lemarié-Rieusset noticed that the condition $\delta \neq \frac{5}{2}, \frac{7}{2}$ can be removed. Their paper is technical and relies on the theory of *local* Muckenhoupt weights.

The second purpose of this paper is to provide a simpler proof of their result. Indeed, in section 4 we remark that the Fourier transform of X_δ is closely related to the realization "à la Bourdaud" [3] of the homogeneous Sobolev space $\dot{H}^\delta$. The operators involved in (IE) turn out to be Fourier pointwise multipliers of the realized spaces. Therefore the estimates that are needed to establish the boundedness of the bilinear operator $B(u,v)(t) = \int_0^t \nabla \cdot e^{(t-s)\Delta}\mathbb{P}(u \otimes v)(s)\, ds$ in $C([0,T], X_\delta)$ become very natural. The conclusion of the proof is a simple application of the contraction mapping theorem.

The idea of using Bourdaud's results on realized spaces in this way (or analogous results by Youssfi [20] for the realized homogeneous Besov spaces) seems to be new. Since this argument does not directly rely on the divergence-free condition or the matricial structure of $\mathbb{P}$, it can be easily applied to more general equations. Moreover, we feel that providing evidence of the relation between the localization problem of the velocity field and Furioli and Terraneo's molecules provides a better understanding of [12].

The spatial localization of the velocity field in different weighted-Lebesgue spaces is studied in [19]. After the first version of this paper was completed, the author was notified by H.-O. Bae and B. J. Jin that their preprint [2] also improves the spatial decay results of [13] and [14] and provides solutions to (NS) in $L^2(\mathbb{R}^3, (1+|x|^2)^\alpha dx)$, $0 \leq \alpha < \frac{5}{2}$. Their method is a refinement of the weighted estimates of He and Xin [14] and is quite different than ours. The assumptions on the data are also different: Bae and Jin deal with less regular data, but they put more stringent assumptions on their spatial localization.

2. Some properties of the space . For $\delta \geq 0$ we introduce the space L^2_δ of all functions $f \in L^2(\mathbb{R}^3, (1+|x|^2)^\delta dx)$ such that $\int x^\alpha f(x)\, dx = 0$ for all $\alpha \in \mathbb{N}^3$, with $0 \leq |\alpha| < \delta - \frac{3}{2}$ (where $|\alpha| = \alpha_1 + \alpha_2 + \alpha_3$). There are of course no moment conditions for $0 \leq \delta \leq \frac{3}{2}$. Note that L^2_δ is well defined because of the embedding of $L^2(\mathbb{R}^3, (1+|x|^2)^\delta dx)$ into $L^1(\mathbb{R}^3, w_\delta(x)dx)$, for $\delta > \frac{3}{2}$. Here and below, for $\delta > \frac{3}{2}$, we set $w_\delta(x) = (1+|x|)^{[\delta-3/2]}$ if $\delta - \frac{3}{2} \notin \mathbb{N}$ (where $[\cdot]$ denotes the integer part), and $w_\delta(x) = (1+|x|)^{\delta-5/2}$ otherwise.

LEMMA 2.1. *Let $\delta \geq 0$. We have $f \in L^2_\delta$ if and only if f can be decomposed as*

$$f = g + \sum_{j=0}^{\infty} f_j,$$

where g and f_j belong to $L^2(\mathbb{R}^3)$, $\operatorname{supp} g \subset \{|x| \leq 1\}$, $\operatorname{supp} f_j \subset \{2^{j-1} \leq |x| \leq 2^{j+1}\}$, and

$$\begin{aligned} &\|f_j\|_2 \leq \epsilon_j 2^{-j\delta}, \qquad \text{with} \quad \epsilon_j \in \ell^2(\mathbb{N}), \\ &\int x^\alpha g(x)\,dx = \int x^\alpha f_j(x)\,dx = 0, \qquad \text{if } \alpha \in \mathbb{N}^3, \quad 0 \leq |\alpha| < \delta - \frac{3}{2}, \end{aligned} \tag{2.1}$$

and where the series converges a.e. in $\mathbb{R}^3$ and in L^2_δ.

Proof. The result is obvious for $0 \leq \delta \leq \frac{3}{2}$, so we may assume $\delta > \frac{3}{2}$. We start with a bad choice, namely

$$\tilde{g}(x) = f(x) I_{|x|\leq 1} \quad \text{and} \quad \tilde{f}_j(x) = f(x) I_{2^j \leq |x| \leq 2^{j+1}} \qquad (j = 0, 1, \ldots),$$

where I denotes the indicator function. Letting $\tilde{f}_{-1} = \tilde{g}$, we set

$$J(j,\alpha) = \int x^\alpha \tilde{f}_j(x)\,dx.$$

Since $|\alpha| < \delta - \frac{3}{2}$, the series $\sum_j J(j,\alpha)$ converges and $\sum_{j=-1}^{\infty} J(j,\alpha) = 0$. We now introduce a family of functions $\psi_\beta \in C_0^\infty(\mathbb{R}^3)$, supported in $\frac{1}{2} \leq |x| \leq 1$ and such that

$$\int x^\alpha \psi_\beta(x)\,dx = \delta_{\alpha,\beta} \qquad (\alpha, \beta \in \mathbb{N}^3, \quad |\alpha|, |\beta| < \delta - \tfrac{3}{2}),$$

with $\delta_{\alpha,\beta} = 0$ or 1 if $\alpha \neq \beta$ or $\alpha = \beta$, respectively (we may define ψ_β, e.g., through the tensorial product of suitable $C_0^\infty(\mathbb{R})$ functions).

Now let

$$c(j,\alpha) \equiv J(j,\alpha) + J(j+1,\alpha) + \cdots,$$

and set, for $j = -1, 0, \ldots,$

$$f_j(x) = \tilde{f}_j(x) - \sum_\beta \Big(c(j,\beta) 2^{-(3+|\beta|)j} \psi_\beta(2^{-j}x) - c(j+1,\beta) 2^{-(3+|\beta|)(j+1)} \psi_\beta(2^{-j-1}x) \Big),$$

the summation being taken over all $\beta \in \mathbb{N}^3$ such that $0 \leq |\beta| < \delta - \frac{3}{2}$.

Since $|J(j,\alpha)| \leq 2^{j|\alpha|} 2^{3j/2 - j\delta} \tilde{\epsilon}_j$ for some $\tilde{\epsilon}_j \in \ell^2(\mathbb{N})$, we have $|c(j,\alpha)| \leq 2^{j|\alpha|} 2^{3j/2-j\delta} \bar{\epsilon}_j$, with $\bar{\epsilon}_j \in \ell^2(\mathbb{N})$. One now easily checks that $\sum_{j=-1}^{\infty} f_j = \sum_{j=-1}^{\infty} \tilde{f}_j = f$ a.e. in $\mathbb{R}^3$ and that g and f_j satisfy (2.1).

The converse is immediate. Note that if g and f_j satisfy the above conditions, then $\sum f_j$ must converge also in the $L^1(\mathbb{R}^3, w_\delta(x)dx)$ norm ($\delta > \frac{3}{2}$) by Hölder's inequality, and this ensures the condition on the moments of f. Lemma 2.1 follows. $\square$

A similar decomposition applies to Z_δ.

LEMMA 2.2. *Let $\delta \geq 0$. We have $f \in Z_\delta$ if and only if*

$$f = g + \sum_{j=0}^{\infty} f_j,$$

with

$$\begin{aligned}
&\operatorname{supp} g \subset \{|x| \leq 1\}, && \operatorname{supp} f_j \subset \{2^{j-2} \leq |x| \leq 2^{j+2}\}, && \\
&g, \nabla g, \Delta g \in L^2(\mathbb{R}^3), && ||f_j||_2 \leq \epsilon_j 2^{2j} 2^{-j\delta}, && \epsilon_j \in \ell^2(\mathbb{N}), \\
&||\nabla f_j||_2 \leq \bar{\epsilon}_j 2^j 2^{-j\delta}, && ||\Delta f_j||_2 \leq \tilde{\epsilon}_j 2^{-j\delta}, && \bar{\epsilon}_j, \tilde{\epsilon}_j \in \ell^2(\mathbb{N}),
\end{aligned} \tag{2.2}$$

and where the series converges a.e. in $\mathbb{R}^3$ and in Z_δ.

If $\delta > \frac{7}{2}$, then Z_δ is continuously embedded in $L^1(\mathbb{R}^3)$. In this case we have $\int f = 0$ if and only if we may choose g and f_j satisfying, in addition, $\int g = \int f_j = 0$ ($j = 0, 1, \ldots$).

Proof. It is obvious that if (2.2) holds, then $f = g + \sum_{j=0}^{\infty} f_j$ belongs to Z_δ. Conversely, let φ and ψ be two compactly supported smooth functions, such that 0 does not belong to the support of ψ and $1 \equiv \varphi(x) + \sum_{j=0}^{\infty} \psi(2^{-j}x)$. If we set $g(x) = f(x)\varphi(x)$ and $f_j(x) = f(x)\psi(2^{-j}x)$, then we have $\nabla f_j(x) = \psi(2^{-j}x)\nabla f(x) + 2^{-j}(\nabla\psi)(2^{-j}x)f(x)$ and

$$\Delta f_j(x) = \psi(2^{-j}x)\Delta f(x) + 2^{-j+1}(\nabla\psi)(2^{-j}x) \cdot \nabla f(x) + 2^{-2j}(\Delta\psi)(2^{-j}x)f(x).$$

Decomposition (2.2) then directly follows from the definition of Z_δ.

If $\delta > \frac{7}{2}$, then $Z_\delta \subset L^1(\mathbb{R}^3)$ as checked with Hölder's inequality. In this case, when $\int f = 0$, we can modify the definition of g and f_j reproducing the proof of Lemma 2.1 (with $|\alpha| = |\beta| = 0$) and get the vanishing integral conditions. Lemma 2.2 follows. $\square$

We finish our study of Z_δ with the following lemma.

LEMMA 2.3. *Let $\delta \geq \frac{1}{2}$. Then Z_δ is an algebra with respect to the pointwise product. More precisely, if f and h belong to Z_δ, then $fh \in Z_{2\delta-1/2} \subset Z_\delta$.*

Proof. The condition $\delta \geq \frac{1}{2}$ ensures that if $f \in Z_\delta$, then f vanishes at infinity. Indeed, we have the following bound:

$$|f(x)| \leq C(1 + |x|)^{-\delta+1/2}\epsilon(x), \tag{2.3}$$

where $\epsilon(x)$ is a bounded function vanishing at infinity. This is seen by applying Lemma 2.2 and writing, for $x \in \operatorname{supp} f_j$, $f_j(x) = \int |x-y|^{-1}\Delta f_j(y)\,dy$. Then applying Hölder's inequality and the last of (2.2) we get for $x \in \operatorname{supp} f_j$, $|f_j(x)| \leq \epsilon_j 2^{-j\delta}2^{j/2}$ with $\epsilon_j \in \ell^2(\mathbb{N})$ and our claim follows.

Another useful estimate (which follows interpolating ∇f_j between $||\Delta f_j||_2$ and $||f_j||_\infty$) is

$$\left(\int_{2^j \leq |x| \leq 2^{j+1}} |\nabla f(x)|^4 dx\right)^{1/4} \leq \epsilon_j 2^{j/4 - j\delta}, \qquad \text{with } \epsilon_j \in \ell^2(\mathbb{N}).$$

Using this, we immediately see that if f and h belong to Z_δ, then $fh \in L^2(\mathbb{R}^3, (1 + |x|^2)^{2\delta-5/2}dx)$, $\nabla(fh) \in L^2(\mathbb{R}^3, (1 + |x|^2)^{2\delta-3/2}dx)$, and $\Delta(fh) \in L^2(\mathbb{R}^3, (1 + |x|^2)^{2\delta-\frac{1}{2}}dx)$. Therefore $fh \in Z_{2\delta-1/2} \subset Z_\delta$ and, moreover,

$$||fh||_{Z_{2\delta-1/2}} \leq C||f||_{Z_\delta}||h||_{Z_\delta}. \qquad \square$$

3. Characterization of molecule spaces. We defined the molecule space X_δ in the introduction for $\delta > \frac{3}{2}$. These spaces can be defined also for $\frac{1}{2} < \delta \le \frac{3}{2}$ by simply dropping the moment conditions on Δf (it should be observed that the embedding $X_\delta \subset \Delta^{-1}\mathcal{H}^1(\mathbb{R}^3)$ breaks down for $\delta \le \frac{3}{2}$, but for the sake of brevity we refer to X_δ as a "molecule space" also in this case). In this section we study the relation between the spaces Z_δ and X_δ.

PROPOSITION 3.1. *We have*

$$X_\delta = Z_\delta, \qquad \text{if } \frac{1}{2} < \delta < \frac{7}{2} \quad \text{and} \quad \delta \neq \frac{3}{2}, \frac{5}{2}, \tag{3.1}$$

$$X_\delta = Z_\delta \cap \left\{ f \in L^1(\mathbb{R}^3) \,:\, \int f = 0 \right\}, \qquad \text{if } \frac{7}{2} < \delta < \frac{9}{2} \tag{3.2}$$

(with norm equivalence).

Note that $\delta = \frac{3}{2}, \frac{5}{2}, \frac{7}{2}$ are excluded. The proof of Proposition 3.1 will only give $Z_\delta \subset X_\delta$ in this case. Let us shed some light on this point with an example. Let $j,k = 1,2,3$ and $f \in L^2(\mathbb{R}^3)$ such that $\Delta f = \partial_j \partial_k g$, where $g(x) = (4\pi)^{-3/2} e^{-|x|^2/4}$. Then $f \in X_{7/2}$ but $f \notin X_\delta$ for $\delta > \frac{7}{2}$, since $\int x_j x_k \Delta f(x)\, dx \neq 0$. Moreover, computing the inverse Fourier transform as in [17] from the identity $\widehat{f}(\xi) = \xi_j \xi_k |\xi|^{-2} e^{-|\xi|^2}$, one checks after some computations that the integral $\int |f(x)|^2 (1+|x|^2)^{3/2}\, dx$ diverges so that $f \notin Z_{7/2}$.

Proof of Proposition 3.1. Throughout the proof $\alpha \in \mathbb{N}^3$. Note that because of our restrictions $\delta - \frac{3}{2}$ is not an integer.

First step: The embedding $X_\delta \subset Z_\delta$. Let $f \in X_\delta$. Applying Lemma 2.1 to Δf and using the fact that f vanishes at infinity, we see that we may write

$$f = \frac{c}{|x|} * p + \sum_{j=0}^{\infty} \frac{c}{|x|} * q_j, \tag{3.3}$$

c being an absolute constant. Here p and q_j are compactly supported L^2-functions, satisfying

$$\begin{aligned}
&\operatorname{supp} p \subset \{|x| \le 1\},\\
&\operatorname{supp} q_j \subset \{2^{j-1} \le |x| \le 2^{j+1}\},\\
&||q_j||_2 \le \epsilon_j 2^{-j\delta}, \qquad \epsilon_j \in \ell^2(\mathbb{N}),\\
&\int x^\alpha p(x)\, dx = \int x^\alpha q_j(x)\, dx = 0, \qquad \text{if } |\alpha| < \delta - \frac{3}{2}.
\end{aligned} \tag{3.4}$$

Let us show that, for all $f \in X_\delta$ and $2^j \le |x| \le 2^{j+1}$, we have

$$|f(x)| \le \bar{\epsilon}_j 2^{j/2} 2^{-j\delta}, \qquad \text{with } \bar{\epsilon}_j \in \ell^2(\mathbb{N}). \tag{3.5}$$

To prove (3.5) we set $P = \frac{1}{|\cdot|} * p$, $Q_j = \frac{1}{|\cdot|} * q_j$ $(j = 0, 1, \ldots)$, and $d = [\delta - \frac{3}{2}]$ (we set $d = -1$ if $\frac{1}{2} < \delta < \frac{3}{2}$). Then we have

$$|Q_j(x)| \le C \epsilon_j 2^{-j(\delta - 1/2)}, \qquad \text{if } |x| \le 4 \cdot 2^j, \tag{3.6}$$

$$|Q_j(x)| \le C |x|^{-(d+2)} \epsilon_j 2^{(d+5/2-\delta)j}, \qquad \text{if } |x| \ge 4 \cdot 2^j. \tag{3.7}$$

The first bound follows from the localization of q_j and Hölder's inequality. Let us prove (3.7); we start by introducing the Taylor polynomial $y \mapsto T_x(y)$ of degree d

centered at x of the function $1/|y|$ (we set $T_x(y) \equiv 0$ for $\frac{1}{2} < \delta < \frac{3}{2}$). Then for $|x| \geq 4 \cdot 2^j$ and $y \in \operatorname{supp} q_j$ we may write, using the last of (3.4),

$$|Q_j(x)| = \left| \int \left(\frac{1}{|x-y|} - T_x(-y) \right) q_j(y)\, dy \right| \leq C|x|^{-d-2} \int |y|^{d+1} |q_j(y)|\, dy.$$

Here the inequality follows from the Taylor formula and the fact that the $(d+1)$-order derivatives of $y \mapsto 1/|y|$ are bounded in a ball centered at x and radium $|x|/2$, up to a constant, by $|x|^{-d-2}$. The bound (3.7) now follows from Hölder's inequality.

Similar arguments allow us to see that $|P(x)| \leq C(1+|x|)^{-(d+2)}$. Summing up on these inequalities immediately yields (3.5).

Condition (3.6) also ensures that $\int_{|x|\leq 1} |f|^2$ is finite. Then using (3.5) we get

$$\int |f(x)|^2 (1+|x|^2)^{\delta-2}\, dx < \infty. \tag{3.8}$$

We now need some bounds for ∇f. We start from $-\nabla f(x) = \frac{cx}{|x|^3} * p + \sum_{j=0}^\infty \frac{cx}{|x|^3} * q_j$, and we set $R_j = (x/|x|^3) * q_j$ $(j = 0, 1, \ldots)$ and $R_{-1} = (x/|x|^3) * p$. Then for $j \geq -1$,

$$|R_j(x)| \leq C|x|^{-(d+3)} \epsilon_j 2^{(d+5/2-\delta)j}, \qquad \text{if } |x| \geq 4 \cdot 2^j, \tag{3.9}$$

$$\left(\int_{|x|\leq 4\cdot 2^j} |R_j(x)|^4\, dx \right)^{1/4} \leq C\epsilon_j 2^{j/4} 2^{-j\delta}. \tag{3.10}$$

Indeed, (3.9) again easily follows using the vanishing of the moments of q_j and the Taylor formula. The proof of (3.10) deserves a more detailed explanation: for $|x| \leq 4 \cdot 2^j$ we write

$$\frac{x}{|x|^3} * q_j(x) = \theta_j * q_j(x), \qquad \text{where} \quad \theta_j(x) = \frac{x}{|x|^3} I_{\{|x|\leq 6\cdot 2^j\}}.$$

Then (3.10) comes from $||\theta_j * q_j||_4 \leq ||\theta_j||_{4/3} ||q_j||_2 \leq C 2^{j/4} ||q_j||_2$.

Now, for $j \geq 1$ we write $-\frac{1}{c}\nabla f = \sum_{k=-1}^{j-2} R_k + \sum_{k=j-1}^\infty R_k \equiv A_j + B_j$. Using (3.9) we get

$$\left(\int_{2^j \leq |x| \leq 2^{j+1}} |A_j(x)|^2\, dx \right)^{1/2} \leq C\tilde\epsilon_j 2^j 2^{-j\delta}, \qquad \text{with } \tilde\epsilon_j \in \ell^2(\mathbb{N}).$$

On the other hand, applying Hölder and Minkowski inequalities and (3.10) yields

$$\left(\int_{2^j \leq |x| \leq 2^{j+1}} |B_j(x)|^2\, dx \right)^{1/2} \leq C'\tilde\epsilon_j 2^j 2^{-j\delta}, \qquad \text{with } \tilde\epsilon_j \in \ell^2(\mathbb{N}).$$

Since we obviously have $\int_{|x|\leq 2} |\nabla f|^2 < \infty$, we thus see that

$$\int |\nabla f(x)|^2 (1+|x|^2)^{\delta-1}\, dx < \infty. \tag{3.11}$$

This last inequality, condition (3.8), and the definition of the X_δ norm yield the injection $X_\delta \subset Z_\delta$.

Second step: The elements of X_δ, $\frac{7}{2} < \delta < \frac{9}{2}$ have vanishing integral. Assume now $\frac{7}{2} < \delta < \frac{9}{2}$. Then the moments of p and q_j vanish up to the order two. Moreover, our previous estimates imply that P and Q_j belong to $L^1(\mathbb{R}^3)$. We thus see, e.g., via the Fourier transform (using the fact that $\widehat{p}(\xi)$ and $\widehat{q}_j(\xi)$ vanish at the origin together with their derivatives up to the order two and letting $\xi \to 0$) that $\int P(x)\,dx = \int Q_j(x)\,dx = 0$ for all positive integers j. Moreover, the series $\sum Q_j$ converges in the L^1-norm by (3.6)–(3.7) yielding $\int f = 0$.

Third step: The converse inclusion. Let $f \in Z_\delta$. In the case $\frac{7}{2} < \delta < \frac{9}{2}$ we assume $\int f = 0$. The bound $||f||_{X_\delta} < \infty$ is obvious. By Lemma 2.2 we have $f = g + \sum_{j=0}^{\infty} f_j$, such that (2.2) holds (with $\int g = \int f_j = 0$ for all $j \geq 0$ if $\frac{7}{2} < \delta < \frac{9}{2}$).

We claim that

$$\int x^\alpha \Delta g(x)\,dx = \int x^\alpha \Delta f_j(x)\,dx = 0, \qquad 0 \leq |\alpha| < \delta - \tfrac{3}{2},$$

for all $j \geq 0$ (there are no moment conditions for $\frac{1}{2} < \delta < \frac{3}{2}$). Indeed, since g and f_j are compactly supported, when applying the Green formula all the boundary terms vanish and we obtain (e.g., for f_j, when $\frac{7}{2} < \delta < \frac{9}{2}$)

$$\int x^\alpha \Delta f_j(x)\,dx = 2 \int f_j(x)\,dx, \qquad \text{if} \quad x^\alpha = x_1^2,\ x_2^2, \text{ or } x_3^2,$$

and $\int x^\alpha \Delta f_j(x)\,dx = 0$; otherwise ($|\alpha| \leq [\delta - \frac{3}{2}]$). Our claim then follows.

Moreover, by Hölder's inequality,

$$\sum_{j=0}^{\infty} ||x^\alpha \Delta f_j||_1 < \infty, \qquad 0 \leq |\alpha| < \delta - \tfrac{3}{2}.$$

Summing on j we get $\int x^\alpha \Delta f(x)\,dx = 0$.

To conclude that $f \in X_\delta$ it remains to check that f vanishes at infinity. This was done in (2.3). ◻

4. Proof of Theorem 1.1. The boundedness of the operator $\nabla e^{t\Delta}\mathbb{P}$ in X_δ ($\frac{3}{2} < \delta < \frac{9}{2}$), $\delta \neq \frac{5}{2}, \frac{7}{2}$ is a fundamental step of [12]. Lemma 4.1 provides a short proof of this fact. Our main tool will be the *realization* of homogeneous Sobolev spaces introduced by Bourdaud. Note that $\nabla e^{t\Delta}\mathbb{P}$ is a matrix operator acting on vector fields. But its matricial structure has no special role in what follows, since we shall establish all the relevant estimates componentwise.

LEMMA 4.1. *Let $\frac{1}{2} < \delta < \frac{9}{2}$, $\delta \neq \frac{3}{2}, \frac{5}{2}, \frac{7}{2}$. The operator $\nabla e^{t\Delta}\mathbb{P}$ is bounded from Z_δ to X_δ for all $t > 0$, with operator norm $O(t^{-1/2})$ as $t \to 0$.*

Proof. Let $f \in Z_\delta$. If $\frac{7}{2} < \delta < \frac{9}{2}$, then we introduce a function h such that

$$f(x) = cg(x) + h(x), \qquad \text{where } g(x) = (4\pi)^{-3/2} e^{-|x|^2/4}$$

and the constant c is chosen in a such way that $\int h(x)\,dx = 0$. If, instead, $\frac{1}{2} < \delta < \frac{7}{2}$, $\delta \neq \frac{3}{2}, \frac{5}{2}$, then we simply set $f(x) = h(x)$. In any case, we deduce from Proposition 3.1 that $h \in X_\delta$ and $||h||_{X_\delta} \leq C||f||_{Z_\delta}$ for some constant C depending only on δ.

We start showing that $\nabla e^{t\Delta}\mathbb{P}g$ belongs to X_δ for all $0 \leq \delta < \frac{9}{2}$. Note that the components of $(\nabla e^{t\Delta}\mathbb{P}g)\widehat{\ }(\xi)$ are given by

$$\mathrm{i}\xi_h\left(\delta_{j,k} - \frac{\xi_j \xi_k}{|\xi|^2}\right) \exp(-(t+1)|\xi|^2) \qquad (j,h,k = 1,2,3, \quad \delta_{j,k} = 0 \text{ or } 1),$$

and the inverse Fourier transform can be easily computed (see, e.g., [17]). We immediately find that $\nabla e^{t\Delta}\mathbb{P}g$ is a smooth function in $\mathbb{R}^3$, such that

$$|\partial^\alpha \nabla e^{t\Delta}\mathbb{P}g(x)| \le C_\alpha (1+|x|)^{-(4+|\alpha|)} \qquad \text{for all } \alpha \in \mathbb{N}^3.$$

This bound implies that $e^{t\Delta}\mathbb{P}\nabla g \in Z_\delta$, for $0 \le \delta < \frac{9}{2}$. But $\int \nabla e^{t\Delta}\mathbb{P}g = 0$ (the Fourier transform of the integrand vanishes at the origin) and thus $\nabla e^{t\Delta}\mathbb{P}g$ belongs, more precisely, to X_δ.

Let us now prove that $\nabla e^{t\Delta}\mathbb{P}h$ does also belong to X_δ. The only difficulty is for $\frac{3}{2} < \delta < \frac{9}{2}$, $\delta \ne \frac{5}{2}, \frac{7}{2}$. Indeed, if $\frac{1}{2} < \delta < \frac{3}{2}$, then one observes that the weight $(1+|x|^2)^\delta$ belongs to the Muckenhoupt class A_2; see [18]. This implies that $\mathbb{P}$, and more generally the Riesz transforms, are bounded in $L^2(\mathbb{R}^3, (1+|x|^2)^\delta dx)$. Since $\nabla e^{t\Delta}\mathbb{P}$ and the Laplacian commute, the result easily follows applying this remark to Δh.

To deal with the case $\delta > \frac{3}{2}$, we start by recalling that the Sobolev space H^δ is defined by

$$||q||^2_{H^\delta} \equiv \int |\widehat{q}(\xi)|^2 (1+|\xi|^2)^\delta \, d\xi$$

and that $H^\delta \subset C^{\delta - 3/2}$ (the Hölder–Zygmund space). Thus, stating that h belongs to X_δ is equivalent to stating that

$$q(\xi) \equiv |\xi|^2 \widehat{h}(\xi) \in H^\delta \qquad \text{and} \qquad \partial^\alpha q(0) = 0 \quad \text{ for all } 0 \le |\alpha| \le \left[\delta - \frac{3}{2}\right].$$

These two conditions on q can be expressed by saying that q belongs to $L^2(\mathbb{R}^3) \cap \dot{H}^\delta_{rel}$, where $\dot{H}^\delta_{rel}$ is the realization of the homogeneous Sobolev space $\dot{H}^\delta$ (see Bourdaud [3]). Recall that $\dot{H}^\delta_{rel}$ can be injected into $\mathcal{S}'(\mathbb{R}^3)$ (this would not be true for $\dot{H}^\delta$, which is instead a space of tempered distributions modulo polynomials) and hence the notion of pointwise multipliers makes sense in the realized space. It follows from the result of [3] that $m(\xi) \equiv \xi_j/|\xi|$ is a multiplier for $\dot{H}^\delta_{rel}$ (any homogeneous function of degree 0 which is smooth outside the origin is indeed a multiplier for this space).

Since $h \in X_\delta$, the components of $|\xi|^2 \widehat{\mathbb{P}h}(\xi)$, which are given by $(\delta_{j,k} - \xi_j \xi_k |\xi|^{-2}) q(\xi)$, belong to $L^2(\mathbb{R}^3) \cap \dot{H}^\delta_{rel}$. Hence, $\mathbb{P}h \in X_\delta$. Moreover, $\mathrm{i}\xi_h e^{-t|\xi|^2} \in \mathcal{S}(\mathbb{R}^3)$ is also a multiplier of $\dot{H}^\delta_{rel}$ (with norm $c/\sqrt{t}$). Then we get $||\nabla e^{t\Delta}\mathbb{P}f||_{X_\delta} \le Ct^{-1/2}||f||_{Z_\delta}$ and Lemma 4.1 is thus proven. □

Our last lemma deals with the case $\delta = \frac{1}{2}$.

Lemma 4.2. *The operator $\nabla e^{t\Delta}\mathbb{P}$ is bounded in $Z_{1/2}$ for all $t > 0$, with operator norm $O(t^{-1/2})$ as $t \to 0$.*

Proof. Following Miyakawa's notations, we denote by $F(x,t)$ the kernel of the operator $\nabla e^{t\Delta}\mathbb{P}$. Then we know that $F(x,t) = t^{-2}\Phi(x/\sqrt{t})$, where Φ is smooth in $\mathbb{R}^3$ and $|\Phi(x)| \le C(1+|x|)^{-4}$ (see again [17] for more details). Let us show that $||\Phi * f||_{Z_{1/2}} \le C||f||_{Z_{1/2}}$. Then the conclusion will follow from a simple rescaling argument.

Let $f \in Z_{1/2}$ and write $f = g + \sum_{j=0}^\infty f_j$, where g and f_j satisfy (2.2). We also know that $|f_j(x)| \le \epsilon_j$ for all $x \in \operatorname{supp} f_j$ (we saw this right after (2.3)), with $\epsilon_j \in \ell^2(\mathbb{N})$. We obviously have $\Phi * g \in Z_{1/2}$ and $\Phi * f \in L^\infty(\mathbb{R}^3)$. Now let $2^k \le |x| \le 2^{k+1}$ ($k \in \mathbb{N}$, $k \ge 3$). Using the decay of Φ we see that $\Phi * f_j$ is bounded by $C\epsilon_j 2^{-4k}2^{3j}$ ($j \le k-3$) and $C\epsilon_j 2^{-j}$ ($j \ge k+3$). Thus $|\Phi * f(x)| \le \tilde{\epsilon}_k$, with $\tilde{\epsilon}_k \in \ell^2(\mathbb{N})$, and we conclude that $\int |\Phi * f(x)|^2 (1+|x|^2)^{-3/2} \, dx$ is finite.

Moreover, by Young's inequality,

$$\begin{aligned}
&\int |\nabla\Phi * f(x)|^2(1+|x|^2)^{-1/2}dx \\
&\qquad \le C\biggl(\int |\Phi(x)|(1+|x|)^{1/2}dx\biggr)^2 \int |\nabla f(x)|^2(1+|x|^2)^{-1/2}dx, \\
&\int |\Delta\Phi * f(x)|^2(1+|x|^2)^{1/2}dx \\
&\qquad \le C\biggl(\int |\Phi(x)|(1+|x|)^{1/2}dx\biggr)^2 \int |\Delta f(x)|^2(1+|x|^2)^{1/2}dx
\end{aligned}$$

(we also used $(1+|x|)^{-1/2} \le c(1+|y|)^{-1/2}(1+|x-y|)^{1/2}$ in the first inequality and $(1+|x|)^{1/2} \le c(1+|y|)^{1/2}(1+|x-y|)^{1/2}$ in the second inequality). Lemma 4.2 now follows. ∎

Proof of Theorem 1.1. The proof is based on the application of Kato's standard iteration argument in the space $C([0,T],Z_\delta)$. Let us write (IE) in the compact form $u(t) = e^{t\Delta}a - B(u,u)$, where $B(u,v) = \int_0^t \nabla \cdot e^{(t-s)\Delta}\mathbb{P}(u \otimes v)(s)\,ds$.

By Lemmas 4.1 and 4.2, and using the fact that Z_δ is a pointwise algebra for $\delta \ge \frac{1}{2}$, we see that the bilinear operator B is bounded in $C([0,T],Z_\delta)$ for $\frac{1}{2} \le \delta < \frac{9}{2}$, $\delta \ne \frac{3}{2}, \frac{5}{2}, \frac{7}{2}$, (the continuity with respect to the time variable is straightforward). Moreover, $\|B(u,v)\|_\delta \le C_T \|u\|_\delta\, \|v\|_\delta$, where $\|w\|_\delta \equiv \sup_{t\in[0,T]} ||w(t)||_{Z_\delta}$ and $C_T = O(T^{1/2})$ as $t \to 0$. We get the same conclusion for $\delta = \frac{3}{2}, \frac{5}{2}, \frac{7}{2}$ if we use the continuous embedding $Z_\delta \subset Z_{\delta'}$ $(\delta \ge \delta')$ and the stronger version of Lemma 2.3. For example,

$$\|B(u,v)\|_{3/2} \le C\|B(u,v)\|_2 \le CT^{1/2}\|u\|_{5/4}\|v\|_{5/4} \le CT^{1/2}\|u\|_{3/2}\|v\|_{3/2},$$

and a similar argument can be used for $\delta = \frac{5}{2}, \frac{7}{2}$.

Since $e^{t\Delta}a$ belongs to $C([0,T],Z_\delta)$ if $a \in Z_\delta$ for all $\delta \ge 0$, as is easily checked, we see that the fixed point argument applies in $C([0,T],Z_\delta)$, at least if $T > 0$ is small enough. ∎

We state as a corollary a slight improvement of Furioli and Terraneo's theorem.

Corollary 4.3. *Let $\frac{1}{2} < \delta < \frac{9}{2}$ and $a \in X_\delta$. Then there exists $T > 0$ such that* (IE) *can be uniquely solved in $C([0,T],X_\delta)$.*

Proof. The result is true for $\frac{1}{2} < \delta < \frac{7}{2}$ and $\delta \ne \frac{3}{2}, \frac{5}{2}$ because of the identification between X_δ and Z_δ. For $\frac{7}{2} < \delta < \frac{9}{2}$ we use the fact that every divergence-free vector field which is in $L^1(\mathbb{R}^3)$ must have a vanishing integral. Hence the existence and unicity result for (IE) in X_δ again follows from the corresponding result in Z_δ and the last conclusion of Proposition 3.1.

In the case $\delta = \frac{3}{2}$ we can observe that if $u, v \in C([0,T],X_{3/2})$, then $B(u,v)$ belongs, e.g., to $C([0,T],Z_2)$, which is contained in $C([0,T],X_{3/2})$. A similar argument can be used for $\delta = \frac{5}{2}, \frac{7}{2}$ and the conclusion easily follows. ∎

As claimed in the introduction, the restriction $\delta < \frac{9}{2}$ cannot be removed in Theorem 1.1. Indeed, if u is a solution to (IE) such that $u \in C([0,T],Z_{9/2})$, for some $T > 0$, then the initial datum must satisfy the conditions of Dobrokhotov and Shafarevich: $\int (a_j a_k) = 0$ if $j \ne k$ and $\int a_1^2 = \int a_2^2 = \int a_3^2$. This is due to the fact that the localization condition $a \in L^2(\mathbb{R}^3, (1+|x|^2)^{5/2}dx)$ is not conserved during the evolution (see [6]). For the same reason, the condition $a \in X_\delta$, $\delta > \frac{9}{2}$ breaks down (in general). But we do not know if the condition $a \in X_{9/2}$ (or $a \in X_{1/2}$) propagates.

We conclude by observing that spatially localized flows $u(t)$, belonging to Z_δ and X_δ with $\delta > \frac{9}{2}$, do, however, exist. Examples of such flows can be found in [4].

Acknowledgment. The author would like to thank the three referees for their remarks and useful comments, which helped him to improve the first version of this paper in a number of ways.

REFERENCES

[1] C. Amrouche, V. Girault, M. E. Schonbek, and T. P. Schonbek, *Pointwise decay of solutions and of higher derivatives to Navier–Stokes equations*, SIAM J. Math. Anal., 31 (2000), pp. 740–753.

[2] H.-O. Bae and B. J. Jin, *Temporal and Spatial Decays for the Navier–Stokes Equation*, preprint, Ajou University, Suwon, Republic of Korea.

[3] G. Bourdaud, *Localisation et multiplicateurs des espaces de Sobolev homogènes*, Manuscripta Math., 60 (1988), pp. 93–130.

[4] L. Brandolese, *Space-time decay of Navier–Stokes flows invariant under rotations*, Math. Ann., 329 (2004), pp. 685–706.

[5] L. Brandolese, *Weighted L^2-spaces and strong solutions of the Navier–Stokes equations in $\mathbb{R}^3$*, in Trends in Partial Differential Equations of Mathematical Physics, Progr. Nonlinear Differential Equations Appl. 61, Birkhäuser, Basel, 2005, pp. 27–35.

[6] L. Brandolese and Y. Meyer, *On the instantaneous spreading for the Navier–Stokes system in the whole space*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 273–285.

[7] R. Coifman, P.-L. Lions, Y. Meyer, and S. Semmes, *Compensated compactness and Hardy spaces*, J. Math. Pures Appl., 72 (1993), pp. 247–286.

[8] R. Coifman and G. Weiss, *Extension of the Hardy spaces and their use in analysis*, Bull. Amer. Math. Soc., 83 (1977), pp. 569–645.

[9] R. Farwig and H. Sohr, *Weighted energy inequalities for the Navier–Stokes equations in exterior domains*, Appl. Anal., 58 (1995), pp. 157–173.

[10] R. Farwig and H. Sohr, *Global estimates in weighted spaces of weak solutions of the Navier–Stokes equations in exterior domains*, Arch. Math. (Basel), 67 (1996), pp. 319–330.

[11] G. Furioli, P.-G. Lemarié-Rieusset, and E. Terraneo, *Unicité dans $L^3(\mathbb{R}^3)$ et d'autres espaces fonctionnels limites*, Rev. Mat. Iberoamericana, 16 (2000), pp. 269–315.

[12] G. Furioli and E. Terraneo, *Molecules of the Hardy space and the Navier–Stokes equations*, Funkcial. Ekvac., 45 (2002), pp. 141–160.

[13] C. He, *Weighted estimates for nonstationary Navier–Stokes equations*, J. Differential Equations, 148 (1998), pp. 422–444.

[14] C. He and Z. Xin, *On the decay properties of solutions to the nonstationary Navier–Stokes equations in $\mathbb{R}^3$*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 597–619.

[15] P. G. Lemarié-Rieusset, *Recent Developments in the Navier–Stokes Problem*, Chapman & Hall, CRC Press, Boca Raton, FL, 2002.

[16] T. Miyakawa, *Application of Hardy space techniques to the time-decay problem for incompressible Navier–Stokes flows in $\mathbb{R}^n$*, Funkcial. Ekvac., 41 (1998), pp. 383–434.

[17] T. Miyakawa, *On space time decay properties of nonstationary incompressible Navier–Stokes flows in $\mathbb{R}^n$*, Funkcial. Ekvac., 32 (2000), pp. 541–557.

[18] E. M. Stein, *Harmonic Analysis. Real-Variable Methods, Orthogonality and Oscillatory Integrals*, Princeton University Press, Princeton, NJ, 1993.

[19] F. Vigneron, *Spatial decay of the velocity field of an incompressible viscous fluid in $\mathbb{R}^d$*, Nonlinear Anal., to appear.

[20] A. Youssfi, *Localisation des espaces de Besov homogènes*, Indiana Univ. Math. J., 37 (1988), pp. 565–587.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 685–711

WILSON BASES FOR GENERAL TIME-FREQUENCY LATTICES*

GITTA KUTYNIOK† AND THOMAS STROHMER‡

Abstract. Motivated by a recent generalization of the Balian–Low theorem and by new research in wireless communications, we analyze the construction of Wilson bases for general time-frequency lattices. We show that orthonormal Wilson bases for $L^2(\mathbb{R})$ can be constructed for any time-frequency lattice whose volume is $\frac{1}{2}$. We then focus on the spaces $\ell^2(\mathbb{Z})$ and $\mathbb{C}^L$ which are the preferred settings for numerical and practical purposes. We demonstrate that with a properly adapted definition of Wilson bases the construction of orthonormal Wilson bases for general time-frequency lattices also holds true in these discrete settings. In our analysis we make use of certain metaplectic transforms. Finally, we discuss some practical consequences of our theoretical findings.

Key words. Wilson basis, metaplectic transform, Gabor frame, Schrödinger representation, time-frequency lattice

AMS subject classifications. 94A12, 42C15

DOI. 10.1137/S003614100343723X

1. Introduction. Gabor systems have become a popular tool, both in theory and in applications; e.g., see [12, 18, 13]. However, one drawback is that due to the *Balian–Low theorem* it is impossible to construct (orthogonal) Gabor bases for $L^2(\mathbb{R})$ with good time-frequency localization [18]. In [33, 9] it has been shown that a modification of Gabor bases, so-called *Wilson bases*, provide a means to circumvent the Balian–Low theorem. Indeed, there exist orthogonal Wilson bases for $L^2(\mathbb{R})$ whose basis functions have exponential decay in time and frequency. These Wilson bases can be constructed from certain tight Gabor frames with redundancy 2.

Gabor frames are usually associated with rectangular time-frequency lattices, but they can also be defined for general nonseparable lattices; see, e.g., [10, 17, 11, 28]. Recently it has been shown that such a generalization of Gabor frames to general time-frequency lattices does not enable us to overcome the Balian–Low theorem [19, 2]. This leads naturally to the question of whether it is possible to extend the construction of Wilson bases to general time-frequency lattices.

Another motivation for the research presented in this paper has its origin in wireless communication. Orthogonal frequency division multiplexing (OFDM) is a wireless transmission technology employing a set of transmission functions which is usually associated with a rectangular time-frequency lattice [14]. The connection to Gabor theory is given by the fact that the collection of transmission pulses in OFDM can be interpreted as a Gabor system; see [27, 30]. The density of the associated rectangular time-frequency lattice can be seen as a measure of the spectral efficiency in terms of number of bits transmitted per Hertz per second. The necessary condition of linear independence of the transmission functions implies that we are dealing with either an undersampled or a critically sampled Gabor system.

*Received by the editors November 4, 2003; accepted for publication (in revised form) June 4, 2004; published electronically November 9, 2005.
http://www.siam.org/journals/sima/37-3/43723.html
†Mathematical Institute, Justus-Liebig-University Giessen, 35392 Giessen, Germany (gitta.kutyniok@math.uni-giessen.de). This author's research was supported by Forschungspreis 2003 der Universität Paderborn.
‡Department of Mathematics, University of California, Davis, CA 95616-8633 (strohmer@math.ucdavis.edu). This author's research was supported by NSF DMS grant 0208568.

For wireless channels that are time-dispersive (due to multipath) and frequency-dispersive (due to the Doppler effect), good time-frequency localization of the transmission pulses is essential to mitigate the interferences caused by the dispersion of the channel [27, 30]. The ideal set of transmission pulses should therefore possess (i) good time-frequency localization and (ii) maximize the spectral efficiency; i.e., the transmission functions should correspond to a (critically sampled) Gabor basis for $L^2(\mathbb{R})$. As we know, the Balian–Low theorem prohibits these conditions to be fulfilled simultaneously.

Recently it has been shown that in case of time-frequency dispersive channels the performance of OFDM can be improved when using general time-frequency lattices, in particular hexagonal-type lattices [31]. In a nutshell, lattices that are adapted to the shape of the Wigner distribution of the transmission pulses allow for a better "packing" of the time-frequency plane, which in turn can be used to either achieve higher data rates or improve interference robustness of the associated so-called Lattice-OFDM system.

One variation of OFDM (for rectangular lattices) is called offset quadrature amplitude modulation (OQAM) OFDM. It corresponds to using a Wilson basis as a set of transmission functions [4]. OQAM-OFDM achieves maximal spectral efficiency and allows for transmission functions with good time-frequency localization. As in the case of standard OFDM, it would be potentially useful for time-frequency dispersive channels to extend OQAM-OFDM to general time-frequency lattices in order to improve the robustness of OQAM-OFDM against interference even further. Thus we again arrive at the problem of constructing Wilson bases for general nonseparable time-frequency lattices.

Yet another motivation comes from filter bank theory, more precisely cosine-modulated filter banks [6]. We know that discrete-time Wilson bases correspond to a special class of cosine-modulated filter banks (see [6]). In light of the improvements gained by using general time-frequency lattices in OFDM [31], it would be interesting to analyze if the construction of cosine-modulated filter banks can be extended to general time-frequency lattices. A positive answer to this question might lead to a more efficient encoding of signals and images.

Since our goal to construct Wilson bases for general time-frequency lattices is in part motivated by applied problems and since any numerical implementation of Wilson bases is based on a discrete model, our analysis will not only concern $L^2(\mathbb{R})$ but also comprise the spaces $\ell^2(\mathbb{Z})$ and $\mathbb{C}^L$. Furthermore, $\ell^2(\mathbb{Z})$ is the appropriate setting when Wilson bases are utilized as filter banks, since in this case one deals with sampled, thus discrete-time, signals.

1.1. Notation. We assume that the reader is familiar with the theory of Gabor frames and refer to [18] for background and details.

A lattice Λ in $\mathbb{R}^d$ is a discrete subgroup with compact quotient; i.e., there exists a matrix $A \in GL(d, \mathbb{R})$ such that $\Lambda = A\mathbb{Z}^d$. The matrix A is called the (nonunique) generator matrix for Λ. The volume of Λ is $\mathrm{vol}(\Lambda) = |\det(A)|$. Two lattices, which play a crucial role in OFDM design (see [31]), are the rectangular lattice Λ_R and the hexagonal lattice Λ_H. A generator matrix for Λ_R is given by

$$A_R = \begin{bmatrix} T & 0 \\ 0 & F \end{bmatrix},$$

and a generator matrix for Λ_H is given by

$$A_H = \begin{bmatrix} \frac{\sqrt{2}}{\sqrt[4]{3}}T & \frac{\sqrt{2}}{2\sqrt[4]{3}}T \\ 0 & \frac{\sqrt[4]{3}}{\sqrt{2}}F \end{bmatrix},$$

where $T, F > 0$. An easy calculation shows that both lattices Λ_R and Λ_H have the same volume TF. A normal form for matrices, which we will use in the following, is the so-called Hermite normal form [20]. We say that a matrix $A = [\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}]$ is in *Hermite normal form* if $c = 0$, $a, d > 0$, and $0 \le b < a$. For example, both matrices A_R and A_H are in Hermite normal form.

For $(x, y) \in \mathbb{R}^2$ and $g \in L^2(\mathbb{R})$, let $g_{x,y}$ be defined by

$$g_{x,y}(t) = e^{2\pi i y t} g(t - x).$$

We denote by $\mathcal{G}(g, \Lambda)$ the system of functions given by

$$g_{\lambda,\mu}(t) = g(t - \lambda) e^{2\pi i t \mu}, \qquad (\lambda, \mu) \in \Lambda.$$

As usual, the *redundancy* of $\mathcal{G}(g, \Lambda)$ is given by $\frac{1}{\mathrm{vol}(\Lambda)}$.

As in [18], we define the *Schrödinger representation* $\rho : \mathbb{H} \to \mathcal{U}(L^2(\mathbb{R}))$ by

$$\rho(x, y, z) g(t) = e^{2\pi i z} e^{-\pi i x y} e^{2\pi i y t} g(t - x).$$

Note that

$$g_{x,y} = e^{\pi i x y} \rho(x, y, 1) g. \tag{1.1}$$

Furthermore, we use the following notation from [18] (with slight changes):

$$\mathcal{J} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad \mathcal{B}_b = \begin{bmatrix} b & 0 \\ 0 & \frac{1}{b} \end{bmatrix}, \quad \mathcal{C}_c = \begin{bmatrix} 1 & 0 \\ c & 1 \end{bmatrix}.$$

$\mathcal{F}$ and $^\wedge$ denote the Fourier transform, $\mathcal{F}^{-1}$ and $^\vee$ denote the inverse Fourier transform. The dilation is given by $\mathcal{D}_b f(t) = |b|^{\frac{1}{2}} f(bt)$ and the "chirp" operator is defined via $\mathcal{N}_c f(t) = e^{-\pi i c t^2} f(t)$, where $b, c \in \mathbb{R}$ and $f \in L^2(\mathbb{R})$.

Metaplectic transforms will turn out to be a very useful tool in our analysis. For the study of the discrete and finite case, we need a result on metaplectic transforms from [24]. Since this thesis is not easy to access, we present the result here in the slightly weaker version we will use, together with the necessary definitions and notations. The result will be stated in the situation of a general locally compact abelian group G with dual group $\widehat{G}$, group multiplication denoted by $+$, and action of $\widehat{G}$ on G denoted by $\langle x, \chi \rangle$ for $x \in G$ and $\chi \in \widehat{G}$. The cases we are interested in later on are $G = \mathbb{Z}$ and $G = \mathbb{Z}_L$. We will usually write a metaplectic transform σ in the matrix notation $\sigma = [\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}] \in \mathrm{Hom}(G \times \widehat{G})$, which means that $\alpha \in \mathrm{Hom}(G)$, $\beta \in \mathrm{Hom}(\widehat{G}, G)$, $\gamma \in \mathrm{Hom}(G, \widehat{G})$, and $\delta \in \mathrm{Hom}(\widehat{G})$. Then the *adjoint* σ^* is defined by $\langle (x, \chi), \sigma^*(y, \pi) \rangle = \langle \sigma(x, \chi), (y, \pi) \rangle$ for all $(x, \chi), (y, \pi) \in G \times \widehat{G}$. Let η be defined by $\eta = [\begin{smallmatrix} 0 & -I_{\widehat{G}} \\ I_G & 0 \end{smallmatrix}] \in \mathrm{Hom}(G \times \widehat{G}, \widehat{G} \times G)$, where the above definition concerning the matrix notation has to be adapted in an obvious way, and where I_G and $I_{\widehat{G}}$ denote the identity on G and $\widehat{G}$, respectively. Then σ is called *symplectic* if $\sigma^* \eta \sigma = \eta$. If $\zeta \in \mathrm{Hom}(G \times \widehat{G}, \widehat{G} \times G)$, then ψ is a *second degree character* of $G \times \widehat{G}$ associated to

ζ if $\psi((x,\chi)+(y,\pi)) = \psi(x,\chi)\psi(y,\pi)\,\langle (x,\chi),\zeta(y,\pi)\rangle$ for all $(x,\chi),(y,\pi) \in G \times \widehat{G}$. Moreover, for $(x,\chi) \in G \times \widehat{G}$ and $g \in L^2(G)$, let $g_{x,\chi}$ be defined by

$$g_{x,\chi}(t) = \chi(t)g(t-x),$$

which generalizes the previous definition.

We can now state the version of [24, Theorem 1.1.28] which we will employ in sections 3 and 4.

THEOREM 1.1. *Let* $\sigma := [\begin{smallmatrix} \alpha & \beta \\ \gamma & \delta \end{smallmatrix}] \in \mathrm{Hom}(G \times \widehat{G})$ *be symplectic and let* ψ_ζ *be a second degree character of* $G \times \widehat{G}$ *associated to*

$$\zeta := \sigma^* \begin{bmatrix} 0 & 0 \\ I_G & 0 \end{bmatrix} \sigma - \begin{bmatrix} 0 & 0 \\ I_G & 0 \end{bmatrix} \in \mathrm{Hom}(G \times \widehat{G}, \widehat{G} \times G).$$

If U *is defined by*

$$Uf(t) := \int_{\widehat{G}} f(\alpha t + \beta\omega)\psi_\zeta^{-1}(t,\omega)d\omega,$$

then we have

$$(Uf)_{x,\chi}(t) = \psi_\zeta^{-1}(x,\chi)Uf_{\sigma(x,\chi)}(t), \quad (x,\chi) \in G \times \widehat{G}.$$

2. Wilson bases for general lattices—the continuous case. We first show that all lattices in $\mathbb{R}^2$ which are important for applications, such as the rectangular lattice, the hexagonal lattice, and lattices whose generator matrix has rational entries, possess a uniquely determined matrix in Hermite normal form, which we will call the *canonical generator matrix*. In particular, we characterize exactly those lattices which possess a generator matrix in Hermite normal form.

We then show that it is possible to construct orthonormal Wilson bases for time-frequency lattices Λ with $\mathrm{vol}(\Lambda) = \frac{1}{2}$ which possess a generator matrix in Hermite normal form. In principle Wilson systems can be defined for lattices Λ with $\mathrm{vol}(\Lambda) \neq \frac{1}{2}$; however, so far all known constructions of *orthogonal Wilson bases* for $L^2(\mathbb{R})$ are strictly tied to lattices with volume $\frac{1}{2}$. In light of this fact, throughout this paper a Wilson system will always be associated with a time-frequency lattice of volume $\frac{1}{2}$.

LEMMA 2.1. *Let* Λ *be a lattice in* $\mathbb{R}^2$. *Then the following conditions are equivalent.*

(i) $P_2(\Lambda)$ *is discrete, where* $P_2 : \mathbb{R}^2 \to \mathbb{R}$, $(x,y) \mapsto y$.

(ii) *There exists a generator matrix* A *for* Λ *which is in Hermite normal form.*

If one of these conditions is satisfied, the matrix A *is uniquely determined.*

Proof. Let

$$A' = \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix}$$

be an arbitrary generator matrix for Λ.

First we prove that (ii) implies (i). By (ii), there exists a matrix

$$A = \begin{bmatrix} a & b \\ 0 & d \end{bmatrix},$$

which is in Hermite normal form and which satisfies $A\mathbb{Z}^2 = \Lambda$. Thus $P_2(\Lambda) = d\mathbb{Z}$, which yields (i).

Next we show (i) $\Rightarrow$ (ii). For this, we construct a matrix

$$A = \begin{bmatrix} a & b \\ 0 & d \end{bmatrix},$$

which satisfies the claimed properties. Without loss of generality we assume that $d' \neq 0$ (if $d' = 0$ and $c' \neq 0$ we could change the columns of A'). We begin with the following observation. Assume that $\frac{c'}{d'}$ is not rational. Since $P_2(\Lambda)$ is a nontrivial discrete, additive subgroup of $\mathbb{R}$, hence a lattice, there exists $s \in \mathbb{R}\backslash\{0\}$ such that $P_2(\Lambda) = s\mathbb{Z}$. Thus $c' = sm$ and $d' = sn$ for some $m, n \in \mathbb{Z}$, a contradiction. This implies that the quotient $\frac{c'}{d'}$ is rational. Setting $\frac{c'}{d'} = \frac{k}{l}$, $k, l \in \mathbb{Z}$ with $\gcd(k, l) = 1$, and factoring out $\frac{d'}{l}$, without loss of generality we can assume that A' is of the form

$$A' = r \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix},$$

with $a', b', r \in \mathbb{R}$ and $c', d' \in \mathbb{Z}$. Now we proceed as follows. First we set $p := \gcd(c', d') > 0$. We then obtain

$$A' \begin{bmatrix} \pm\frac{d'}{p} \\ \mp\frac{c'}{p} \end{bmatrix} = \begin{bmatrix} \pm\frac{r\det(A')}{p} \\ 0 \end{bmatrix} \in \Lambda,$$

since $\frac{c'}{p}, \frac{d'}{p} \in \mathbb{Z}$. Hence we can define a by $a := \left|\frac{r\det(A')}{p}\right| > 0$.

In a second step we compute b and d. Since $\frac{c'}{p}$ and $\frac{d'}{p}$ are relative prime, there exist $m, n \in \mathbb{Z}$ such that $\frac{c'}{p}m + \frac{d'}{p}n = 1$ (see [22, Theorem 4.4]). Hence $c'm + d'n = p$ and we obtain

$$A' \begin{bmatrix} m \\ n \end{bmatrix} = \begin{bmatrix} r(a'm + b'n) \\ rp \end{bmatrix} \in \Lambda.$$

Now let $k \in \mathbb{Z}$ be chosen in such a way that $0 \leq r(a'm + b'n) + ka < a$ and define $b := ka + r(a'm + b'n)$ and $d := rp$. Without loss of generality we can assume that $d > 0$, since otherwise we just take $-m$ and $-n$ instead of m and n. Then

$$\begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} r(a'm + b'n) \\ rp \end{bmatrix} + k \begin{bmatrix} a \\ 0 \end{bmatrix} \in \Lambda$$

and $|ad| = |\frac{r\det(A')}{p}|rp = |r^2\det(A')|$. This proves that A generates Λ and is in Hermite normal form.

At last we prove that the matrix in condition (ii) is uniquely determined. For this, assume there exist $a, b, d, a', b', d' \in \mathbb{R}$ with $a, a', d, d' > 0$, $0 \leq b < a$, and $0 \leq b' < a'$ such that

$$\Lambda = A\mathbb{Z}^2 = A'\mathbb{Z}^2, \tag{2.1}$$

where

$$A = \begin{bmatrix} a & b \\ 0 & d \end{bmatrix} \quad \text{and} \quad A' = \begin{bmatrix} a' & b' \\ 0 & d' \end{bmatrix}.$$

By (2.1), there exist $m, n \in \mathbb{Z}$ with $a' = ma + nb$ and $0 = nd$, which implies that $a' = ma$. Again by (2.1), we can find $k, l \in \mathbb{Z}$ such that $b' = ka + lb$ and $d' = ld$. Since

$\mathrm{vol}(\Lambda) = |ad| = |a'd'|$, we obtain $|ml| = 1$. Now $a, a', d, d' > 0$ implies that $a = a'$, $d = d'$, and $l = 1$. Finally, applying this to $b' = ka + lb$ and using that $0 \leq b < a$ and $0 \leq b' < a'$ yields $b = b'$. Thus, we have shown that $A = A'$, which completes the proof. □

In the following, we restrict our attention to lattices which possess a generator matrix in Hermite normal form. All results in the situation $L^2(\mathbb{R})$ could be derived (in the same manner, but with much more technical effort) for general lattices, but with little or no practical benefit.

Definition 2.2. *Let Λ be a lattice in $\mathbb{R}^2$ which satisfies the conditions of Lemma* 2.1. *Then the uniquely determined generator matrix A of Lemma* 2.1 *is called the* canonical generator matrix *for Λ.*

Now let Λ be a lattice with $\mathrm{vol}(\Lambda) = \frac{1}{2}$ which possesses a generator matrix in Hermite normal form. Using the definition of a canonical generator matrix, we define a Wilson system associated with Λ as follows.

Definition 2.3. *If $\mathcal{G}(g, \Lambda) \subseteq L^2(\mathbb{R})$ is a Gabor system of redundancy* 2 *and*

$$A = \begin{bmatrix} a & b \\ 0 & d \end{bmatrix}$$

is the canonical generator matrix for the lattice Λ, then the associated Wilson system $\mathcal{W}(g, \Lambda, L^2(\mathbb{R})) = \{\psi^\Lambda_{m,n}\}_{m\in\mathbb{Z}, n\geq 0}$ *consists of the functions*

$$\begin{aligned} \psi^\Lambda_{m,0} &= g_{2ma,0} && \text{if } n = 0, \\ \psi^\Lambda_{m,n} &= \tfrac{1}{\sqrt{2}} e^{-\pi i b d n^2} (g_{ma+nb,nd} + g_{ma-nb,-nd}) && \text{if } n \neq 0,\ m+n \text{ even}, \\ \psi^\Lambda_{m,n} &= \tfrac{i}{\sqrt{2}} e^{-\pi i b d n^2} (g_{ma+nb,nd} - g_{ma-nb,-nd}) && \text{if } n \neq 0,\ m+n \text{ odd}. \end{aligned}$$

If the system $\mathcal{W}(g, \Lambda, L^2(\mathbb{R}))$ is an orthonormal basis for $L^2(\mathbb{R})$ we call it a Wilson (orthonormal) basis.

We will see that this definition reduces to the usual definition of Wilson systems. For this, we consider the rectangular lattice

$$\Gamma = \{(\tfrac{m}{2}, n)\}_{m,n\in\mathbb{Z}}.$$

It is an easy calculation to show that the canonical generator matrix for Γ is

$$A = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus, for each $g \in L^2(\mathbb{R})$, the Wilson system $\mathcal{W}(g, \Gamma, L^2(\mathbb{R}))$ consists indeed of the functions

$$\begin{aligned} \psi^\Gamma_{m,0} &= g_{m,0} && \text{if } n = 0, \\ \psi^\Gamma_{m,n} &= \tfrac{1}{\sqrt{2}} (g_{m/2,n} + g_{m/2,-n}) && \text{if } m+n \text{ is even}, \\ \psi^\Gamma_{m,n} &= \tfrac{i}{\sqrt{2}} (g_{m/2,n} - g_{m/2,-n}) && \text{if } m+n \text{ is odd}, \end{aligned}$$

which coincides with the usual definition of Wilson systems; cf., for instance, [18, Definition 8.5.1]. Notice that we will fix the notation Γ for the remainder.

We will make use of the following well-known theorem about a Wilson system for rectangular lattices to constitute an orthonormal basis (e.g., cf. [3, Theorem 4.1]).

Theorem 2.4. *Suppose that $g \in L^2(\mathbb{R})$ is such that*

(i) $\hat{g}$ *is real-valued and*

(ii) $\{g_{m/2,n}\}_{m,n\in\mathbb{Z}}$ *is a tight Gabor frame for* $L^2(\mathbb{R})$ *with frame bound* 2.

Then the system $\mathcal{W}(g,\Gamma,L^2(\mathbb{R}))$ *is a Wilson orthonormal basis for* $L^2(\mathbb{R})$.

We are now ready to extend the construction of Wilson bases for time-frequency lattices which possess a generator matrix in Hermite normal form.

THEOREM 2.5. *Let* Λ *be a lattice in* $\mathbb{R}^2$ *with* $\mathrm{vol}(\Lambda)=\frac{1}{2}$ *and canonical generator matrix*

$$A=\begin{bmatrix} a & b\\ 0 & d\end{bmatrix}.$$

Define U *by*

$$U:=\mathcal{D}_{1/d}\circ\mathcal{F}\circ\mathcal{N}_{-b/d}\circ\mathcal{F}^{-1}. \tag{2.2}$$

Let $g\in L^2(\mathbb{R})$ *be such that*

(i) $\widehat{Ug}$ *is real-valued and*

(ii) $\{g_{ma+nb,nd}\}_{m,n\in\mathbb{Z}}$ *is a tight frame for* $L^2(\mathbb{R})$ *with frame bound* 2.

Then the system $\mathcal{W}(g,\Lambda,L^2(\mathbb{R}))$ *is a Wilson orthonormal basis for* $L^2(\mathbb{R})$.

Proof. We will reduce our claim to Theorem 2.4 by using a metaplectic transform. We define $\mathcal{A}$ by

$$\mathcal{A}=\begin{bmatrix} d & -b\\ 0 & 2a\end{bmatrix}.$$

Then, for all $m,n\in\mathbb{Z}$, we have

$$\mathcal{A}(ma+nb,nd)=\left(\tfrac{1}{2}m,n\right). \tag{2.3}$$

Since $\mathcal{A}\in\mathrm{Sp}(2,\mathbb{R})$, we can apply [15, Theorem 4.51] and write

$$\mathcal{A}=B_d(-\mathcal{J})C_{b/d}\mathcal{J}.$$

By [18, Example 9.4.1] we obtain

$$\rho(x,y,1)g=U^{-1}\rho(\mathcal{A}(x,y),1)(Ug), \tag{2.4}$$

with U defined in (2.2). Using (1.1), (2.3), and (2.4), we compute

$$\begin{aligned} g_{ma+nb,nd}&=e^{\pi i(ma+nb)nd}\rho(ma+nb,nd,1)g\\ &=e^{\pi i(ma+nb)nd}U^{-1}\rho(\mathcal{A}(ma+nb,nd),1)(Ug)\\ &=e^{\pi i(ma+nb)nd}U^{-1}\rho(\tfrac{1}{2}m,n,1)(Ug). \end{aligned}$$

Using (1.1) again, we obtain

$$g_{ma+nb,nd}=e^{\pi i(ma+nb)nd}e^{-\pi i\frac{1}{2}mn}U^{-1}(Ug)_{\frac{m}{2},n}=e^{\pi ibdn^2}U^{-1}(Ug)_{\frac{m}{2},n}. \tag{2.5}$$

Since multiplication by a phase factor and applying a unitary operator to a tight frame preserves tightness (and frame bounds), it follows that condition (ii) is equivalent to $\{(Ug)_{m/2,n}\}_{m,n\in\mathbb{Z}}$ being a tight frame with frame bound 2. We need not deal with condition (i), since this states already that $\widehat{Ug}$ is real-valued. Applying Theorem 2.4 yields that the Wilson system $\mathcal{W}(Ug,\Gamma,L^2(\mathbb{R}))$ is an orthonormal basis.

Using now the metaplectic transform, i.e., (2.5), and the fact that U is a unitary operator, finishes the proof. □

We will conclude this section by providing an example which shows how we can compute a Wilson basis with excellent time-frequency localization for a special lattice, but this calculation can also be done for an arbitrary lattice.

Here we consider the hexagonal lattice Λ_H with generator matrix

$$A_H = \begin{bmatrix} \frac{\sqrt{2}}{\sqrt[4]{3}} & \frac{\sqrt{2}}{2\sqrt[4]{3}} \\ 0 & \frac{\sqrt[4]{3}}{\sqrt{2}} \end{bmatrix}.$$

This lattice was also used in [31]. Observe that since we have $0 \leq \frac{\sqrt{2}}{2\sqrt[4]{3}} < \frac{\sqrt{2}}{\sqrt[4]{3}}$, the matrix A_H is already the canonical generator matrix for Λ_H. First we define the function $h \in L^2(\mathbb{R})$ by

$$h(x) = (2\nu)^{\frac{1}{4}} e^{-\nu\pi x^2}.$$

By [18, Theorem 7.5.3], the set $\{h_{m/2,n}\}_{m,n\in\mathbb{Z}}$ is a frame for $L^2(\mathbb{R})$. Let S denote its frame operator and consider the function $\varphi \in L^2(\mathbb{R})$ given by

$$\varphi = \sqrt{2}\mathcal{F} \circ S^{-\frac{1}{2}} h. \tag{2.6}$$

Using [3, Theorem 4.6], this function coincides with the function considered in [9, section 4]. There it was shown that φ satisfies conditions (i) and (ii) of Theorem 2.4 and hence yields a Wilson basis in the sense of Theorem 2.4. Moreover, φ has exponential decay in time and frequency. To obtain a generating function for a Wilson basis with respect to Λ_H in the sense of Theorem 2.5, first observe that by the proof of Theorem 2.5, we only need to compute the function $g = U^{-1}\varphi$, where U is defined in (2.2). Then g automatically satisfies conditions (i) and (ii) of Theorem 2.5, and hence the system $\mathcal{W}(g, A_H, L^2(\mathbb{R}))$ is a Wilson orthonormal basis by Theorem 2.5. Thus we define $g \in L^2(\mathbb{R})$ by

$$g = \mathcal{F} \circ \mathcal{N}_{\frac{1}{\sqrt{3}}} \circ \mathcal{F}^{-1} \circ \mathcal{D}_{\frac{\sqrt[4]{3}}{\sqrt{2}}} \varphi = \sqrt{2}\mathcal{F} \circ \mathcal{N}_{\frac{1}{\sqrt{3}}} \circ \mathcal{D}_{\frac{\sqrt{2}}{\sqrt[4]{3}}} \circ S^{-\frac{1}{2}} h.$$

Let us mention that the function g has exponential decay in time and frequency. Thus we obtain a Wilson basis with respect to the lattice Λ with very good time-frequency localization. As already mentioned above, this procedure can be applied to an arbitrary lattice, hence we obtain a Wilson basis with excellent time-frequency localization for any lattice.

3. Wilson bases for general lattices—the discrete case. In this section we analyze the construction of Wilson bases for general time-frequency lattices for functions defined on $\ell^2(\mathbb{Z})$. The reasons for considering the setting $\ell^2(\mathbb{Z})$ are that, on the one hand, several applications such as filter bank design in digital signal processing deal directly with a discrete setting [6] and, on the other hand, even those problems that arise in the "continuous" setting of $L^2(\mathbb{R})$ require a discrete model for their numerical treatment. Thus, with these practical aspects in mind, throughout this section we naturally consider only lattices whose generator matrices have rational entries, since any implementation is intrinsically restricted to such "rationally" generated lattices. Another natural setting for numerical implementations is of course

$\mathbb{C}^L$ (which can be identified with the space of L-periodic sequences). We will analyze that case in the next section.

Before we proceed we define Gabor systems and Wilson systems on $\ell^2(\mathbb{Z})$ for general time-frequency lattices Λ with $\mathrm{vol}(\Lambda) = \frac{1}{2}$. First we prove that each lattice possesses a generator matrix of some particular form.

LEMMA 3.1. *Let Λ be a lattice in $\mathbb{Z} \times \mathbb{R}$ with generator matrix A given by*

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \text{with } a, b \in \mathbb{Z},\ c, d \in \mathbb{Q}, \text{ and } \det(A) = \frac{1}{2},$$

and denote $c = \frac{p}{q}, d = \frac{p'}{q'}$ with $\gcd(p,q) = \gcd(p',q') = 1$, $p, q, p', q' \in \mathbb{Z}$. Then Λ possesses a uniquely determined generator matrix of the form

$$A' = \begin{bmatrix} \frac{N}{2} & b' \\ 0 & \frac{1}{N} \end{bmatrix}, \tag{3.1}$$

where $N = \frac{qq'}{\gcd(pq',p'q)}$ and $b' \in \mathbb{Z}$, $0 \le b' < \frac{N}{2}$.

Proof. We assume that $c \neq 0$, otherwise (3.1) is automatically satisfied. We first show that A can be written as

$$A = \begin{bmatrix} a & b \\ \frac{r}{N} & \frac{s}{N} \end{bmatrix}, \tag{3.2}$$

with integers r, s, N, such that $\gcd(r,s) = 1$.

Let $c = \frac{p}{q}, d = \frac{p'}{q'}$ with $p, p', q, q' \in \mathbb{Z}$, denote $N' := qq', \tilde{c} := pq', \tilde{d} := p'q$, and write $z = \gcd(\tilde{c}, \tilde{d})$. Since $a, b, \tilde{c}, \tilde{d} \in \mathbb{Z}$ and since

$$\mathrm{vol}(\Lambda) = \frac{1}{2} \Rightarrow a\tilde{d} - b\tilde{c} = \frac{N'}{2},$$

it follows that $\frac{N'}{2} \in \mathbb{Z}$. A necessary and sufficient condition for the equation $a\tilde{d} - b\tilde{c} = \frac{N'}{2}$ to have an integer solution in a and b is that $\gcd(\tilde{c}, \tilde{d}) | \frac{N'}{2}$ (see [22, Theorem 8.1]), hence $z | \frac{N'}{2}$. Denote $z' := \frac{N'}{2z}, r := \frac{\tilde{c}}{z}, s := \frac{\tilde{d}}{z}$. Then $c = \frac{r}{2z'}, d = \frac{s}{2z'}$ with $z' \in \mathbb{Z}$, and $\gcd(r,s) = 1$. By a proper choice of the signs of c and d we can always assume that z' is positive. By writing $N := 2z' \in \mathbb{Z}$ we see that A can indeed be written as in (3.2).

Now, assuming that A is of the form (3.2), we compute

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} Nd \\ -Nc \end{bmatrix} = \begin{bmatrix} N(ad - bc) \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{N}{2} \\ 0 \end{bmatrix}.$$

Since $\gcd(r,s) = 1$ there exist integers m, n with $mr + ns = 1$. For such a pair (m,n) we denote $b' = am + bn$ and obtain

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} m \\ n \end{bmatrix} = \begin{bmatrix} am + bn \\ \frac{r}{N}m + \frac{s}{N}n \end{bmatrix} = \begin{bmatrix} b' \\ \frac{1}{N} \end{bmatrix}.$$

If $b' < 0$ or $b' \ge \frac{N}{2}$, we substitute the vector obtained by

$$\begin{bmatrix} b' \\ \frac{1}{N} \end{bmatrix} + k \begin{bmatrix} \frac{N}{2} \\ 0 \end{bmatrix} = \begin{bmatrix} b' + k\frac{N}{2} \\ \frac{1}{N} \end{bmatrix},$$

where $k \in \mathbb{Z}$ is chosen in such a way that $0 \leq b' + k\frac{N}{2} < \frac{N}{2}$. Consequently the matrix

$$\begin{bmatrix} \frac{N}{2} & b' \\ 0 & \frac{1}{N} \end{bmatrix} \tag{3.3}$$

generates the lattice Λ. Finally, since $\frac{N}{2}$ and b' are integers and $0 \leq b' < \frac{N}{2}$, the generator matrix in (3.3) is indeed of the form (3.1).

The fact that this is a unique representation follows immediately from the condition $0 \leq b' < \frac{N}{2}$. $\square$

Definition 3.2. *Let Λ be a lattice in $\mathbb{Z} \times \mathbb{R}$. Then the uniquely determined matrix A' of Lemma* 3.1 *is called the* canonical generator matrix *for Λ.*

In the following we will regard such a lattice as being in $\mathbb{Z} \times \mathbb{T}$ by considering $\Lambda = \{\frac{N}{2}m + bn, e^{2\pi i \frac{n}{N}}\}_{m\in\mathbb{Z},n=0,\dots,N-1}$. This is a very natural approach since, for all $k \in \mathbb{Z}$, we have

$$(\tfrac{N}{2}m + bn, \tfrac{1}{N}n + k) = (\tfrac{N}{2}(m - 2bk) + b(n + kN), \tfrac{1}{N}(n + kN)).$$

Hence the lattice $A'\mathbb{Z}^2$ is invariant under adding integers to the second component. Moreover, it is sufficient to restrict to the index set $\mathbb{Z} \times \{0, \dots, N-1\}$ since, for all $0 \leq n' < N$ and $k \in \mathbb{Z}$,

$$(\tfrac{N}{2}m + b(n' + kN), (n' + kN) \mod N) = (\tfrac{N}{2}(m + 2bk) + bn', n'),$$

which implies

$$\{(\tfrac{N}{2}m + bn, n \mod N)\}_{m,n\in\mathbb{Z}} = \{(\tfrac{N}{2}m + bn, n)\}_{m\in\mathbb{Z},n=0,\dots,N-1}$$

in the sense of sets.

Using the definition of canonical generator matrices we can now define Gabor systems for $\ell^2(\mathbb{Z})$.

Definition 3.3. *Let Λ be a lattice in $\mathbb{Z} \times \mathbb{T}$ with canonical generator matrix A given by*

$$A = \begin{bmatrix} \frac{N}{2} & b \\ 0 & \frac{1}{N} \end{bmatrix},$$

and let $g \in \ell_2(\mathbb{Z})$. Then the associated Gabor system $\{g_{m\frac{N}{2}+nb,n\frac{1}{N}}\}_{m\in\mathbb{Z},n=0,\dots,N-1}$ is given by

$$g_{m\frac{N}{2}+nb,n\frac{1}{N}}(l) = g(l - (m\tfrac{N}{2} + nb))e^{\frac{2\pi i l n}{N}}, \qquad l \in \mathbb{Z}.$$

We first give the definition of a Wilson basis associated with a lattice with diagonal canonical generator matrix, i.e., with $b = 0$ (in this special case the definition coincides with the one given in [5]).

Definition 3.4. *Let Λ be a lattice in $\mathbb{Z} \times \mathbb{T}$ with canonical generator matrix A given by*

$$A = \begin{bmatrix} \frac{N}{2} & 0 \\ 0 & \frac{1}{N} \end{bmatrix},$$

and let $g \in \ell_2(\mathbb{Z})$. Then the Wilson system $\mathcal{W}(g, \Lambda, \ell_2(\mathbb{Z})) = \{\psi_{m,n}\}_{m\in\mathbb{Z},n=0,\dots,\frac{N}{2}}$ is given by

$$\psi^{\Lambda}_{m,n} = g_{mN,n\frac{1}{N}} \qquad \text{if } m \in \mathbb{Z},\ n = 0, \tfrac{N}{2},$$

and for $m \in \mathbb{Z}$, $n = 1, \dots, \frac{N}{2} - 1$,

$$\begin{aligned}\psi^{\Lambda}_{m,n} &= \tfrac{1}{\sqrt{2}}(g_{m\frac{N}{2},n\frac{1}{N}} + g_{m\frac{N}{2},-n\frac{1}{N}}) \quad \textit{if } m+n \textit{ is even},\\ \psi^{\Lambda}_{m,n} &= \tfrac{i}{\sqrt{2}}(g_{m\frac{N}{2},n\frac{1}{N}} - g_{m\frac{N}{2},-n\frac{1}{N}}) \quad \textit{if } m+n \textit{ is odd}.\end{aligned}$$

The Zak transform, which can be defined for any locally compact abelian group (cf. [26]), will be employed to prove equivalent conditions for the Wilson system to form an orthonormal basis. In particular, we need the Zak transform on $\mathbb{T}$ with respect to the uniform lattice $K = \{e^{2\pi i \frac{2k}{N}} : k = 0, \dots, \frac{N}{2} - 1\}$ in $\mathbb{T}$, which is defined on the set of square-integrable functions on $\{e^{2\pi i t} : t \in [0, \frac{2}{N})\} \times \{0, \dots, \frac{N}{2} - 1\}$ by

$$Zf(e^{2\pi i t}, y) = \sum_{k=0}^{\frac{N}{2}-1} f(e^{2\pi i (t + \frac{2k}{N})}) e^{2\pi i \frac{2k}{N} y}.$$

The proof of the following proposition is inspired by the proof of [9, Proposition 5.2].

PROPOSITION 3.5. *Let* $g \in \ell_2(\mathbb{Z})$ *be such that* $\hat{g}$ *is real-valued and consider the lattice* Λ *with canonical generator matrix given by*

$$\begin{bmatrix} \frac{N}{2} & 0 \\ 0 & \frac{1}{N} \end{bmatrix}.$$

Then the following conditions are equivalent.

(i) $\{g_{m\frac{N}{2},n\frac{1}{N}}\}_{m\in\mathbb{Z},n=0,\dots,N-1}$ *is a tight frame for* $\ell_2(\mathbb{Z})$ *with frame bound* 2.

(ii) *We have* $|Z\hat{g}(e^{2\pi i t}, y)|^2 + |Z\hat{g}(e^{2\pi i (t+\frac{1}{N})}, y)|^2 = N$ *a.e.*

(iii) *For all* $j \in \{0, \dots, N-1\}$, *we have* $\sum_{l=0}^{N-1} \hat{g}(e^{2\pi i (t + \frac{l}{N})}) \hat{g}(e^{2\pi i (t + \frac{l+2j}{N})}) = N\delta_{j,0}$ *a.e.*

(iv) $\mathcal{W}(g, \Lambda, \ell_2(\mathbb{Z}))$ *is an orthonormal basis for* $\ell_2(\mathbb{Z})$.

Proof. Throughout this proof we choose the normalized Haar measure on $\mathbb{T}$, i.e., $m(E) = \int_0^1 1_E(e^{2\pi i t}) dt$, for all measurable $E \subseteq \mathbb{T}$ and the counting measure on its dual group $\widehat{\mathbb{T}} = \mathbb{Z}$. This choice ensures that the Plancherel formula for $\mathbb{T}$ holds.

Since we will mainly work in the Fourier domain, we first need to compute the Fourier transform of the elements of the Gabor system for the following calculations:

$$\begin{aligned}\widehat{g_{m\frac{N}{2},n\frac{1}{N}}}(e^{2\pi i t}) &= \sum_{l\in\mathbb{Z}} e^{2\pi i \frac{n}{N} l} g(e^{2\pi i (l - m\frac{N}{2})}) e^{-2\pi i l t}\\ &= e^{2\pi i \frac{mn}{2}} e^{-2\pi i t m \frac{N}{2}} \sum_{l\in\mathbb{Z}} g(e^{2\pi i l}) e^{-2\pi i l (t - \frac{n}{N})}\\ &= (-1)^{mn} \hat{g}_{n\frac{1}{N}, -m\frac{N}{2}}(e^{2\pi i t}).\end{aligned}$$

(i) $\Leftrightarrow$ (ii): Since the Fourier transform is a unitary operator, the Gabor system $\{g_{m\frac{N}{2},\frac{1}{N}n}\}_{m\in\mathbb{Z},n=0,\dots,N-1}$ is a tight frame for $\ell_2(\mathbb{Z})$ with frame bound 2 if and only if the Gabor system $\{\hat{g}_{n\frac{1}{N},m\frac{N}{2}}\}_{m\in\mathbb{Z},n=0,\dots,N-1}$ is a tight frame for $L^2(\mathbb{T})$ with frame bound 2. Then we write this set as the disjoint union $\{\hat{g}_{n\frac{2}{N},m\frac{N}{2}}\}_{m\in\mathbb{Z},n=0,\dots,\frac{N}{2}-1} \cup \{(T_{-\frac{1}{N}}\hat{g})_{n\frac{2}{N},m\frac{N}{2}}\}_{m\in\mathbb{Z},n=0,\dots,\frac{N}{2}-1} =: G_1 \cup G_2$, and let S_i denote the frame operator for G_i, $i = 1, 2$.

First we compute the frame operator S_1. For all $f \in L^2(\mathbb{T})$, using the Poisson summation formula [16, Theorem 4.42] applied to $H = \{e^{2\pi i k \frac{2}{N}} : k = 0, \dots, \frac{N}{2} - 1\}$, we obtain

$$
\begin{aligned}
S_1 f(e^{2\pi i t}) &= \sum_{m\in\mathbb{Z}} \sum_{n=0}^{\frac{N}{2}-1} \left\langle f, \hat{g}_{n\frac{2}{N}, m\frac{N}{2}} \right\rangle \hat{g}_{n\frac{2}{N}, m\frac{N}{2}}(e^{2\pi i t}) \\
&= \sum_{m\in\mathbb{Z}} \sum_{n=0}^{\frac{N}{2}-1} \int_0^1 f(e^{2\pi i s}) \overline{\hat{g}(e^{2\pi i (s - n\frac{2}{N})})} e^{-2\pi i m \frac{N}{2} s} ds \, \hat{g}(e^{2\pi i (t - n\frac{2}{N})}) e^{2\pi i m \frac{N}{2} t} \\
&= \sum_{n=0}^{\frac{N}{2}-1} \left[\sum_{m\in\mathbb{Z}} (\widehat{f\overline{T_{n\frac{2}{N}}\hat{g}}})(m\tfrac{N}{2}) e^{2\pi i m \frac{N}{2} t} \right] \hat{g}(e^{2\pi i (t - n\frac{2}{N})}) \\
&= \sum_{n=0}^{\frac{N}{2}-1} \tfrac{2}{N} \sum_{k=0}^{\frac{N}{2}-1} (f\overline{T_{n\frac{2}{N}}\hat{g}})(e^{2\pi i (t + k\frac{2}{N})}) \hat{g}(e^{2\pi i (t - n\frac{2}{N})}).
\end{aligned}
$$

Applying now the Zak transform yields

$$
\begin{aligned}
Z(S_1 f)(e^{2\pi i t}, y) &= \sum_{l=0}^{\frac{N}{2}-1} \tfrac{2}{N} \sum_{n,k=0}^{\frac{N}{2}-1} f(e^{2\pi i (t + \frac{2k+2l}{N})}) \overline{\hat{g}(e^{2\pi i (t + \frac{2k+2l-2n}{N})})} \hat{g}(e^{2\pi i (t + \frac{2l-2n}{N})}) e^{2\pi i \frac{2l}{N} y} \\
&= \tfrac{2}{N} \sum_{l,n,k=0}^{\frac{N}{2}-1} f(e^{2\pi i (t + \frac{2l+2n}{N})}) \overline{\hat{g}(e^{2\pi i (t + \frac{2l}{N})})} \hat{g}(e^{2\pi i (t + \frac{2l-2k}{N})}) e^{2\pi i \frac{2(l-k+n)}{N} y} \\
&= \tfrac{2}{N} \sum_{l=0}^{\frac{N}{2}-1} \overline{\hat{g}(e^{2\pi i (t + \frac{2l}{N})})} \left[\sum_{k=0}^{\frac{N}{2}-1} \hat{g}(e^{2\pi i (t + \frac{2l-2k}{N})}) e^{-2\pi i \frac{2k}{N} y} \right] \\
&\qquad \cdot \left[\sum_{n=0}^{\frac{N}{2}-1} f(e^{2\pi i (t + \frac{2n+2l}{N})}) e^{2\pi i \frac{2n}{N} y} \right] e^{2\pi i \frac{2l}{N} y} \\
&= \tfrac{2}{N} \overline{Z(\hat{g})(e^{2\pi i t}, y)} Z(\hat{g})(e^{2\pi i t}, y) Z(f)(e^{2\pi i t}, y).
\end{aligned}
$$

To compute the Zak transform of S_2, we can use the previous calculation, which yields

$$
\begin{aligned}
Z(S_2 f)(e^{2\pi i t}, y) &= \tfrac{2}{N} Z(f)(e^{2\pi i t}, y) |Z(T_{-\frac{1}{N}}\hat{g})(e^{2\pi i t}, y)|^2 \\
&= \tfrac{2}{N} Z(f)(e^{2\pi i t}, y) |Z(\hat{g})(e^{2\pi i (t + \frac{1}{N})}, y)|^2,
\end{aligned}
$$

since

$$
Z(T_{-\frac{1}{N}}\hat{g})(e^{2\pi i t}, y) = \sum_{k=0}^{\frac{N}{2}-1} \hat{g}(e^{2\pi i (t + \frac{2k}{N} + \frac{1}{N})}) e^{2\pi i \frac{2k}{N} y} = Z(\hat{g})(e^{2\pi i (t + \frac{1}{N})}, y).
$$

Hence (i) is equivalent to

$$
\begin{aligned}
2Z(f)(e^{2\pi i t}, y) &= Z((S_1 + S_2) f)(e^{2\pi i t}, y) \\
&= \tfrac{2}{N} Z(f)(e^{2\pi i t}, y) \left[|Z(\hat{g})(e^{2\pi i t}, y)|^2 + |Z(\hat{g})(e^{2\pi i (t + \frac{1}{N})}, y)|^2 \right] \text{ a.e.,}
\end{aligned}
$$

which holds if and only if (ii) is satisfied.

(ii) $\Leftrightarrow$ (iii): The following properties of the Zak transform will be exploited several times. The reconstruction formula

$$\sum_{y=0}^{\frac{N}{2}-1} Zf(e^{2\pi it},y) = \sum_{k=0}^{\frac{N}{2}-1} f(e^{2\pi i(t+\frac{2k}{N})}) \sum_{y=0}^{\frac{N}{2}-1} e^{2\pi i\frac{2k}{N}y} = \tfrac{N}{2} f(e^{2\pi it})$$

holds a.e., since $\sum_{y=0}^{\frac{N}{2}-1} e^{2\pi i\frac{2k}{N}y} \neq 0$ if and only if $k=0$ by [21, Lemma 23.29] and, if $k=0$, then $\sum_{y=0}^{\frac{N}{2}-1} e^{2\pi i\frac{2k}{N}y} = \frac{N}{2}$. Moreover, we will use that

$$Z\hat{g}(e^{2\pi i(t+\frac{2l}{N})},y) = \sum_{k=0}^{\frac{N}{2}-1} f(e^{2\pi i(t+\frac{2k+2l}{N})})e^{2\pi i\frac{2k}{N}y} = e^{-2\pi i\frac{2l}{N}y} Z\hat{g}(e^{2\pi it},y).$$

The idea is to write the equation in (iii) in terms of the Zak transform. Using the fact that $\hat{g}$ is real-valued, we compute

$$\begin{aligned}
&\sum_{l=0}^{N-1} \hat{g}(e^{2\pi i(t+\frac{l}{N})})\hat{g}(e^{2\pi i(t+\frac{l+2j}{N})})\\
&= \frac{4}{N^2}\sum_{l=0}^{N-1}\sum_{x=0}^{\frac{N}{2}-1} Z\hat{g}(e^{2\pi i(t+\frac{l}{N})},x)\sum_{y=0}^{\frac{N}{2}-1}\overline{Z\hat{g}(e^{2\pi i(t+\frac{l+2j}{N})},y)}\\
&= \frac{4}{N^2}\sum_{k,x,y=0}^{\frac{N}{2}-1}\Big[Z\hat{g}(e^{2\pi i(t+\frac{2k}{N})},x)\overline{Z\hat{g}(e^{2\pi i(t+\frac{2k+2j}{N})},y)}\\
&\quad + Z\hat{g}(e^{2\pi i(t+\frac{2k+1}{N})},x)\overline{Z\hat{g}(e^{2\pi i(t+\frac{2k+2j+1}{N})},y)}\Big]\\
&= \frac{4}{N^2}\sum_{x,y=0}^{\frac{N}{2}-1}\left[\sum_{k=0}^{\frac{N}{2}-1} e^{-2\pi i\frac{2k}{N}(x-y)}\right] e^{2\pi i\frac{2j}{N}y}\Big[Z\hat{g}(e^{2\pi it},x)\overline{Z\hat{g}(e^{2\pi it},y)}\\
&\quad + Z\hat{g}(e^{2\pi i(t+\frac{1}{N})},x)\overline{Z\hat{g}(e^{2\pi i(t+\frac{1}{N})},y)}\Big]\\
&= \frac{2}{N}\sum_{x=0}^{\frac{N}{2}-1}\Big[|Z\hat{g}(e^{2\pi it},x)|^2 + |Z\hat{g}(e^{2\pi i(t+\frac{1}{N})},x)|^2\Big] e^{2\pi i\frac{2j}{N}x}\\
&= \frac{2}{N}\Big[|Z\hat{g}(e^{2\pi it},\cdot)|^2 + |Z\hat{g}(e^{2\pi i(t+\frac{1}{N})},\cdot)|^2\Big]^{\vee}(j),
\end{aligned}$$

where the inverse Fourier transform is taken in $\mathbb{Z}_{N/2}$. This shows that (iii) is equivalent to

$$\frac{2}{N}\Big[|Z\hat{g}(e^{2\pi it},\cdot)|^2 + |Z\hat{g}(e^{2\pi i(t+\frac{1}{N})},\cdot)|^2\Big]^{\vee}(j) = N\delta_{j,0}. \tag{3.4}$$

If (ii) holds, then

$$\frac{2}{N}\Big[|Z\hat{g}(e^{2\pi it},\cdot)|^2 + |Z\hat{g}(e^{2\pi i(t+\frac{1}{N})},\cdot)|^2\Big]^{\vee}(j) = \frac{2}{N}N\sum_{x=0}^{\frac{N}{2}-1} e^{2\pi i\frac{2j}{N}x} = N\delta_{j,0},$$

which is (3.4). On the other hand, the inverse Fourier transform is injective. This proves that (3.4) holds if and only if (ii) is true and thus (ii) $\Leftrightarrow$ (iii).

(iii) $\Leftrightarrow$ (iv): First we remark that $\mathcal{W}(g, \Lambda, \ell_2(\mathbb{Z}))$ is an orthonormal basis if and only if the set

$$\Psi := \{T_{nN} f_m : m = 1, \dots, N,\ n \in \mathbb{Z}\},$$

where

$$\begin{aligned} f_1(x) &= g(x), \\ f_N(x) &= g_{0,\frac{N}{2}}(x), \\ f_{2l+k}(x) &= \frac{(-1)^{kl}}{\sqrt{2}}(g_{k\frac{N}{2},\frac{l}{N}} + (-1)^{k+l} g_{k\frac{N}{2},-\frac{l}{N}}),\ l = 1, \dots, \tfrac{N}{2}-1,\ k = 0, 1, \end{aligned}$$

is an orthonormal basis, since these elements differ from the elements in $\mathcal{W}(g, \Lambda, \ell_2(\mathbb{Z}))$ only by factors of absolute value 1. Next, notice that to prove (iv) it is sufficient and necessary that

$$\|T_{nN} f_m\|_2 = 1, \quad m = 1, \dots, L,\ n \in \mathbb{Z} \tag{3.5}$$

and

$$\sum_{m=1}^{L} \sum_{n \in \mathbb{Z}} \langle h_1, T_{nN} f_m \rangle \langle T_{nN} f_m, h_2 \rangle = \langle h_1, h_2 \rangle \quad \text{for all } h_1, h_2 \in \ell_2(\mathbb{Z}). \tag{3.6}$$

We start by dealing with (3.5). Using the Plancherel theorem, we compute

$$1 = \|T_{nN} f_1\|_2^2 = \|\hat{g}\|_2^2 = \int_0^1 \hat{g}(e^{2\pi i t}) \hat{g}(e^{2\pi i t}) dt,$$

$$1 = \|T_{nN} f_N\|_2^2 = \|\hat{g}_{\frac{N}{2},0}\|^2 = \int_0^1 \hat{g}(e^{2\pi i t}) \hat{g}(e^{2\pi i t}) dt,$$

and, for $m = 2, \dots, N-1$,

$$\begin{aligned} 1 &= \|T_{nN} f_m\|_2^2 \\ &= \|\tfrac{1}{\sqrt{2}}(\hat{g}_{\frac{l}{N},-k\frac{N}{2}} + (-1)^{k+l} \hat{g}_{-\frac{l}{N},-k\frac{N}{2}})\|_2^2 \\ &= \frac{1}{2}\int_0^1 |e^{-2\pi i \frac{kN}{2} t}|^2 |\hat{g}(e^{2\pi i (t-\frac{l}{N})}) + (-1)^{k+l} \hat{g}(e^{2\pi i (t+\frac{l}{N})})|^2 dt \\ &= \frac{1}{2}\int_0^1 \Big[|\hat{g}(e^{2\pi i (t-\frac{l}{N})})|^2 + |\hat{g}(e^{2\pi i (t+\frac{l}{N})})|^2 + (-1)^{k+l} \hat{g}(e^{2\pi i (t-\frac{l}{N})}) \overline{\hat{g}(e^{2\pi i (t+\frac{l}{N})})} \\ &\quad + (-1)^{k+l} \overline{\hat{g}(e^{2\pi i (t-\frac{l}{N})})} \hat{g}(e^{2\pi i (t+\frac{l}{N})})\Big] dt. \end{aligned}$$

Since $\hat{g}$ is real-valued, we can continue the last computation and obtain that

$$1 = \|T_{nN} f_m\|_2^2 = \|\hat{g}\|_2^2 + (-1)^{k+l} \int_0^1 \hat{g}(e^{2\pi i t}) \hat{g}(e^{2\pi i (t+\frac{2l}{N})}) dt.$$

Combining the above computations, we have proved that (3.5) holds if and only if

$$(3.7)\qquad \int_0^1 \hat{g}(e^{2\pi it})\hat{g}(e^{2\pi i(t+\frac{2j}{N})})dt = \delta_{j,0} \quad \text{for all } j \in \{0,\dots,N-1\}.$$

Now we turn to the study of condition (3.6). Using the Plancherel formula and the Poisson summation formula [16, Theorem 4.42] applied to $H = \{e^{2\pi i\frac{k}{N}} : k = 0,\dots,N-1\}$, we obtain

$$\begin{aligned}
&\sum_{m=1}^{N}\sum_{n\in\mathbb{Z}} \langle h_1, T_{nN}f_m\rangle \langle T_{nN}f_m, h_2\rangle \\
&= \sum_{m=1}^{N}\sum_{n\in\mathbb{Z}} \left\langle \widehat{h_1}, \widehat{T_{nN}f_m}\right\rangle \left\langle \widehat{T_{nN}f_m}, \widehat{h_2}\right\rangle \\
&= \sum_{m=1}^{N}\sum_{n\in\mathbb{Z}} \int_0^1 (\widehat{h_1}\overline{\widehat{f_m}})(e^{2\pi it}) \int_0^1 (\widehat{f_m}\overline{\widehat{h_2}})(e^{2\pi is})e^{-2\pi isNn}ds\, e^{2\pi itNn}dt \\
&= \sum_{m=1}^{N} \int_0^1 (\widehat{h_1}\overline{\widehat{f_m}})(e^{2\pi it}) \left[\sum_{n\in\mathbb{Z}} (\widehat{f_m}\overline{\widehat{h_2}})\hat{}\,(Nn)e^{2\pi itNn}\right] dt \\
&= \sum_{m=1}^{N} \int_0^1 (\widehat{h_1}\overline{\widehat{f_m}})(e^{2\pi it}) \tfrac{1}{N}\sum_{r=0}^{N-1} (\widehat{f_m}\overline{\widehat{h_2}})(e^{2\pi i(t+\frac{r}{N})})\, dt,
\end{aligned}$$

which equals $\langle h_1, h_2\rangle$ if and only if

$$(3.8)\qquad \sum_{m=1}^{N} \overline{\widehat{f_m}(e^{2\pi it})}\widehat{f_m}(e^{2\pi i(t+\frac{r}{N})}) = N\delta_{r,0} \quad \text{for all } r \in \{0,\dots,N-1\}.$$

Setting $\mathbb{L} := \{-\frac{N}{2}+1,\dots,-1,1,\dots,\frac{N}{2}-1\}$, we compute

$$\begin{aligned}
&\sum_{m=1}^{N} \overline{\widehat{f_m}(e^{2\pi it})}\widehat{f_m}(e^{2\pi i(t+\frac{r}{N})}) \\
&= \hat{g}(e^{2\pi it})\hat{g}(e^{2\pi i(t+\frac{r}{N})}) + \hat{g}_{\frac{N}{2},0}(e^{2\pi it})\hat{g}_{\frac{N}{2},0}(e^{2\pi i(t+\frac{r}{N})}) + \frac{1}{2}\sum_{l=1}^{\frac{N}{2}-1}\sum_{k=0}^{1} [\overline{\hat{g}_{\frac{l}{N},-k\frac{N}{2}}(e^{2\pi it})} \\
&\quad + (-1)^{k+l}\overline{\hat{g}_{-\frac{l}{N},-k\frac{N}{2}}(e^{2\pi it})}][\hat{g}_{\frac{l}{N},-k\frac{N}{2}}(e^{2\pi i(t+\frac{r}{N})}) + (-1)^{k+l}\hat{g}_{-\frac{l}{N},-k\frac{N}{2}}(e^{2\pi i(t+\frac{r}{N})})] \\
&= \hat{g}(e^{2\pi it})\hat{g}(e^{2\pi i(t+\frac{r}{N})}) + \hat{g}(e^{2\pi i(t-\frac{N}{2})})\hat{g}(e^{2\pi i(t+\frac{r}{N}-\frac{N}{2})}) \\
&\quad + \frac{1}{2}\sum_{l=1}^{\frac{N}{2}-1}\sum_{k=0}^{1} e^{-2\pi ik\frac{N}{2}\frac{r}{N}}[\hat{g}(e^{2\pi i(t-\frac{l}{N})}) \\
&\quad + (-1)^{k+l}\hat{g}(e^{2\pi i(t+\frac{l}{N})})][\hat{g}(e^{2\pi i(t-\frac{l}{N}+\frac{r}{N})}) + (-1)^{k+l}\hat{g}(e^{2\pi i(t+\frac{l}{N}+\frac{r}{N})})] \\
&= \hat{g}(e^{2\pi it})\hat{g}(e^{2\pi i(t+\frac{r}{N})}) + \hat{g}(e^{2\pi i(t-\frac{N}{2})})\hat{g}(e^{2\pi i(t+\frac{r}{N}-\frac{N}{2})}) \\
&\quad + \frac{1}{2}\sum_{l=1}^{\frac{N}{2}-1}\sum_{k=0}^{1} (-1)^{kr}[\hat{g}(e^{2\pi i(t-\frac{l}{N})})\hat{g}(e^{2\pi i(t-\frac{l}{N}+\frac{r}{N})}) + \hat{g}(e^{2\pi i(t+\frac{l}{N})})\hat{g}(e^{2\pi i(t+\frac{l}{N}+\frac{r}{N})}) \\
&\quad + (-1)^{k+l}(e^{2\pi i(t-\frac{l}{N})})\hat{g}(e^{2\pi i(t+\frac{l}{N}+\frac{r}{N})}) + \hat{g}(e^{2\pi i(t+\frac{l}{N})})\hat{g}(e^{2\pi i(t-\frac{l}{N}+\frac{r}{N})}))]
\end{aligned}$$

$$= \hat{g}(e^{2\pi it})\hat{g}(e^{2\pi i(t+\frac{r}{N})}) + \sum_{l\in\mathbb{L}} \hat{g}(e^{2\pi i(t+\frac{l}{N})})\hat{g}(e^{2\pi i(t+\frac{l}{N}+\frac{r}{N})}) \left[\frac{1+(-1)^r}{2}\right]$$
$$+ \sum_{l\in\mathbb{L}} (-1)^l \hat{g}(e^{2\pi i(t+\frac{l}{N})})\hat{g}(e^{2\pi i(t-\frac{l}{N}+\frac{r}{N})}) \left[\frac{1+(-1)^{r+1}}{2}\right]$$
$$+\hat{g}(e^{2\pi i(t-\frac{N}{2})})\hat{g}(e^{2\pi i(t+\frac{r}{N}-\frac{N}{2})}).$$

If r is even, i.e., $r = 2j$, we obtain

$$\sum_{m=1}^{N} \overline{\widehat{f_m}(e^{2\pi it})}\widehat{f_m}(e^{2\pi i(t+\frac{r}{N})}) = \sum_{l=0}^{N-1} \hat{g}(e^{2\pi i(t+\frac{l}{N})})\hat{g}(e^{2\pi i(t+\frac{l}{N}+\frac{2j}{N})}),$$

and if r is odd, i.e., $r = 2j+1$, we obtain

$$\begin{aligned}
&\sum_{m=1}^{N} \overline{\widehat{f_m}(e^{2\pi it})}\widehat{f_m}(e^{2\pi i(t+\frac{r}{N})}) \\
&= \hat{g}(e^{2\pi it})\hat{g}(e^{2\pi i(t+\frac{r}{N})}) + \hat{g}(e^{2\pi i(t-\frac{N}{2})})\hat{g}(e^{2\pi i(t+\frac{r}{N}-\frac{N}{2})}) \\
&\quad + \sum_{l\in\mathbb{L}} (-1)^l \hat{g}(e^{2\pi i(t+\frac{l}{N})})\hat{g}(e^{2\pi i(t-\frac{l}{N}+\frac{r}{N})}) \\
&= 0.
\end{aligned}$$

This shows that (3.6) holds if and only if

$$\sum_{l=0}^{N-1} \hat{g}(e^{2\pi i(t+\frac{l}{N})})\hat{g}(e^{2\pi i(t+\frac{l+2j}{N})}) = N\delta_{j,0}.$$

Moreover, this equation implies equation (3.7), since

$$\begin{aligned}
\int_0^1 \hat{g}(e^{2\pi it})\hat{g}(e^{2\pi i(t+\frac{2j}{N})})dt &= \int_0^{\frac{1}{N}} \sum_{l=0}^{N-1} \hat{g}(e^{2\pi i(t+\frac{l}{N})})\hat{g}(e^{2\pi i(t+\frac{l+2j}{N})})dt \\
&= \int_0^{\frac{1}{N}} N\delta_{j,0}dt = \delta_{j,0}.
\end{aligned}$$

This shows (iii) $\Leftrightarrow$ (iv), and hence the theorem is proved. $\square$

Now we will study the case of a general time-frequency lattice.

PROPOSITION 3.6. *Let Λ be a lattice in $\mathbb{Z}\times\mathbb{T}$ with $\mathrm{vol}(\Lambda) = \frac{1}{2}$ with canonical generator matrix A given by*

$$A = \begin{bmatrix} \frac{N}{2} & b \\ 0 & \frac{1}{N} \end{bmatrix}.$$

Let $g \in \ell_2(\mathbb{Z})$, let $m_0, n_0 \in \mathbb{Z}$ be chosen such that $\frac{N}{2}m_0 + bn_0 = \gcd(\frac{N}{2}, b) =: c$, and let U be defined on $\ell_2(\mathbb{Z})$ by

$$Uf(k) = f(k)e^{\pi i \frac{n_0}{cN}k^2}.$$

Then

$$g_{m\frac{N}{2}+nb,n\frac{1}{N}}(l) = C(m,n)U(U^{-1}g)_{m\frac{N}{2}+nb,-m\frac{n_0}{2c}-n\frac{bn_0}{cN}+n\frac{1}{N}}(l), \quad l \in \mathbb{Z},$$

where $C(m,n) = e^{\pi i \frac{n_0}{cN}(m\frac{N}{2}+nb)^2}$.

Proof. Let $\sigma \in \mathrm{Hom}(\mathbb{Z}\times\mathbb{T})$ be defined by

$$\sigma = \begin{bmatrix} I_{\mathbb{Z}} & 0 \\ -\frac{n_0}{cN} & I_{\mathbb{T}} \end{bmatrix}.$$

It is easy to check that σ is symplectic on $\mathbb{Z}\times\mathbb{T}$. In order to apply Theorem 1.1, we need to compute a second degree character of $\mathbb{Z}\times\mathbb{T}$ associated to $\zeta = \sigma^*\kappa_0\sigma - \kappa_0$, where κ_0 is defined by

$$\kappa_0 = \begin{bmatrix} 1 & 1 \\ I_{\mathbb{Z}} & 0 \end{bmatrix} \in \mathrm{Hom}(\mathbb{Z}\times\mathbb{T}, \mathbb{T}\times\mathbb{Z}).$$

First we note that $\sigma^*(z,n) = (ze^{2\pi i(-\frac{n_0}{cN})n}, n)$, since

$$\langle (m,t), \sigma^*(z,n)\rangle = \langle \sigma(m,t),(z,n)\rangle = z^m(e^{2\pi i(-\frac{n_0}{cN})m}t)^n = \left\langle (m,t), (ze^{2\pi i(-\frac{n_0}{cN})n}, n)\right\rangle.$$

Thus

$$\zeta(n,z) = (\sigma^*\kappa_0\sigma - \kappa_0)(n,z) = \sigma^*\kappa_0(n, e^{2\pi i(-\frac{n_0}{cN})n}z) - (1,n) = (e^{2\pi i(-\frac{n_0}{cN})n}, 0).$$

The map

$$\psi : \mathbb{Z}\times\mathbb{T}\to\mathbb{T}, \quad \psi(m,t) = e^{-\pi i \frac{n_0}{cN}m^2}$$

is a second degree character associated to ζ as the following calculation shows:

$$\begin{aligned} \psi(m,t)\psi(n,z)\langle (m,t),\zeta(n,z)\rangle &= e^{-\pi i\frac{n_0}{cN}m^2} e^{-\pi i\frac{n_0}{cN}n^2} e^{2\pi i(-\frac{n_0}{cN})mn} \\ &= e^{-\pi i \frac{n_0}{cN}(m+n)^2} \\ &= \psi((m,t)+(n,z)). \end{aligned}$$

Next notice that

$$\sigma(m\tfrac{N}{2}+nb, n\tfrac{1}{N}) = (m\tfrac{N}{2}+nb, -m\tfrac{n_0}{2c} - n\tfrac{bn_0}{cN} + n\tfrac{1}{N}). \tag{3.9}$$

Now we can apply Theorem 1.1, which proves the claim. □

Next we define a Wilson basis associated with a lattice with arbitrary canonical generator matrix. For this, the following mapping will turn out to be very useful.

LEMMA 3.7. *Let* $\frac{N}{2}, b \in \mathbb{Z}$ *with* $0 \le b < \frac{N}{2}$, *and let* $m_0, n_0 \in \mathbb{Z}$ *be chosen such that* $\frac{N}{2}m_0 + bn_0 = \gcd(\frac{N}{2}, b) =: c$. *Further, let* $d := \mathrm{lcm}(\frac{N}{2}, b)$. *Then the mapping* $\varphi : \mathbb{Z}^2 \to \mathbb{Z}^2$ *defined by*

$$\varphi(m,n) = \begin{cases} (m,n) & : \ b = 0, \\ (mm_0 - \frac{2d}{N}n, mn_0 + \frac{d}{b}n) & : \ b \neq 0 \end{cases}$$

is bijective and, for all $m \in \mathbb{Z}$, *we have*

$$\{(m, n \bmod 2c) : (m,n) \in \varphi^{-1}(\mathbb{Z}\times\{0,\dots,N-1\})\} = \{m\}\times\{0,\dots,2c-1\}$$

with

$$|\{n : (m,n)\in\varphi^{-1}(\mathbb{Z}\times\{0,\dots,N-1\})\}| = 2c.$$

Proof. We only need to study the case $b \neq 0$. For this, let $(m,n),(m',n') \in \mathbb{Z}^2$ be such that $\varphi(m,n) = \varphi(m',n')$. Then

$$\tfrac{N}{2}(mm_0 - \tfrac{2d}{N}n) + b(mn_0 + \tfrac{d}{b}n) = \tfrac{N}{2}(m'm_0 - \tfrac{2d}{N}n') + b(m'n_0 + \tfrac{d}{b}n'),$$

which holds if and only if

$$m(\tfrac{N}{2}m_0 + bn_0) = m'(\tfrac{N}{2}m_0 + bn_0),$$

and hence $m = m'$. This implies

$$(-\tfrac{2d}{N}n, \tfrac{d}{b}n) = (-\tfrac{2d}{N}n', \tfrac{d}{b}n'),$$

which yields $n = n'$. This proves that φ is injective.

To show that φ is surjective, let $(k,l) \in \mathbb{Z}^2$ and consider $M := \frac{N}{2}k + bl$. It is well known that there exists some $m \in \mathbb{Z}$ with $M = mc$. Furthermore, we have

$$\{(p,q) \in \mathbb{Z}^2 : \tfrac{N}{2}p + bq = mc\} = \{(mm_0 - \tfrac{2d}{N}n, mn_0 + \tfrac{d}{b}n) : n \in \mathbb{Z}\},$$

since $\frac{N}{2}p + bq = \frac{N}{2}p' + bq'$ if and only if $\frac{N}{2}(p-p') = b(q'-q)$. This yields the existence of some $n \in \mathbb{Z}$ with

$$\varphi(m,n) = (mm_0 - \tfrac{2d}{N}n, mn_0 + \tfrac{d}{b}n) = (k,l).$$

Secondly, we will prove the second part of the lemma. First observe that $m,n \in \mathbb{Z}$ satisfy

$$\varphi(m,n) \in \mathbb{Z} \times \{0,\dots,N-1\} \tag{3.10}$$

if and only if they satisfy

$$-\tfrac{b}{d}n_0m \le n \le -\tfrac{b}{d}n_0m + \tfrac{b}{d}(N-1) = 2c - \tfrac{b}{d}n_0m - \tfrac{b}{d}.$$

Hence, for each fixed $m \in \mathbb{Z}$, the set of $n \in \mathbb{Z}$ such that (3.10) is satisfied equals

$$S_m := \{\lceil -\tfrac{b}{d}n_0m \rceil, \dots, \lfloor 2c - \tfrac{b}{d}n_0m - \tfrac{b}{d} \rfloor\}.$$

To finish the proof we claim that

$$|S_m| = 2c \quad \text{for all } m \in \mathbb{Z}. \tag{3.11}$$

For this, fix $m \in \mathbb{Z}$ and let $k \in \mathbb{Z}$ and $l \in \{0,\dots,\frac{d}{b}-1\}$ be such that $-n_0m = k\frac{d}{b} + l$. Then we obtain $\lceil -\frac{b}{d}n_0m \rceil = k$ if $l = 0$ and otherwise $\lceil -\frac{b}{d}n_0m \rceil = k+1$. Moreover, we have

$$\lfloor 2c - \tfrac{b}{d}n_0m - \tfrac{b}{d} \rfloor = \left\lfloor 2c + k + \frac{l-1}{\frac{d}{b}} \right\rfloor,$$

which equals $2c + k - 1$ if $l = 0$ and otherwise $2c + k$. Thus the second part of the lemma is proved. ∎

Note that the following definition reduces to Definition 3.4 in the case of a diagonal canonical generator matrix.

DEFINITION 3.8. *Let Λ be a lattice in $\mathbb{Z} \times \mathbb{T}$ with canonical generator matrix A given by*

$$A = \begin{bmatrix} \frac{N}{2} & b \\ 0 & \frac{1}{N} \end{bmatrix}.$$

Let $g \in \ell_2(\mathbb{Z})$, and let φ be defined as in Lemma 3.7. *Then the Wilson system $\mathcal{W}(g,\Lambda,\ell_2(\mathbb{Z})) = \{\psi_{m,n}\}_{m\in\mathbb{Z},n=0,\dots,\frac{N}{2}}$ is given by*

$$\psi^{\Lambda}_{m,n} = g_{\varphi_1(2m,n)\frac{N}{2}+\varphi_2(2m,n)b,\varphi_2(2m,n)\frac{1}{N}} \qquad \text{if } m\in\mathbb{Z},\ n=0,\tfrac{N}{2},$$

and for $m \in \mathbb{Z}$, $n = 1,\dots,\frac{N}{2}-1$,

$$\begin{aligned}
\psi^{\Lambda}_{m,n} &= \tfrac{1}{\sqrt{2}}(g_{\varphi_1(m,n)\frac{N}{2}+\varphi_2(m,n)b,\varphi_2(m,n)\frac{1}{N}} \\
&\quad + g_{\varphi_1(m,-n)\frac{N}{2}+\varphi_2(m,-n)b,\varphi_2(m,-n)\frac{1}{N}}) \quad \text{if } m+n \text{ is even,}\\
\psi^{\Lambda}_{m,n} &= \tfrac{i}{\sqrt{2}}(g_{\varphi_1(m,n)\frac{N}{2}+\varphi_2(m,n)b,\varphi_2(m,n)\frac{1}{N}} \\
&\quad - g_{\varphi_1(m,-n)\frac{N}{2}+\varphi_2(m,-n)b,\varphi_2(m,-n)\frac{1}{N}}) \quad \text{if } m+n \text{ is odd.}
\end{aligned}$$

The following theorem gives an equivalent condition for a Wilson system with respect to an arbitrary time-frequency lattice to form an orthonormal basis in terms of a frame condition for the associated Gabor system.

THEOREM 3.9. *Let Λ be a lattice in $\mathbb{Z}\times\mathbb{T}$ with canonical generator matrix A given by*

$$A = \begin{bmatrix} \frac{N}{2} & b \\ 0 & \frac{1}{N} \end{bmatrix}.$$

Let $g \in \ell_2(\mathbb{Z})$ be such that $\widehat{U^{-1}g}$ is real-valued, let $M := 2c$, and let U and φ be defined as in Proposition 3.6 *and Lemma* 3.7, *respectively. Then the following conditions are equivalent.*

(i) $\{g_{m\frac{N}{2}+nb,n\frac{1}{N}}\}_{m\in\mathbb{Z},n=0,\dots,N-1}$ *is a tight frame for $\ell_2(\mathbb{Z})$ with frame bound* 2.

(ii) $\{(U^{-1}g)_{m\frac{M}{2},n\frac{1}{M}}\}_{m\in\mathbb{Z},n=0,\dots,M-1}$ *is a tight frame for $\ell_2(\mathbb{Z})$ with frame bound* 2.

(iii) $\mathcal{W}(U^{-1}g,\frac{M}{2}\mathbb{Z}\times\frac{1}{M}\{0,\dots,M-1\},\ell_2(\mathbb{Z}))$ *is an orthonormal basis for $\ell_2(\mathbb{Z})$.*

(iv) $\mathcal{W}(g,\Lambda,\ell_2(\mathbb{Z}))$ *is an orthonormal basis for $\ell_2(\mathbb{Z})$.*

Proof. Let σ be defined as in the proof of Proposition 3.6 and let $\varphi = (\varphi_1,\varphi_2)$. Then we compute

$$\begin{aligned}
&\sigma(\varphi_1(m,n)\tfrac{N}{2}+\varphi_2(m,n)b,\varphi_2(m,n)\tfrac{1}{N})\\
&= (\varphi_1(m,n)\tfrac{N}{2}+\varphi_2(m,n)b, -\varphi_1(m,n)\tfrac{n_0}{2c}-\varphi_2(m,n)\tfrac{bn_0}{cN}+\varphi_2(m,n)\tfrac{1}{N})\\
&= (m(\tfrac{N}{2}m_0+bn_0)+n(-\tfrac{2d}{N}\tfrac{N}{2}+\tfrac{d}{b}b), m(-\tfrac{n_0}{2c}m_0-\tfrac{bn_0}{cN}n_0+\tfrac{1}{N}n_0)\\
&\quad +n(\tfrac{2d}{N}\tfrac{n_0}{2c}-\tfrac{d}{b}\tfrac{bn_0}{cN}+\tfrac{d}{b}\tfrac{1}{N}))\\
&= (mc, mn_0(\tfrac{1}{N}-\tfrac{1}{cN}(\tfrac{N}{2}m_0+bn_0))+n\tfrac{d}{bN})\\
&= (m\tfrac{M}{2}, n\tfrac{1}{M}),
\end{aligned}$$

where in the last step we used $cd = \frac{N}{2}b$. Using Lemma 3.7, the equivalence of (i) and (ii) now follows immediately from Proposition 3.6 and (3.9), since U is unitary and $|C(m,n)| = 1$. Proposition 3.5 proves (ii) $\Leftrightarrow$ (iii). Therefore it remains to prove the equivalence of (iii) and (iv). For this, we will use the following implication of Proposition 3.6:

$$\begin{aligned}
U(U^{-1}g)_{m\frac{M}{2},n\frac{1}{M}} &= U(U^{-1}g)_{\sigma(\varphi_1(m,n)\frac{N}{2}+\varphi_2(m,n)b,\varphi_2(m,n)\frac{1}{N})}\\
&= C(\varphi(m,n))^{-1} g_{\varphi_1(m,n)\frac{N}{2}+\varphi_2(m,n)b,\varphi_2(m,n)\frac{1}{N}}.
\end{aligned}$$

Further, notice that $C(\varphi(m,n))^{-1}$ does not depend on the sign of n, since

$$C(\varphi(m,n))^{-1} = e^{-\pi i \frac{n_0}{cN}(\varphi_1(m,n)\frac{N}{2}+\varphi_2(m,n)b)^2} = e^{-\pi i \frac{n_0}{cN} m^2 (m_0 \frac{N}{2}+n_0 b)^2} = e^{-\pi i \frac{n_0}{N} m^2 c^2}.$$

The definition of a Wilson basis, the fact that U is a unitary operator, and the fact that $|C(\varphi(m,n))| = 1$ yields the result. □

4. Wilson bases for general lattices—the finite case. The space $\mathbb{C}^L$ has several advantages over $\ell^2(\mathbb{Z})$ when constructing numerical methods for practical time-frequency analysis, which often allow a further acceleration of numerical algorithms; e.g., see [1, 29, 32].

Before defining Gabor systems and Wilson systems for $\mathbb{C}^L$ for general time-frequency lattices, we first prove that each such lattice not only possesses a uniquely determined generator matrix in Hermite normal form (which was already proved in [20]), but in our situation this matrix attains a special form.

LEMMA 4.1. *Let Λ be a lattice in $\mathbb{Z}_L \times \mathbb{Z}_L$ with generator matrix A given by*

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \text{with } a,b,c,d \in \mathbb{N}, \text{ and } \det(A) = \frac{L}{2},$$

and denote $p = \gcd(c,d)$ if $c \neq 0$ and $p = d$ if $c = 0$. Then Λ possesses a uniquely determined generator matrix of the form

$$A' = \begin{bmatrix} \frac{L}{2p} & b' \\ 0 & p \end{bmatrix}, \tag{4.1}$$

where $p = \gcd(c,d)$ and $0 \leq b' < \frac{L}{2p}$.

Proof. Let $p = \gcd(c,d)$ with $c = qp, d = rp$ and note that $p|\frac{L}{2}$ since $\det(A) = ad - bc = (ar - bq)p = \frac{L}{2}$. Since $\frac{d}{p} = r$ and $\frac{-c}{p} = -q$, we have

$$A \begin{bmatrix} r \\ -q \end{bmatrix} = \begin{bmatrix} \frac{L}{2p} \\ 0 \end{bmatrix}. \tag{4.2}$$

Furthermore, we claim that there exists a $z \in \mathbb{Z}$ with $0 \leq z < \frac{L}{2p}$ such that the point $\left[\begin{smallmatrix} z \\ p \end{smallmatrix}\right]$ belongs to Λ. This can be seen as follows: The condition $\left[\begin{smallmatrix} z \\ p \end{smallmatrix}\right] \in \Lambda$ is equivalent to the existence of $m, n \in \mathbb{Z}$ such that

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} m \\ n \end{bmatrix} = \begin{bmatrix} z \\ p \end{bmatrix}. \tag{4.3}$$

Consider the equation $cm + dn = p$ and substitute $c = qp, d = rp$, then $qpm + rpn = p$, hence $qm + rn = 1$. Since r and q are relative prime, there exist $m, n \in \mathbb{Z}$ such that $qm + rn = 1$ (see, e.g., [22, Theorem 4.4]). Thus (4.3) holds for $z \in \mathbb{Z}$, but we still have to show that it holds under the condition $0 \leq z < \frac{L}{2p}$. We can write $z = b' + k\frac{L}{2p}$ with $0 \leq b' < \frac{L}{2p}$ and $k \in \mathbb{Z}$. Hence

$$\begin{bmatrix} z \\ p \end{bmatrix} = \begin{bmatrix} b' \\ p \end{bmatrix} + k \begin{bmatrix} \frac{L}{2p} \\ 0 \end{bmatrix}.$$

Since $\left[\begin{smallmatrix} \frac{L}{2p} \\ 0 \end{smallmatrix}\right] = A\left[\begin{smallmatrix} r \\ -q \end{smallmatrix}\right]$ by (4.2), it follows that $\left[\begin{smallmatrix} b' \\ p \end{smallmatrix}\right] \in \Lambda$. Consequently the matrix

$$A' = \begin{bmatrix} \frac{L}{2p} & b' \\ 0 & p \end{bmatrix}$$

(which satisfies $\det(A') = \frac{L}{2}$) generates Λ.

The fact that this matrix is uniquely determined is an immediate consequence from the condition $0 \leq b' < \frac{L}{2p}$. $\square$

DEFINITION 4.2. *Let Λ be a lattice in $\mathbb{Z}_L \times \mathbb{Z}_L$. Then the uniquely determined matrix A' of Lemma* 4.1 *is called the* canonical generator matrix *for Λ.*

Using the notion of a canonical generator matrix, we first give the definition of a Gabor system.

DEFINITION 4.3. *Let Λ be a lattice in $\mathbb{Z}_L \times \mathbb{Z}_L$ with canonical generator matrix A given by*

$$A = \begin{bmatrix} \frac{L}{2p} & b \\ 0 & p \end{bmatrix}.$$

Set $M = 2p, N = \frac{L}{p}$ and let g be some L-periodic function on $\mathbb{Z}$. Then the associated Gabor system is given by $\{g_{ma+nb,nd}\}_{m=0,\dots,M-1,n=0,\dots,N-1}$, where

$$g_{ma+nb,nd}(l) = g(l-(ma+nb))e^{2\pi i lnd/L}, \qquad l = 0,\dots,L-1.$$

Next we define a Wilson basis associated with a lattice with diagonal canonical generator matrix in the following way.

DEFINITION 4.4. *Let Λ be a lattice in $\mathbb{Z}_L \times \mathbb{Z}_L$ with canonical generator matrix A given by*

$$A = \begin{bmatrix} \frac{L}{2p} & 0 \\ 0 & p \end{bmatrix}$$

and let g be some L-periodic function on $\mathbb{Z}$. Then the Wilson system $\mathcal{W}(g,\Lambda,\mathbb{C}^L) = \{\psi_{m,n}\}_{(m,n)\in I}$, where $I = \{0,\dots,p-1\} \times \{0, \frac{L}{2p}\} \cup \{0,\dots,2p-1\} \times \{1,\dots,\frac{L}{2p}-1\}$, is given by

$$\psi^{\Lambda}_{m,n} = g_{m\frac{L}{p},np} \qquad \text{if } m = 0,\dots,p-1,\ n = 0, \tfrac{L}{2p},$$

and for $m = 0,\dots,2p-1$, $n = 1,\dots,\frac{L}{2p}-1$,

$$\psi^{\Lambda}_{m,n} = \tfrac{1}{\sqrt{2}}(g_{m\frac{L}{2p},np} + g_{m\frac{L}{2p},-np}) \quad \text{if } m+n \text{ is even,}$$
$$\psi^{\Lambda}_{m,n} = \tfrac{i}{\sqrt{2}}(g_{m\frac{L}{2p},np} - g_{m\frac{L}{2p},-np}) \quad \text{if } m+n \text{ is odd.}$$

Also in the finite case we will employ the Zak transform. This time we will use the Zak transform on the group $\mathbb{Z}_L$ with respect to the uniform lattice $K = \{2pk : k = 0,\dots,\frac{L}{2p}-1\}$ in $\mathbb{Z}_L$, which is defined on the set of square-integrable functions on the set $\{0,\dots,2p-1\} \times \{0,\dots,\frac{L}{2p}-1\}$ by

$$Zf(x,y) = \sum_{k=0}^{\frac{L}{2p}-1} f(x+2pk)e^{2\pi i \frac{2pk}{L}y},$$

where we associate $\mathbb{Z}_L$ with $\{0,\dots,L-1\}$.

The following proposition is the analogue to Proposition 3.5 for the space $\mathbb{C}^L$.

PROPOSITION 4.5. *Let g be some L–periodic function on $\mathbb{Z}$ such that $\hat{g}$ is real-valued and consider the lattice Λ with canonical generator matrix given by*

$$\begin{bmatrix} \frac{L}{2p} & 0 \\ 0 & p \end{bmatrix}.$$

Then the following conditions are equivalent.

(i) $\{g_{m\frac{L}{2p},np}\}_{m=0,\dots,2p-1,n=0,\dots,\frac{L}{p}-1}$ *is a tight frame for* $\mathbb{C}^L$ *with frame bound* 2.
(ii) *We have* $|Z\hat{g}(x,y)|^2 + |Z\hat{g}(x+p,y)|^2 = \frac{1}{p}$ *a.e.*
(iii) *For all* $j = 0,\dots,\frac{L}{2p}-1$ *and* $y \in \mathbb{Z}_L$, *we have* $\sum_{l=0}^{\frac{L}{p}-1} \hat{g}(y+lp)\hat{g}(y+lp+2jp) = \frac{1}{p}\delta_{j,0}$.
(iv) $\mathcal{W}(g,\Lambda,\mathbb{C}^L)$ *is an orthonormal basis for* $\mathbb{C}^L$.

Proof. The proof, while lengthy, is very similar to the proof of Proposition 3.5. In fact, with obvious adaptations, such as using the normalized Haar measure on $\mathbb{Z}_L$, i.e., $m(E) = \frac{1}{L}\sum_{x\in\mathbb{Z}_L} 1_E(x)$ for all $E \subseteq \mathbb{Z}_L$, and replacing Zak transforms and Fourier transforms by their corresponding finite counterparts, the proof carries over almost line by line. We therefore leave this part to the reader. ∎

Now we will turn our attention to general time-frequency lattices. Here the situation is slightly more involved compared to $\ell^2(\mathbb{Z})$.

Let Λ be a lattice in $\mathbb{Z}_L \times \mathbb{Z}_L$ with canonical generator matrix A given by

$$A = \begin{bmatrix} \frac{L}{2p} & b \\ 0 & p \end{bmatrix}.$$

Then we choose $\alpha,\beta,m_0,n_0 \in \mathbb{Z}$ such that

$$\alpha\frac{L}{2p}m_0 + \alpha b n_0 + \beta p n_0 \tag{4.4}$$

attains its minimal positive value. Assume that there exists a choice of α,β,m_0,n_0 such that $(\frac{L}{2p}m_0 + bn_0)(pn_0) < 0$ and $(\alpha\frac{L}{2p})(\alpha b+\beta p) > 0$. In the other cases we have to change the signs of the later defined γ and δ accordingly. In the following we will restrict our analysis to the case where $|\alpha| = 1$. For the remainder of this section let α,β,m_0,n_0 be defined in this way. Now we regard α and β as elements of $\mathbb{Z}_L$. For the sake of brevity we set

$$c := \gcd(\alpha\tfrac{L}{2p}, \alpha b + \beta p), \quad d := \operatorname{lcm}(\alpha\tfrac{L}{2p}, \alpha b + \beta p),$$

and

$$s := \gcd(\tfrac{L}{2p}m_0 + bn_0, pn_0), \quad t := \operatorname{lcm}(\tfrac{L}{2p}m_0 + bn_0, pn_0).$$

The minimality condition for (4.4) shows that

$$\alpha\frac{L}{2p}m_0 + \alpha b n_0 + \beta p n_0 = c = s. \tag{4.5}$$

We further define $\gamma,\delta \in \mathbb{Z}_L$ by

$$\gamma := \frac{t}{\frac{L}{2p}m_0 + bn_0} \quad \text{and} \quad \delta := -\frac{t}{pn_0},$$

and $\sigma \in \operatorname{Hom}(\mathbb{Z}_L \times \mathbb{Z}_L)$ by

$$\sigma = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}.$$

PROPOSITION 4.6. *Let* Λ *be a lattice in* $\mathbb{Z}_L \times \mathbb{Z}_L$ *with canonical generator matrix* A *given by*

$$A = \begin{bmatrix} \frac{L}{2p} & b \\ 0 & p \end{bmatrix}.$$

Let σ be defined as in the preceding paragraph, and let U on the space of L-periodic functions on $\mathbb{Z}$ be defined by

$$Uf(k) = \sum_{l \in \mathbb{Z}_L} f(\alpha k + \beta l) e^{-\pi i(\alpha\gamma k^2 + \beta\delta l^2)(L+1)/L} e^{-2\pi i \beta\gamma k l/L}.$$

Then

$$g_{m\frac{L}{2p}+nb,np}(l) = C(m,n) U (U^{-1} g)_{\sigma(m\frac{L}{2p}+nb,np)}(l),$$

where $C(m,n) = e^{-\pi i(\alpha\gamma m^2+\beta\delta n^2)(L+1)/L} e^{-2\pi i\beta\gamma mn/L}$.

Before moving on to the proof of this statement we point out that the operator U in Proposition 4.6 is no longer a simple chirp operator as for the case $\ell^2(\mathbb{Z})$; cf. Proposition 3.6. This difference and the different form of σ necessitate a somewhat different proof for the case $\mathbb{C}^L$.

Proof. In order to apply Theorem 1.1 we need to check whether $\sigma = (\sigma_1, \sigma_2)$ is symplectic. For this, we have to show that, for all $(x,y),(x',y') \in \mathbb{Z}_L \times \mathbb{Z}_L$,

$$e^{-2\pi i\sigma_2(x,y)\sigma_1(x',y')/L} e^{2\pi i \sigma_2(x',y')\sigma_1(x,y)/L} = e^{-2\pi i x'y/L} e^{2\pi i xy'/L}. \tag{4.6}$$

We have

$$\begin{aligned} &e^{-2\pi i\sigma_2(x,y)\sigma_1(x',y')/L} e^{2\pi i \sigma_2(x',y')\sigma_1(x,y)/L} \\ &= e^{-2\pi i(\gamma x+\delta y)(\alpha x'+\beta y')/L} e^{2\pi i(\gamma x'+\delta y')(\alpha x+\beta y)/L} \\ &= e^{2\pi i(\alpha\delta-\beta\gamma)(xy'-x'y)/L} \end{aligned}$$

and

$$\begin{aligned} \alpha\delta - \beta\gamma &= -\frac{\alpha t}{pn_0} - \frac{\beta t}{\frac{L}{2p}m_0 + bn_0} \\ &= \frac{-t}{(\frac{L}{2p}m_0+bn_0)(pn_0)}(\alpha(\tfrac{L}{2p}m_0+bn_0)+\beta pn_0). \end{aligned}$$

By (4.5),

$$\alpha\delta - \beta\gamma = \frac{-st}{(\frac{L}{2p}m_0+bn_0)(pn_0)} = 1,$$

since $(\frac{L}{2p}m_0+bn_0)(pn_0) < 0$. This proves that (4.6) is satisfied, which shows that σ is indeed symplectic. Moreover, a short computation analogous to the one in the proof of Proposition 3.6 shows that

$$\zeta(k,l) = (\sigma^*\kappa_0\sigma - \kappa_0)(k,l) = (\alpha\gamma k + \beta\gamma l, \beta\gamma k + \beta\delta l).$$

Now it is easy to check (compare also [24, Example 1.1.34 (iii)]) that

$$\psi : \mathbb{Z}_L^2 \to \mathbb{T}, \quad \psi(k,l) = e^{\pi i(\alpha\gamma k^2+\beta\delta l^2)(L+1)/L} e^{2\pi i\beta\gamma kl/L}$$

is a second degree character associated to ζ. Applying Theorem 1.1 now finishes the proof. $\square$

As in the discrete case we need to define a special bijective map in order to give the definition of a Wilson basis associated with a lattice with arbitrary canonical generator matrix.

LEMMA 4.7. *Let* $\frac{L}{2p}, b \in \mathbb{Z}_L$ *with* $0 \leq b < \frac{L}{2p}$ *and let* $\alpha, \beta, m_0, n_0, c, d$ *be defined as before. Then the mapping* $\varphi : \mathbb{Z}^2 \to \mathbb{Z}^2$ *defined by*

$$\varphi(m,n) = \begin{cases} (m,n) & : \quad b = 0, \\ (mm_0 - \frac{2pd}{\alpha L} n, mn_0 + \frac{d}{\alpha b + \beta p} n) & : \quad b \neq 0 \end{cases}$$

is bijective and we have

$$\{(m \bmod \tfrac{L}{c}, n \bmod 2c) : (m,n) \in \varphi^{-1}(\{0, \dots, 2p-1\} \times \{0, \dots, \tfrac{L}{p} - 1\})\}$$

$$= \{0, \dots, \tfrac{L}{c} - 1\} \times \{0, \dots, 2c - 1\}.$$

Proof. The proof of this lemma is very similar to the proof of Lemma 3.7; we therefore omit it. □

Note that the following definition reduces to Definition 4.4 in the case of a diagonal canonical generator matrix.

DEFINITION 4.8. *Let* Λ *be a lattice in* $\mathbb{Z}_L \times \mathbb{Z}_L$ *with canonical generator matrix* A *given by*

$$A = \begin{bmatrix} \frac{L}{2p} & b \\ 0 & p \end{bmatrix}.$$

Let g *be some* L*-periodic function on* $\mathbb{Z}$*, and let* φ *be defined as in Lemma* 4.7*. Then the Wilson system* $\mathcal{W}(g, \Lambda, \mathbb{C}^L) = \{\psi_{m,n}\}_{(m,n) \in I}$*, where* $I = \{0, \dots, p-1\} \times \{0, \frac{L}{2p}\} \cup \{0, \dots, 2p-1\} \times \{1, \dots, \frac{L}{2p} - 1\}$*, is given by*

$$\psi^{\Lambda}_{m,n} = g_{\varphi_1(2m,n)\frac{L}{2p} + \varphi_2(2m,n)b, \varphi_2(2m,n)p} \qquad \text{if } m = 0, \dots, p-1, \ n = 0, \tfrac{L}{2p},$$

and for $m = 0, \dots, 2p-1$*,* $n = 1, \dots, \frac{L}{2p} - 1$*,*

$$\begin{aligned} \psi^{\Lambda}_{m,n} &= \tfrac{1}{\sqrt{2}}(g_{\varphi_1(m,n)\frac{L}{2p} + \varphi_2(m,n)b, \varphi_2(m,n)p} \\ &\quad + g_{\varphi_1(m,-n)\frac{L}{2p} + \varphi_2(m,-n)b, \varphi_2(m,-n)p}) \quad \text{if } m+n \text{ is even}, \\ \psi^{\Lambda}_{m,n} &= \tfrac{i}{\sqrt{2}}(g_{\varphi_1(m,n)\frac{L}{2p} + \varphi_2(m,n)b, \varphi_2(m,n)p} \\ &\quad - g_{\varphi_1(m,-n)\frac{L}{2p} + \varphi_2(m,-n)b, \varphi_2(m,-n)p}) \quad \text{if } m+n \text{ is odd}. \end{aligned}$$

The following theorem is the analogue to Theorem 3.9 for the space $\mathbb{C}^L$.

THEOREM 4.9. *Let* Λ *be a lattice in* $\mathbb{Z}_L \times \mathbb{Z}_L$ *with canonical generator matrix* A *given by*

$$A = \begin{bmatrix} \frac{L}{2p} & b \\ 0 & p \end{bmatrix}.$$

Let g *be some* L*-periodic function on* $\mathbb{Z}$ *such that* $\widehat{U^{-1}g}$ *is real-valued, let* $M := 2p$*,* $N := \frac{L}{p}$*,* $q := \frac{L}{2c}$*,* $\tilde{M} := 2q$*,* $\tilde{N} := \frac{L}{q}$*, and let* U *and* φ *be defined as in Proposition* 4.6 *and Lemma* 4.7*, respectively. Then the following conditions are equivalent.*

(i) $\{g_{m\frac{L}{2p}+nb,np}\}_{m=0,\dots,M-1,n=0,\dots,N-1}$ *is a tight frame for* $\mathbb{C}^L$ *with frame bound* 2.

(ii) $\{(U^{-1}g)_{m\frac{L}{2q},nq}\}_{m=0,\dots,\tilde{M}-1,n=0,\dots,\tilde{N}-1}$ *is a tight frame for* $\mathbb{C}^L$ *with frame bound* 2.

(iii) $\mathcal{W}(U^{-1}g,\frac{L}{2q}\mathbb{Z}_L\times q\mathbb{Z}_L,\mathbb{C}^L)$ *is an orthonormal basis for* $\mathbb{C}^L$.

(iv) $\mathcal{W}(g,\Lambda,\mathbb{C}^L)$ *is an orthonormal basis for* $\mathbb{C}^L$.

Proof. Let σ be defined as before and let $\varphi=(\varphi_1,\varphi_2)$. Then we compute

$$\begin{aligned}
&\sigma(\varphi_1(m,n)\tfrac{L}{2p}+\varphi_2(m,n)b,\varphi_2(m,n)p)\\
&=(\alpha(\varphi_1(m,n)\tfrac{L}{2p}+\varphi_2(m,n)b)+\beta\varphi_2(m,n)p,\gamma(\varphi_1(m,n)\tfrac{L}{2p}+\varphi_2(m,n)b)+\delta\varphi_2(m,n)p)\\
&=(m(\alpha\tfrac{L}{2p}m_0+\alpha bn_0+\beta pn_0)+n(-\alpha\tfrac{L}{2p}\tfrac{2pd}{\alpha L}+\alpha b\tfrac{d}{\alpha b+\beta p}+\beta p\tfrac{d}{\alpha b+\beta p}),\\
&\quad m(\gamma\tfrac{L}{2p}m_0+\gamma bn_0+\delta pn_0)+n(-\gamma\tfrac{L}{2p}\tfrac{2pd}{\alpha L}+\gamma b\tfrac{d}{\alpha b+\beta p}+\delta p\tfrac{d}{\alpha b+\beta p}))\\
&=(mc,m(\tfrac{t}{\frac{L}{2p}m_0+bn_0}(\tfrac{L}{2p}m_0+bn_0)-\tfrac{t}{pn_0}pn_0)\\
&\quad+n(-\tfrac{t}{\frac{L}{2p}m_0+bn_0}\tfrac{d}{\alpha}+\tfrac{t}{\frac{L}{2p}m_0+bn_0}\tfrac{bd}{\alpha b+\beta p}-\tfrac{t}{pn_0}\tfrac{pd}{\alpha b+\beta p}))\\
&=(mc,ntd(\tfrac{-\alpha bpn_0-\beta p^2n_0+\alpha bpn_0-\alpha p\frac{L}{2p}m_0-\alpha pbn_0}{\alpha(\frac{L}{2p}m_0+bn_0)(\alpha b+\beta p)pn_0}))\\
&=(mc,-ntdp\tfrac{\alpha\frac{L}{2p}m_0+\alpha bn_0+\beta pn_0}{\alpha(\frac{L}{2p}m_0+bn_0)(\alpha b+\beta p)pn_0})\\
&=(mc,-ntp\tfrac{cd}{\alpha(\frac{L}{2p}m_0+bn_0)(\alpha b+\beta p)pn_0})\\
&=(mc,-nt\tfrac{\frac{L}{2}}{(\frac{L}{2p}m_0+bn_0)pn_0})\\
&=(mc,n\tfrac{L}{2c}),
\end{aligned}$$

where in the last step we used (4.5) and $st=-(\frac{L}{2p}m_0+bn_0)pn_0$. Since $|\alpha|=1$, we have $c=s=\gcd(\alpha\frac{L}{2p},\alpha b+\beta p)$ is a factor of $\frac{L}{2p}$ and hence of $\frac{L}{2}$. Using Lemma 4.7, the equivalence of (i) and (ii) follows immediately from Proposition 4.6, since U is unitary and $|C(m,n)|=1$. Proposition 4.5 proves (ii) $\Leftrightarrow$ (iii). Therefore it remains to prove the equivalence of (iii) and (iv). For this, we will use the following implication of Proposition 4.6:

$$\begin{aligned}
U(U^{-1}g)_{m\frac{L}{2q},nq}&=U(U^{-1}g)_{\sigma(\varphi_1(m,n)\frac{L}{2p}+\varphi_2(m,n)b,\varphi_2(m,n)p)}\\
&=C(\varphi(m,n))^{-1}g_{\varphi_1(m,n)\frac{L}{2p}+\varphi_2(m,n)b,\varphi_2(m,n)p}.
\end{aligned}$$

An easy but tedious calculation shows that $C(\varphi(m,n))^{-1}$ does not depend on the sign of n. The definition of a Wilson basis, the fact that U is a unitary operator, and the fact that $|C(\varphi(m,n))|=1$ yields the result. ∎

Tight Gabor frames in $\mathbb{C}^L$ can be constructed in the same way as for $\ell^2(\mathbb{Z})$ and $L^2(\mathbb{R})$ by using the "inverse square root trick." Furthermore, it has been shown in [25] that for properly localized windows the dual window constructed in $\mathbb{C}^L$ by "sampling and periodization" of the frame $\{g_{ma,nb}\}$ converges to the dual window $S^{-1}g$ with increasing sampling rate and increasing periodization interval; see [25] for details. This result can be easily extended to tight windows. We refer also to [23, 30, 7] for related results and leave the details to the reader. To obtain tight Gabor frames in $\mathbb{C}^L$ that satisfy the required conditions of the theorem above and have good time-frequency localization, one can thus essentially proceed analogous to the example at the end of section 2.

5. Conclusion. We have demonstrated that orthonormal Wilson bases for $L^2(\mathbb{R})$ (with excellent time-frequency localization) can be constructed for general time-frequency lattices. Of course any numerical implementation has to be done in a discrete setting. Somewhat longer proofs establish a similar result for the spaces $\ell^2(\mathbb{Z})$ and $\mathbb{C}^L$ for nonrectangular time-frequency lattices. The approach based on metaplectic transforms used in this paper suggests that the main results can be extended to the setting of symplectic time-frequency lattices on general locally compact abelian groups.

Furthermore, our results imply that from a practical viewpoint it is indeed possible to extend OQAM-OFDM or cosine-modulated filter banks to general time-frequency lattices. Moreover, we expect that the benefits of using general time-frequency lattices will be even more pronounced for images and higher-dimensional signals. Our expectation is based on the fact that in the theory of sphere packings (and sphere coverings) the advantages of the optimal sphere packing over the packing associated with the rectangular lattices increases significantly with the dimension of the space [8].

An interesting research problem is thus to investigate how to extend the results in this paper to $L^2(\mathbb{R}^d)$ for nonsymplectic lattices as well as to find optimal time-frequency lattices in $\mathbb{R}^{2d}$ for $d > 1$. One possibility for defining an "optimal" time-frequency lattice is to fix the function g to be a Gaussian, say, and then find that time-frequency lattice of fixed density which minimizes the condition number of the associated Gabor frame operator as indicated in [31].

Acknowledgment. We thank the referees for valuable comments and suggestions which led to an improvement of the results and presentation in this paper.

REFERENCES

[1] L. Auslander and R. Tolimieri, *On finite Gabor expansions of signals*, in Signal Processing, Part I: Signal Processing Theory, L. Auslander, T. Kailath, and S.K. Mitter, eds., Springer-Verlag, New York, 1990, pp. 13–23.

[2] J.J. Benedetto, W. Czaja, and A.Y. Maltsev, *The Balian-Low theorem for the symplectic form on* $\mathbb{R}^{2d}$, J. Math. Phys., 44 (2003), pp. 1735–1750.

[3] J.J. Benedetto, C. Heil, and D.F. Walnut, *Differentiation and the Balian–Low theorem*, J. Fourier Anal. Appl., 1 (1995), pp. 355–403.

[4] H. Bölcskei, *Orthogonal frequency division multiplexing based on offset QAM*, in Advances in Gabor Analysis, H.G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Cambridge, MA, 2002, pp. 321–352.

[5] H. Bölcskei, H.G. Feichtinger, K. Gröchenig, and F. Hlawatsch, *Discrete-time Wilson expansions*, in Proc. IEEE Int. Sympos. Time-Frequency Time-Scale Analysis, Paris, 1996, pp. 525-528.

[6] H. Bölcskei and F. Hlawatsch, *Oversampled modulated filter banks*, in Gabor Analysis and Algorithms: Theory and Applications, H.G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Cambridge, MA, 1998, pp. 295–322.

[7] O. Christensen and T. Strohmer, *Methods for approximation of the inverse (Gabor) frame operator*, in Advances in Gabor Analysis, Birkhäuser Boston, Cambridge, MA, 2003, pp. 171–195.

[8] J.H. Conway and N.J.A. Sloane, eds., *Sphere Packings, Lattices and Groups*. Grundlehren Math. Wiss. 290, Springer-Verlag, Berlin, Heidelberg, 1993.

[9] I. Daubechies, S. Jaffard, and J.-L. Journé, *A simple Wilson orthonormal basis with exponential decay*, SIAM J. Math. Anal., 22 (1991), pp. 554–572.

[10] H.G. Feichtinger, O. Christensen, and T. Strohmer, *A group-theoretical approach to Gabor analysis*, Optical Engineering, 34 (1995), pp. 1697–1704.

[11] H.G. Feichtinger and W. Kozek, *Quantization of TF lattice-invariant operators on elementary LCA groups*, in Gabor Analysis and Algorithms: Theory and Applications, H.G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Cambridge, MA, 1998, pp. 233–266.

[12] H.G. Feichtinger and T. Strohmer, eds., *Gabor Analysis and Algorithms: Theory and Applications*, Birkhäuser Boston, Cambridge, MA, 1998.

[13] H.G. Feichtinger and T. Strohmer, eds., *Advances in Gabor analysis*, Birkhäuser Boston, Cambridge, MA, 2002.

[14] B. Le Floch, M. Alard, and C. Berrou, *Coded orthogonal frequency division multiplex*, Proc. IEEE, 83 (1995), pp. 982–996.

[15] G.B. Folland, *Harmonic Analysis in Phase Space*, Ann. of Math., Stud. 122, Princeton University Press, Princeton, NJ, 1989.

[16] G.B. Folland, *A Course in Abstract Harmonic Analysis*, CRC Press, Boca Raton, FL, 1995.

[17] K. Gröchenig, *Aspects of Gabor analysis on locally compact abelian groups*, in Gabor Analysis and Algorithms: Theory and Applications, H.G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Cambridge, MA, 1998, pp. 211–231.

[18] K. Gröchenig, *Foundations of Time-Frequency Analysis*, Birkhäuser Boston, Cambridge, MA, 2001.

[19] K. Gröchenig, D. Han, C. Heil, and G. Kutyniok, *The Balian–Low theorem for symplectic lattices in higher dimensions*, Appl. Comput. Harmon. Anal., 13 (2002), pp. 169–176.

[20] C. Hermite, *Extraits de lettres de M. Ch. Hermite à M. Jacobi sur differents objets de la théorie des nombres, Deuxième lettre*, Reine Angewandte Mathematik, 40 (1850), pp. 279–290.

[21] E. Hewitt and K.A. Ross, *Abstract Harmonic Analysis*, Springer-Verlag, New York, 1970.

[22] L.K. Hua, *Introduction to Number Theory*, Springer-Verlag, Berlin, 1982.

[23] A.J.E.M. Janssen, *From continuous to discrete Weyl-Heisenberg frames through sampling*, J. Fourier Anal. Appl., 3 (1997), pp. 583–596.

[24] N. Kaiblinger, *Metaplectic Representation, Eigenfunctions of Phase Space Shifts, and Gelfand–Shilov spaces for LCA Groups*, Ph.D. thesis, University of Vienna, Austria, 1999.

[25] N. Kaiblinger, *Approximation of the Fourier transform and the dual Gabor window*, J. Fourier Anal. Appl., 11 (2005), pp. 25–42.

[26] E. Kaniuth and G. Kutyniok, *Zeros of the Zak transform on locally compact abelian groups*, Proc. Amer. Math. Soc., 126 (1998), pp. 3561–3569.

[27] W. Kozek and A. Molisch, *Nonorthogonal pulseshapes for multicarrier communications in doubly dispersive channels*, IEEE J. Sel. Areas Comm., 16 (1998), pp. 1579–1589.

[28] G. Kutyniok, *Time-Frequency Analysis on Locally Compact Groups*, Ph.D. thesis, University of Paderborn, Germany, 2000.

[29] T. Strohmer, *Numerical algorithms for discrete Gabor expansions*, in Gabor Analysis and Algorithms: Theory and Applications, H.G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Cambridge, MA, 1998, pp. 267–294.

[30] T. Strohmer, *Approximation of dual Gabor frames, window decay, and wireless communications*, Appl. Comput. Harmon. Anal., 11 (2001), pp. 243–262.

[31] T. Strohmer and S. Beaver, *Optimal OFDM system design for time-frequency dispersive channels*, IEEE Trans. Comm., 51 (2003), pp. 1111–1122.

[32] R. Tolimieri and M. An, *Time-Frequency Representations*, Birkhäuser Boston, Cambridge, MA, 1998.

[33] K.G. Wilson, *Generalized Wannier Functions*, preprint, Cornell University, Ithaca, NY, 1987.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 712–736

A HIGHER ORDER ASYMPTOTIC PROBLEM RELATED TO PHASE TRANSITIONS*

ROGER MOSER†

Abstract. An asymptotic analysis of a family of functionals used in the van der Waals–Cahn–Hilliard theory of phase transitions gives rise to a generalized area functional in the limit. We examine a family of related higher order functionals on a three-dimensional domain. The expected limit in this case is a generalization of the Willmore functional. An analysis of the problem under a monotonicity assumption supports this conjecture.

Key words. phase transitions, Willmore functional, generalized surfaces

AMS subject classifications. 49Q20, 35J60, 80A22

DOI. 10.1137/040616760

1. Introduction.

Suppose $\Omega \subset \mathbb{R}^3$ is an open domain. For $\epsilon > 0$ and $u \in H^1_{\mathrm{loc}}(\Omega)$, consider the functional

$$E_\epsilon(u) = \int_\Omega \left(\frac{\epsilon}{2} |\nabla u|^2 + \frac{1}{4\epsilon}(1-u^2)^2 \right) dx.$$

These and similar functionals are used in the van der Waals–Cahn–Hilliard theory of phase transitions as a model for the energy of phase interfaces. In this context, but also as an independent problem, it is interesting to study the asymptotic behavior of E_ϵ as $\epsilon \searrow 0$. Such an analysis was first done by Modica and Mortola [14, 15] in the framework of Γ-convergence. Related results have been obtained by Modica [12, 13], Sternberg [22], and many others. The limit functional for this asymptotic problem turns out to be (up to a constant multiple) an area functional for generalized surfaces (or generalized submanifolds for similar problems in other dimensions) in Ω.

The asymptotic behavior of critical points of E_ϵ under a volume constraint—in other words, solutions of

$$\epsilon \Delta u + \frac{1}{\epsilon}(1-u^2)u = \lambda \quad \text{in } \Omega, \tag{1.1}$$

where λ is a constant—have been studied by Hutchinson and Tonegawa [9]. This theory again identifies the area functional as the limit of E_ϵ, but it also gives some information about the behavior of the left-hand side of (1.1) as $\epsilon \searrow 0$. In fact, the quantity

$$\tau_\epsilon(u) = \epsilon \Delta u + \frac{1}{\epsilon}(1-u^2)u$$

converges to the mean curvature of the limit surface in some sense, at least if solutions of (1.1) or similar equations are considered. The results have been extended by Tonegawa [23] to the case of sufficiently regular nonconstant right-hand sides in (1.1).

*Received by the editors October 11, 2004; accepted for publication (in revised form) April 6, 2005; published electronically November 9, 2005.
http://www.siam.org/journals/sima/37-3/61676.html
†Courant Institute, 251 Mercer Street, New York, NY 10012 (rmoser@cims.nyu.edu).

If the limit of τ_ϵ corresponds to the mean curvature, then one might expect that the functionals

$$T_\epsilon(u) = \frac{1}{4\epsilon}\int_\Omega (\tau_\epsilon(u))^2\,dx$$

for $u \in H^2_{\mathrm{loc}}(\Omega)$ converge to a generalized version of the Willmore functional. (For an immersed C^2-surface $\Sigma \subset \Omega$ with mean curvature vector H, the Willmore functional is defined by

$$W(\Sigma) = \frac{1}{4}\int_\Sigma |H|^2\,d\mathcal{H}^2,$$

where $\mathcal{H}^2$ is the two-dimensional Hausdorff measure; see Willmore [24] for a discussion.) We study the asymptotic behavior of T_ϵ as $\epsilon \searrow 0$ in this paper. Although we cannot answer the question implied above in full generality, our results indicate strongly that the Willmore functional is indeed the limit of T_ϵ.

This problem is related to a conjecture of De Giorgi [6] about the Γ-convergence of similar functionals. A problem of this type has also been studied in a paper by Bellettini and Mugnai [5] for two-dimensional domains and under the assumption of radial symmetry.

For technical reasons, we assume that $\Omega = \Omega' \times \mathbb{R}$ for our main result, where $\Omega' \subset \mathbb{R}^2$ is an open domain. We then work in the space of all functions $u \in H^2_{\mathrm{loc}}(\Omega)$ satisfying

$$\frac{\partial u}{\partial x^3} \geq 0.$$

This condition simplifies the analysis of the problem. Some of the tools we use also work without this restriction, but it appears that a satisfactory analysis of the full problem also requires some additional arguments.

Before we state our results, we introduce some notation and terminology. We denote the Lebesgue measure in Ω by $\mathcal{L}^3$ and the m-dimensional Hausdorff measure by $\mathcal{H}^m$. If we have a countably m-rectifiable set $\Sigma \subset \Omega$, then $T_x\Sigma$ denotes the approximate tangent space of Σ at x. This approximate tangent space exists $\mathcal{H}^m$-almost everywhere on Σ. If $\phi \in C^1(\Omega, \mathbb{R}^3)$, then $\operatorname{div}_\Sigma \phi$ denotes the divergence of ϕ with respect to Σ. That is, if we write proj_V for the (3×3)-matrix of the orthogonal projection onto a linear subspace V of $\mathbb{R}^3$, then

$$\operatorname{div}_\Sigma \phi(x) = \operatorname{trace}(\operatorname{proj}_{T_x\Sigma}\nabla\phi(x))$$

wherever the approximate tangent space exists. We also write $\operatorname{proj}_V^\perp$ for the matrix of the orthogonal projection onto the orthogonal complement of V.

THEOREM 1.1. *Suppose $\Omega = \Omega' \times \mathbb{R}$ for some open set $\Omega' \subset \mathbb{R}^2$. For $\epsilon > 0$, suppose $u_\epsilon \in H^2_{\mathrm{loc}}(\Omega)$ such that*

$$\frac{\partial u_\epsilon}{\partial x^3} \geq 0$$

$\mathcal{L}^3$-almost everywhere in Ω. If

$$\liminf_{\epsilon \searrow 0}(E_\epsilon(u_\epsilon) + T_\epsilon(u_\epsilon)) < \infty,$$

then there exists a sequence $\epsilon_k \searrow 0$ with the following properties.

(i) *There exist a relatively closed, countably* 2*-rectifiable set* $\Sigma \subset \Omega$ *and an upper semicontinuous function* $\theta : \Sigma \to [2\sqrt{2}/3, \infty)$ *such that*

$$\lim_{k\to\infty} \frac{\epsilon_k}{2} \int_\Omega \eta |\nabla u_{\epsilon_k}|^2 \, dx = \lim_{k\to\infty} \frac{1}{4\epsilon_k} \int_\Omega \eta (1-u_{\epsilon_k}^2)^2 \, dx = \frac{1}{2} \int_\Sigma \eta\theta \, d\mathcal{H}^2$$

for every $\eta \in C_0^0(\Omega)$.

(ii) *The sequence* $\{u_{\epsilon_k}\}$ *converges* $\mathcal{L}^3$*-almost everywhere to a function* $u : \Omega \to \{-1, 1\}$ *that is locally constant in* $\Omega\backslash\Sigma$.

(iii) *There exists an* $\mathcal{H}^2$*-measurable function* $H : \Sigma \to \mathbb{R}^3$ *with* $H(x) \perp T_x\Sigma$ *for* $\mathcal{H}^2$*-almost every* $x \in \Sigma$*, such that*

$$\int_\Sigma (\operatorname{div}_\Sigma \phi + \phi \cdot H)\theta \, d\mathcal{H}^2 = 0$$

for every $\phi \in C_0^1(\Omega, \mathbb{R}^3)$ *and*

$$\frac{1}{4} \int_\Sigma |H|^2 \theta \, d\mathcal{H}^2 \le \liminf_{k\to\infty} T_{\epsilon_k}(u_{\epsilon_k}).$$

The proof of Theorem 1.1 rests on two pillars: a theory of a generalized Willmore functional on the one hand, and estimates of the terms of E_ϵ on the other hand. The generalized Willmore functional we use gives in particular a direct link between the functionals T_ϵ and the classical Willmore functional. But to apply the theory successfully, it is important to show that the terms

$$\frac{\epsilon}{2}|\nabla u_\epsilon|^2$$

are bounded by the other contribution to $E_\epsilon(u_\epsilon)$,

$$\frac{1}{4\epsilon}(1-u_\epsilon^2)^2,$$

in an appropriate sense, at least in the limit. Such estimates are also important in the works of Hutchinson and Tonegawa [9], Tonegawa [23], and Bellettini and Mugnai [5]. Apart from these estimates, however, we use relatively little information about the structure of the problem. This means that our methods can also be applied to other problems. For instance, it is not difficult to see that similar results can be obtained if the potential given by the function $(1-u^2)^2$ is replaced by another two-well potential. (We leave it to the reader to verify the details.) Similar tools as we use here have also been applied to a problem related to harmonic maps in [17].

We conclude the introduction by pointing out that the Willmore functional can always be approximated by T_ϵ for sufficiently smooth surfaces. Suppose, e.g., that $U \subset \mathbb{R}^3$ is a bounded, open set with C^2-regular boundary. Choose a function $\chi \in C^2(\mathbb{R}^3)$ with $\chi \equiv 1$ on $\Sigma = \partial U$, such that the distance function $\operatorname{dist}(\,\cdot\,, \Sigma)$ is of class C^2 in a neighborhood of $\operatorname{supp}\chi$. Then the functions

$$u_\epsilon(x) = \begin{cases} \chi(x)\tanh\left(\frac{\operatorname{dist}(x,\Sigma)}{\sqrt{2}\epsilon}\right) & \text{if } x \notin U, \\ -\chi(x)\tanh\left(\frac{\operatorname{dist}(x,\Sigma)}{\sqrt{2}\epsilon}\right) & \text{if } x \in U \end{cases}$$

satisfy

$$\lim_{\epsilon\searrow 0} \frac{\epsilon}{2} \int_\Omega \eta |\nabla u_\epsilon|^2 \, dx = \lim_{\epsilon\searrow 0} \frac{1}{4\epsilon} \int_\Omega \eta(1-u_\epsilon^2)^2 \, dx = \frac{\sqrt{2}}{3} \int_\Sigma \eta \, d\mathcal{H}^2$$

for any $\eta \in C_0^0(\mathbb{R}^3)$ and

$$\frac{2\sqrt{2}}{3} W(\Sigma) = \lim_{\epsilon \searrow 0} T_\epsilon(u_\epsilon).$$

The proof consists of the obvious computations and the observation that the function $v(s) = \tanh(s/(\sqrt{2}\epsilon))$ solves

$$v'' + (1 - v^2)v = 0.$$

2. A generalization of the Willmore functional. We need a few tools from geometric measure theory, among them some concepts from [16, 18, 17] that can be regarded as variants of the notion of varifolds (cf. Simon [20] and Allard [1]). Similar tools have been developed by Ambrosio and Soner [4] and Lin [10]. Yet in none of these papers are they quite in the right form for the problem we study here; therefore we present the theory from the beginning. Because it might also be of independent interest, we consider arbitrary dimensions $n \geq 2$. In this section, we assume that Ω is an open set in $\mathbb{R}^n$.

Let $\mathcal{S}$ be the set of all symmetric, positive semidefinite real $(n \times n)$-matrices. We write $\mathcal{M}(\Omega)$ for the set of all pairs $M = (\mu, \nu)$, such that

- μ is a Radon measure on Ω (with nonnegative values),
- ν is a Radon measure on Ω with values in $\mathcal{S}$, and
- there exists a function $\sigma \in L^\infty(\mu, \mathcal{S})$ such that $\nu = \mu \llcorner \sigma$.

If $M = (\mu, \nu) \in \mathcal{M}(\Omega)$, we define the linear functional

$$\delta M(\phi) = \int_\Omega \left(\operatorname{div} \phi \, d\mu - \frac{\partial \phi^\alpha}{\partial x^\beta} \, d\nu_{\alpha\beta} \right)$$

on $C_0^1(\Omega, \mathbb{R}^n)$. Here and throughout the paper we use a summation convention: a repeated Greek index indicates a summation over its range.

Examples.

(i) Suppose $\Sigma \subset \Omega$ is a countably m-rectifiable set and

$$\theta \in L^1_{\mathrm{loc}}(\mathcal{H}^m \llcorner \Sigma, [0, \infty)).$$

Define $M = (\mu, \nu)$, where

$$\mu = (\mathcal{H}^m \llcorner \Sigma) \llcorner \theta \tag{2.1}$$

and

$$\nu = \mu \llcorner \sigma \tag{2.2}$$

for $\sigma(x) = \operatorname{proj}^\perp_{T_x\Sigma}$. Then

$$\delta M(\phi) = \int_\Sigma (\operatorname{div}_\Sigma \phi) \theta \, d\mathcal{H}^m.$$

(ii) If $\epsilon > 0$ and $u \in H^2_{\mathrm{loc}}(\Omega)$, define

$$\mu = \mathcal{L}^n \llcorner \left(\frac{\epsilon}{2} |\nabla u|^2 + \frac{1}{4\epsilon}(1 - u^2)^2 \right) \tag{2.3}$$

(where $\mathcal{L}^n$ is the Lebesgue measure in Ω), and

$$\nu_{\alpha\beta} = \epsilon\mathcal{L}^n \llcorner \left(\frac{\partial u}{\partial x^\alpha}\frac{\partial u}{\partial x^\beta}\right). \tag{2.4}$$

Then the pair $M = (\mu, \nu)$ satisfies

$$\delta M(\phi) = \int_\Omega \phi \cdot \nabla u\, \tau_\epsilon(u)\, dx.$$

(iii) Other examples of possible applications are given in [16, 18, 17].

The examples indicate that the space $\mathcal{M}(\Omega)$ can provide a connection between the functionals E_ϵ and the (generalized) area functional. The following notion also gives a relation between T_ϵ and the Willmore functional.

Definition 2.1. *The functional W on $\mathcal{M}(\Omega)$ given by*

$$W(M) = \frac{1}{4}\sup\left\{(\delta M(\phi))^2 :\ \phi \in C_0^1(\Omega, \mathbb{R}^n) \text{ with } \int_\Omega \phi^\alpha\phi^\beta\, d\nu_{\alpha\beta} \le 1\right\}$$

is called the generalized Willmore functional*. We write*

$$\mathcal{W}(\Omega) = \{M \in \mathcal{M}(\Omega) :\ W(M) < \infty\}.$$

Examples.

(i) Suppose $\Sigma \subset \Omega$ is countably m-rectifiable and $\theta \in L^1_{\mathrm{loc}}(\mathcal{H}^m \llcorner \Sigma, [0,\infty))$, such that there exists an $\mathcal{H}^m$-measurable function $H : \Sigma \to \mathbb{R}^n$ with

$$H(x) \perp T_x\Sigma \tag{2.5}$$

for $\mathcal{H}^m$-almost every $x \in \Sigma$ and

$$\int_\Sigma (\operatorname{div}_\Sigma \phi + \phi \cdot H)\theta\, d\mathcal{H}^m = 0$$

for every $\phi \in C_0^1(\Omega, \mathbb{R}^n)$. If $M = (\mu, \nu)$ is defined by (2.1) and (2.2), then

$$W(M) = \frac{1}{4}\int_\Sigma |H|^2\theta\, d\mathcal{H}^m.$$

But if (2.5) is not satisfied $\mathcal{H}^m$-almost everywhere, then $W(M) = \infty$.

(ii) If $\epsilon > 0$ and $u \in H^2_{\mathrm{loc}}(\Omega)$, and if $M = (\mu, \nu)$ is given by (2.3) and (2.4), then

$$\begin{aligned}(\delta M(\phi))^2 &= \left(\int_\Omega \phi \cdot \nabla u\, \tau_\epsilon(u)\, dx\right)^2 \\ &\le \left(\int_\Omega (\tau_\epsilon(u))^2\, dx\right)\left(\int_\Omega (\phi \cdot \nabla u)^2\, dx\right) = 4T_\epsilon(u)\int_\Omega \phi^\alpha\phi^\beta\, d\nu_{\alpha\beta}.\end{aligned}$$

Thus $W(M) \le T_\epsilon(u)$.

In the rest of the section we analyze measure pairs in $\mathcal{W}(\Omega)$. We first show that a representation of δM by an $\mathbb{R}^n$-valued function H as in the first example always exists for $M \in \mathcal{W}(\Omega)$.

Proposition 2.2. *Suppose that $M = (\mu, \nu) \in \mathcal{W}(\Omega)$, and let $\sigma \in L^\infty(\mu, \mathcal{S})$ be the function such that $\nu = \mu \llcorner \sigma$. Then there exists a μ-measurable function $H = (H^1, \ldots, H^n) : \Omega \to \mathbb{R}^n$ with $H^\beta \sigma_{\alpha\beta} \in L^1_{\mathrm{loc}}(\mu)$ for $\alpha = 1, \ldots, n$, such that*

$$\delta M(\phi) = -\int_\Omega \phi^\alpha H^\beta \, d\nu_{\alpha\beta} \tag{2.6}$$

for every $\phi \in C^1_0(\Omega, \mathbb{R}^n)$ and

$$H(x) \perp \ker \sigma(x) \tag{2.7}$$

for μ-almost every $x \in \Omega$. Moreover,

$$W(M) = \frac{1}{4} \int_\Omega H^\alpha H^\beta \, d\nu_{\alpha\beta}. \tag{2.8}$$

Any other function satisfying (2.6) *and* (2.7) *agrees with H except possibly on a μ-null set.*

Proof. The functional δM on $C^1_0(\Omega, \mathbb{R}^n)$ satisfies

$$\begin{aligned} |\delta M(\phi)| &\le 2 \left(W(M) \int_\Omega \phi^\alpha \phi^\beta \, d\nu_{\alpha\beta} \right)^{1/2} \\ &\le 2 \|\phi\|_{C^0(\overline{\Omega})} (W(M) \|\sigma\|_{L^\infty(\mu)} \mu(\operatorname{supp} \phi))^{1/2}. \end{aligned} \tag{2.9}$$

Hence there exists a continuous extension to $C^0_0(\Omega, \mathbb{R}^n)$, represented by an $\mathbb{R}^n$-valued Radon measure h on Ω. Moreover, this measure is absolutely continuous with respect to μ. By the Radon–Nikodým theorem there exists a unique function $\hat{H} \in L^1_{\mathrm{loc}}(\mu, \mathbb{R}^n)$ such that $h = \mu \llcorner \hat{H}$. In other words, we have

$$\delta M(\phi) = \int_\Omega \phi \cdot \hat{H} \, d\mu$$

for any $\phi \in C^1_0(\Omega, \mathbb{R}^n)$. Now we consider the first inequality in (2.9) again. Any $\psi \in L^\infty(\mu, \mathbb{R}^n) \cap L^2(\mu, \mathbb{R}^n)$ with

$$\int_\Omega \psi^\alpha \psi^\beta \, d\nu_{\alpha\beta} = 0$$

can be approximated by a sequence of vector fields $\psi_i \in C^1_0(\Omega, \mathbb{R}^n)$ such that, as $i \to \infty$, $\|\psi_i - \psi\|_{L^2(\mu)} \to 0$ (because μ is a Radon measure). Hence

$$\left| \int_\Omega \psi \cdot \hat{H} \, d\mu \right| = \lim_{i \to \infty} \left| \int_\Omega \psi_i \cdot \hat{H} \, d\mu \right| \le 2\sqrt{W(M)} \limsup_{i \to \infty} \left(\int_\Omega \psi_i^\alpha \psi_i^\beta \, d\nu_{\alpha\beta} \right)^{1/2} = 0.$$

It follows that $\hat{H}(x) \perp \ker \sigma(x)$ for μ-almost every $x \in \Omega$. In particular there is a function $H : \Omega \to \mathbb{R}^n$, unique up to a μ-null set, such that (2.7) holds and

$$X \cdot \hat{H}(x) = -X^\alpha H^\beta(x) \sigma_{\alpha\beta}(x)$$

for μ-almost every $x \in \Omega$ and every $X \in \mathbb{R}^n$. This H satisfies (2.6). Only (2.8) remains to be proved.

The inequality

$$W(M) \leq \frac{1}{4} \int_\Omega H^\alpha H^\beta \, d\nu_{\alpha\beta}$$

follows directly from the definition of the functional W. To show the converse inequality, we approximate H by a sequence of functions $H_i \in C_0^1(\Omega, \mathbb{R}^n)$ such that

$$\lim_{i\to\infty} \int_\Omega (H^\alpha - H_i^\alpha)(H^\beta - H_i^\beta) \, d\nu_{\alpha\beta} = 0.$$

Then we have

$$W(M) \geq \frac{\left(\int_\Omega H^\alpha H_i^\beta \, d\nu_{\alpha\beta}\right)^2}{4\int_\Omega H_i^\alpha H_i^\beta \, d\nu_{\alpha\beta}}$$

for every i (provided that the denominator on the right-hand side does not vanish). For $i \to \infty$, we obtain the desired inequality. □

DEFINITION 2.3. *If $M = (\mu, \nu)$ is in $\mathcal{W}(\Omega)$ and $H : \Omega \to \mathbb{R}^n$ satisfies* (2.6) *and* (2.7)*, then H is called the* generalized mean curvature *of M. For any Borel set $A \subset \Omega$, we write*

$$W(M; A) = \frac{1}{4} \int_A H^\alpha H^\beta \, d\nu_{\alpha\beta}.$$

DEFINITION 2.4. *Suppose $M^{(k)} = (\mu^{(k)}, \nu^{(k)}) \in \mathcal{W}(\Omega)$ for $k \in \mathbb{N}$, and $M = (\mu, \nu) \in \mathcal{W}(\Omega)$. We say that $M^{(k)}$* converges weakly *to M and we write $M^{(k)} \rightharpoonup M$ if*

$$\sup_{k\in\mathbb{N}} W(M^{(k)}) < \infty$$

and, for every $\eta \in C_0^0(\Omega)$,

$$\int_\Omega \eta \, d\mu = \lim_{k\to\infty} \int_\Omega \eta \, d\mu^{(k)} \quad \textit{and} \quad \int_\Omega \eta \, d\nu = \lim_{k\to\infty} \int_\Omega \eta \, d\nu^{(k)}.$$

We say that $M^{(k)}$ converges strongly *to M and we write $M^{(k)} \to M$ if $M^{(k)} \rightharpoonup M$ and*

$$W(M) \geq \limsup_{k\to\infty} W(M^{(k)}).$$

Remark. The following is an immediate consequence of the definition of W: if $M^{(k)} \rightharpoonup M$, then

$$W(M) \leq \liminf_{k\to\infty} W(M^{(k)}).$$

Thus if $M^{(k)} \to M$, then

$$W(M) = \lim_{k\to\infty} W(M^{(k)}).$$

The space $\mathcal{W}(\Omega)$ is quite large. For instance, it contains measure pairs corresponding to countably rectifiable sets in Ω of any dimension. For some of the results in this section we have to consider a smaller subspace. We write $\mathcal{W}_m(\Omega)$ for the set of all pairs $M = (\mu, \nu) \in \mathcal{W}(\Omega)$ that satisfy

$$\operatorname{trace} \nu \leq (n-m)\mu. \tag{2.10}$$

In order to obtain some idea of what this condition means, consider an M that belongs to a rectifiable set Σ as in the examples above. Then (2.10) is satisfied if and only if the dimension of Σ is at least m.

For $M = (\mu, \nu) \in \mathcal{W}_m(\Omega)$, in particular if $m = 2$, we want to examine the behavior of quantities of the form $r^{-m}\mu(B_r(x_0))$ as $r \searrow 0$, where $B_r(x_0)$ denotes the open ball in $\mathbb{R}^n$ of center x_0 and radius r. The following lemma gives an inequality which is useful for that purpose. It is a variant of a result that is well known for many related theories. (Cf. Allard [1], Price [19], and Simon [21]. There are many other papers with results of this kind.)

LEMMA 2.5. *Suppose $M = (\mu, \nu) \in \mathcal{W}_m(\Omega)$ has the generalized mean curvature H. For $x_0 \in \Omega$ and $0 < \rho < \operatorname{dist}(x_0, \partial\Omega)$, set*

$$\Phi_M(x_0, \rho) = \rho^{-m}\left(\mu(B_\rho(x_0)) + \frac{1}{m}\int_{B_\rho(x_0)} (x^\alpha - x_0^\alpha)H^\beta(x)\, d\nu_{\alpha\beta}(x)\right).$$

If $0 < s \leq r < \operatorname{dist}(x_0, \partial\Omega)$, then

$$\Phi_M(x_0, s) \leq \Phi_M(x_0, r) + I,$$

where

$$I = \int_{B_r(x_0)\setminus B_s(x_0)} \left(\frac{(x^\alpha - x_0^\alpha)(x^\beta - x_0^\beta)}{|x - x_0|^{m+2}} + \frac{(x^\alpha - x_0^\alpha)H^\beta(x)}{m|x - x_0|^m}\right) d\nu_{\alpha\beta}(x).$$

Proof. We assume for simplicity that $x_0 = 0$. Suppose $\psi \in C_0^1(B_1(0))$ is a nonnegative function. For $\rho > 0$, set $\psi_\rho(x) = \psi(x/\rho)$. Insert now $\phi(x) = \psi_\rho(x)x$ into (2.6). The resulting equation is

$$\begin{aligned}\int_\Omega \psi_\rho(n\, d\mu - \operatorname{trace} d\nu) + \frac{1}{\rho}\int_\Omega x \cdot \nabla\psi(x/\rho)\, d\mu(x)&\\ - \int_\Omega \left(\frac{x^\alpha}{\rho}\frac{\partial\psi}{\partial x^\beta}(x/\rho) - x^\alpha H^\beta(x)\psi_\rho(x)\right) d\nu_{\alpha\beta}(x) &= 0.\end{aligned}$$

Thus (2.10) implies

$$\begin{aligned}m\int_\Omega \psi_\rho d\mu + \frac{1}{\rho}\int_\Omega x \cdot \nabla\psi(x/\rho)\, d\mu(x)&\\ - \int_\Omega \left(\frac{x^\alpha}{\rho}\frac{\partial\psi}{\partial x^\beta}(x/\rho) - x^\alpha H^\beta(x)\psi_\rho(x)\right) d\nu_{\alpha\beta}(x) &\leq 0.\end{aligned}$$

It follows that

$$\begin{aligned}\frac{d}{d\rho}\left(\rho^{-m}\int_\Omega \psi_\rho\, d\mu + \frac{1}{m}\int_\Omega \psi_\rho(x)x^\alpha H^\beta(x)\, d\nu_{\alpha\beta}(x)\right)&\\ \geq -\rho^{-m-2}\int_\Omega \left(x^\alpha\frac{\partial\psi}{\partial x^\beta}(x/\rho) + \frac{1}{m}x \cdot \nabla\psi(x/\rho)x^\alpha H^\beta(x)\right) d\nu_{\alpha\beta}(x).&\end{aligned} \tag{2.11}$$

If ψ is of the form $\psi(x) = \eta(|x|)$ for some $\eta \in C^1(\mathbb{R})$ with $\eta \equiv 0$ in $[1, \infty)$, the right-hand side of (2.11) is

$$-\rho^{-m-2} \int_\Omega \eta'(|x|/\rho) \left(\frac{x^\alpha x^\beta}{|x|} + \frac{1}{m} |x| x^\alpha H^\beta(x) \right) d\nu_{\alpha\beta}(x).$$

For

$$\Phi_\psi(\rho) = \rho^{-m} \left(\int_{B_\rho(x_0)} \psi_\rho d\mu + \frac{1}{m} \int_{B_\rho(x_0)} \psi_\rho x^\alpha H^\beta(x)\, d\nu_{\alpha\beta}(x) \right),$$

it follows that

$$\begin{aligned}\Phi_\psi(r) - \Phi_\psi(s) &\geq - \int_s^r \rho^{-m-2} \int_\Omega \eta'(|x|/\rho) \left(\frac{x^\alpha x^\beta}{|x|} + \frac{1}{m} |x| x^\alpha H^\beta(x) \right) d\nu_{\alpha\beta}(x)\, d\rho \\ &= \int_\Omega \int_s^r \rho^{-m} \frac{d}{d\rho} \eta(|x|/\rho)\, d\rho \left(\frac{x^\alpha x^\beta}{|x|^2} + \frac{1}{m} x^\alpha H^\beta(x) \right) d\nu_{\alpha\beta}(x).\end{aligned}$$

We have

$$\int_s^r \rho^{-m} \frac{d}{d\rho} \eta(|x|/\rho)\, d\rho = r^{-m} \eta(|x|/r) - s^{-m} \eta(|x|/s) + m \int_s^r \rho^{-m-1} \eta(|x|/\rho)\, d\rho. \tag{2.12}$$

If we approximate the characteristic function of the interval $(-\infty, 1)$ by a sequence of C^1-functions and insert these approximations instead of η into (2.12), the right-hand side converges to 0 if $|x| \geq r$ or $|x| < s$ and to

$$r^{-m} + m \int_{|x|}^r \rho^{-m-1}\, d\rho = |x|^{-m}$$

if $|x| \in [s, r)$. The claim of the lemma now follows from Lebesgue's convergence theorem. □

LEMMA 2.6. *Suppose $M = (\mu, \nu) \in \mathcal{W}_2(\Omega)$ and $\sigma \in L^\infty(\mu, \mathcal{S})$ such that $\nu = \mu \llcorner \sigma$. Then for any $\delta > 0$ and for any pair of concentric balls $B_s(x_0) \subset B_r(x_0) \subset \Omega$, the inequality*

$$(1 - \delta) s^{-2} \mu(B_s(x_0)) \leq (1 + \delta) r^{-2} \mu(B_r(x_0)) + \left(\frac{1}{4} + \frac{\|\sigma\|_{L^\infty(\mu)}}{2\delta} \right) W(M; B_r(x_0))$$

holds.

Proof. We use Lemma 2.5 and the estimates

$$\begin{aligned}&\frac{1}{2\rho^2} \left| \int_{B_\rho(x_0)} (x^\alpha - x_0^\alpha) H^\beta(x)\, d\nu_{\alpha\beta}(x) \right| \\ &\qquad \leq \frac{1}{2\rho^2} \left(\int_{B_\rho(x_0)} (x^\alpha - x_0^\alpha)(x^\beta - x_0^\beta)\, d\nu_{\alpha\beta}(x) \right)^{1/2} \left(\int_{B_\rho(x_0)} H^\alpha H^\beta\, d\nu_{\alpha\beta} \right)^{1/2} \\ &\qquad \leq \delta \rho^{-2} \mu(B_\rho(x_0)) + \frac{\|\sigma\|_{L^\infty(\mu)}}{4\delta} W(M; B_\rho(x_0))\end{aligned}$$

and

$$\begin{aligned}&\left| \int_{B_r(x_0) \setminus B_s(x_0)} \frac{(x^\alpha - x_0^\alpha) H^\beta(x)}{2|x - x_0|^2}\, d\nu_{\alpha\beta}(x) \right| \\ &\qquad \leq \int_{B_r(x_0) \setminus B_s(x_0)} \frac{(x^\alpha - x_0^\alpha)(x^\beta - x_0^\beta)}{|x - x_0|^4}\, d\nu_{\alpha\beta}(x) + \frac{1}{16} \int_{B_r(x_0) \setminus B_s(x_0)} H^\alpha H^\beta\, d\nu_{\alpha\beta}.\end{aligned}$$

It now suffices to combine the inequalities. ◻

If $M = (\mu, \nu) \in \mathcal{W}_2(\Omega)$, we conclude from Lemma 2.6 that $\mu(\{x\}) = 0$ for every $x \in \Omega$. Hence

$$\lim_{r \searrow 0} W(M; B_r(x)) = 0.$$

Thus it also follows from Lemma 2.6 that the 2-density

$$\Theta_M(x) = \lim_{r \searrow 0} r^{-2} \mu(B_r(x)) \tag{2.13}$$

exists for every $x \in \Omega$. Moreover, the function Θ_M defined by (2.13) has the following property.

LEMMA 2.7. *Let $M^{(k)} = (\mu^{(k)}, \nu^{(k)}) \in \mathcal{W}_2(\Omega)$ for $k \in \mathbb{N}$, and $M = (\mu, \nu) \in \mathcal{W}_2(\Omega)$ such that $M^{(k)} \to M$. Suppose $\sigma^{(k)} \in L^\infty(\mu^{(k)}, \mathcal{S})$ are the functions such that $\nu^{(k)} = \mu^{(k)} \llcorner \sigma^{(k)}$. Let $x_k \in \Omega$ with $x_k \to x_0 \in \Omega$ as $k \to \infty$. If*

$$\sup_{k \in \mathbb{N}} \|\sigma^{(k)}\|_{L^\infty(\mu^{(k)})} < \infty,$$

then

$$\Theta_M(x_0) \geq \limsup_{k \to \infty} \Theta_{M^{(k)}}(x_k).$$

Remark. It follows in particular that Θ_M is upper semicontinuous.

Proof. Let $\epsilon > 0$. There exists a radius $r > 0$ such that

$$\Theta_M(x_0) \geq r^{-2} \mu(\overline{B_r(x_0)}) - \epsilon$$

and

$$W(M; \overline{B_r(x_0)}) \leq \epsilon.$$

Hence

$$\Theta_M(x_0) \geq r^{-2} \limsup_{k \to \infty} \mu^{(k)}(B_r(x_0)) - \epsilon$$

and

$$\limsup_{k \to \infty} W(M^{(k)}; B_r(x_0)) \leq \epsilon.$$

Now Lemma 2.6 implies that

$$\Theta_{M^{(k)}}(x_0) \leq (1+\delta) r^{-2} \mu^{(k)}(B_r(x_0)) + C\left(1 + \frac{1}{\delta}\right)\epsilon$$

for any $\delta \in (0, 1]$ and any sufficiently large k, where C is a number that depends only on

$$\sup_{k \in \mathbb{N}} \|\sigma^{(k)}\|_{L^\infty(\mu^{(k)})}.$$

Thus

$$\limsup_{k \to \infty} \Theta_{M^{(k)}}(x_0) \leq (1+\delta)(\Theta_M(x_0) + \epsilon) + C\left(1 + \frac{1}{\delta}\right)\epsilon.$$

Since ϵ and δ are arbitrary, the claim follows. ◻

For a fixed $M = (\mu, \nu) \in \mathcal{W}_2(\Omega)$ and a fixed point $x_0 \in \Omega$, we now consider the rescaled measure pairs $M^{x_0,r} = (\mu^{x_0,r}, \nu^{x_0,r}) \in \mathcal{W}_2(\Omega^{x_0,r})$, where $\Omega^{x_0,r} = r^{-1}(\Omega - x_0)$, defined by

$$\int_{\Omega^{x_0,r}} \eta \, d\mu^{x_0,r} = r^{-2} \int_{\Omega} \eta((x - x_0)/r) \, d\mu(x)$$

and

$$\int_{\Omega^{x_0,r}} \eta \, d\nu^{x_0,r} = r^{-2} \int_{\Omega} \eta((x - x_0)/r) \, d\nu(x)$$

for all $\eta \in C_0^0(\Omega^{x_0,r})$. We have

$$\lim_{r \searrow 0} \mu^{x_0,r}(B_R(0)) = \lim_{r \searrow 0} r^{-2} \mu(B_{Rr}(x_0)) = R^2 \Theta_M(x_0)$$

for almost any $R > 0$. If $\phi \in C_0^1(\mathbb{R}^n, \mathbb{R}^n)$, then

$$\delta M^{x_0,r}(\phi) = r^{-1} \delta M(\phi((x - x_0)/r))$$

for any r that is large enough so that $\operatorname{supp} \phi \subset \Omega^{x_0,r}$. If $R > 0$ is such that $\operatorname{supp} \phi \in B_R(0)$, then it follows that

$$|\delta M^{x_0,r}(\phi)| \leq 2 \sup_{\mathbb{R}^n} |\phi| \left(\|\sigma\|_{L^\infty(\mu)} r^{-2} \mu(B_{Rr}(x_0)) W(M; B_{Rr}(x_0)) \right)^{1/2} \to 0$$

as $r \searrow 0$ (where $\sigma \in L^\infty(\mu, \mathcal{S})$ is such that $\nu = \mu \llcorner \sigma$).

We infer that we can choose a sequence $r_k \searrow 0$ such that $M^{x_0,r_k} \to M^*$ for some $M^* = (\mu^*, \nu^*) \in \mathcal{W}_2(\mathbb{R}^n)$ with $\delta M^* = 0$ (in the sense that the restrictions of M^{x_0,r_k} to any bounded domain in $\mathbb{R}^n$ converge strongly to the corresponding restrictions of M^*). Moreover, any given sequence $r_k \searrow 0$ has a subsequence with this property.

Definition 2.8. *Any measure pair $M^* \in \mathcal{W}_2(\Omega)$ obtained by a limit process as described above is called a* tangent measure pair *of M at x_0.*

Lemma 2.9. *Suppose $M^* = (\mu^*, \nu^*)$ is a tangent measure pair of $M = (\mu, \nu) \in \mathcal{W}_2(\Omega)$ at $x_0 \in \Omega$. Then $r^{-2}\mu^*(B_r(0)) = \Theta_M(x_0)$ for any $r > 0$.*

Proof. For all but countably many $r > 0$, we have

$$\mu^*(B_r(0)) = \lim_{k \to \infty} \mu^{x_0,r_k}(B_r(0)) = \lim_{k \to \infty} r_k^{-2} \mu(B_{rr_k}(x_0)) = r^2 \Theta_M(x_0).$$

By Lemma 2.5, the quantity $r^{-2}\mu^*(B_r(0))$ is nondecreasing; hence the identity holds even for all $r > 0$. □

Lemma 2.10. *Suppose $M^* = (\mu^*, \nu^*) \in \mathcal{W}_2(\mathbb{R}^n)$ satisfies $\delta M^* = 0$ and*

$$r^{-2}\mu^*(B_r(0)) = s^{-2}\mu^*(B_s(0)) \tag{2.14}$$

for every pair of positive numbers r, s. Then the set

$$S(M^*) = \{x \in \mathbb{R}^n : \Theta_{M^*}(x) \geq \Theta_{M^*}(0)\}$$

is a linear subspace of $\mathbb{R}^n$. If $\mu^* \neq 0$, then $\dim S(M^*) \leq 2$, and the following three conditions are equivalent:

(i) $\dim S(M^*) = 2$,

(ii) $\mu^*(\mathbb{R}^n \backslash S(M^*)) = 0$,

(iii) $\mu^* = \pi^{-1}\Theta_{M^*}(0)\mathcal{H}^2 \llcorner S(M^*)$ *and* $\nu^* = \operatorname{proj}^{\perp}_{S(M^*)} \mu^*$.

Proof. We may assume $\mu^* \neq 0$. Let $\sigma^* \in L^\infty(\mu^*, \mathcal{S})$ be the function such that $\nu^* = \mu^* \llcorner \sigma^*$. Then (2.14) and Lemma 2.5 imply that $x \in \ker \sigma^*(x)$ for μ^*-almost every $x \in \mathbb{R}^n$. If $x_0 \in \mathbb{R}^n$ and $r > 0$, Lemma 2.5 also implies

$$\Theta_{M^*}(x_0) + \int_{B_r(x_0)} \frac{(x^\alpha - x_0^\alpha)(x^\beta - x_0^\beta)}{|x - x_0|^4}\, d\nu^*_{\alpha\beta}(x) \\ \leq r^{-2}\mu^*(B_r(x_0)) \leq r^{-2}\mu^*(B_{r+|x_0|}(0)) = \left(1 + \frac{|x_0|}{r}\right)^{-2} \Theta_{M^*}(0).$$

Letting $r \to \infty$, we see that $\Theta_{M^*}(x_0) \leq \Theta_{M^*}(0)$, and $\Theta_{M^*}(x_0) = \Theta_{M^*}(0)$ only if $x - x_0 \in \ker \sigma^*(x)$ for μ^*-almost every $x \in \mathbb{R}^n$.

Let $x_0 \in S(M^*)$. Then it follows that $x_0 \in \ker \sigma^*(x)$ for μ^*-almost every $x \in \mathbb{R}^n$. Let $\eta \in C_0^1(\mathbb{R}^n)$; then

$$\frac{d}{dt}\int_{\mathbb{R}^n} \eta(x + tx_0)\, d\mu^*(x) = \int_{\mathbb{R}^n} x_0 \cdot \nabla\eta(x + tx_0)\, d\mu^*(x) \\ = \int_{\mathbb{R}^n} x_0^\alpha \frac{\partial \eta}{\partial x^\beta}(x + tx_0)\, d\nu^*_{\alpha\beta}(x) = 0,$$

because $\delta M^* = 0$. That is, the measure μ^* is invariant under translations in the direction of x_0. It follows that $S(M^*)$ is a linear subspace of $\mathbb{R}^n$. Moreover, this translation invariance and (2.14) imply $\dim S(M^*) \leq 2$, and they also imply the equivalence of (i) and (ii). Obviously, (iii) implies (ii). Thus it remains to show that (ii) implies (iii).

For the rest of the proof, we assume that $S(M^*) = \mathbb{R}^2 \times \{0\}$ for simplicity. We also assume that $\mu^*(\mathbb{R}^n \backslash S(M^*)) = 0$. We already know that $\nu^*_{1\alpha} = \nu^*_{2\alpha} = 0$ for $\alpha = 1, \ldots, n$. Now for $\xi \in C_0^1(\mathbb{R}^2)$ and $\zeta \in C_0^1(\mathbb{R}^{n-2})$, consider a vector field of the form

$$\phi(x^1, \ldots, x^n) = \xi(x^1, x^2)\zeta(x^3, \ldots, x^n)e_\alpha,$$

where $\alpha \geq 3$ and e_α is the αth standard unit vector in $\mathbb{R}^n$. Because $\delta M^* = 0$, we have

$$\int_{\mathbb{R}^2 \times \{0\}} \xi(x^1, x^2) \left(\frac{\partial \zeta}{\partial x^\alpha}(0)\, d\mu^*(x) - \frac{\partial \zeta}{\partial x^\beta}(0)\, d\nu^*_{\alpha\beta}(x)\right) = 0.$$

If we choose ζ such that $(0, 0, \frac{\partial \zeta}{\partial x^3}, \ldots, \frac{\partial \zeta}{\partial x^n}) = e_\beta$, it follows that $\nu^*_{\alpha\beta} = 0$ for $\alpha \neq \beta$ and

$$\nu^*_{33} = \cdots = \nu^*_{nn} = \mu^*.$$

This shows (iii) and concludes the proof. $\square$

PROPOSITION 2.11. *If $M = (\mu, \nu) \in \mathcal{W}_2(\Omega)$, then the set*

$$\Sigma = \{x \in \Omega \colon\ \Theta_M(x) > 0\}$$

is countably 2-rectifiable. If $\sigma \in L^\infty(\mu, \mathcal{S})$ is such that $\nu = \mu \llcorner \sigma$, then $\sigma(x) = \mathrm{proj}^\perp_{T_x\Sigma}(x)$ for μ-almost every $x \in \Sigma$.

Proof. Let Σ_0 be the set of all points $x \in \Sigma$ such that σ and Θ_M are approximately continuous at x with respect to μ. Then $\mu(\Sigma \backslash \Sigma_0) = 0$. Suppose $x_0 \in \Sigma_0$. Let

$M^* = (\mu^*, \nu^*)$ be a tangent measure pair of M at x_0. From Lemmas 2.7 and 2.9 it follows that

$$\Theta_{M^*}(x) \geq \Theta_M(x_0) = \Theta_{M^*}(0) > 0$$

for μ^*-almost every $x \in \mathbb{R}^n$. Thus the set $S(M^*)$ defined in Lemma 2.10 is a two-dimensional linear subspace of $\mathbb{R}^n$. Moreover,

$$\mu^* = \pi^{-1}\Theta_M(x_0)\mathcal{H}^2 \llcorner S(M^*)$$

and

$$\nu^* = \mathrm{proj}^{\perp}_{S(M^*)}\mu^*.$$

Because σ is approximately continuous at x_0, we have $\sigma(x_0) = \mathrm{proj}^{\perp}_{S(M^*)}$. In particular the tangent measure pair M^* is unique. The claims now follow from standard arguments from geometric measure theory (cf. Federer [7] or Simon [20]). $\square$

3. Energy estimates. We now assume again that Ω is a three-dimensional domain, that is, $\Omega \subset \mathbb{R}^3$. For $\epsilon > 0$, we examine functions $u \in H^2_{\mathrm{loc}}(\Omega)$ such that $E_\epsilon(u)$ and $T_\epsilon(u)$ are finite. We first derive a few inequalities that will be useful when we consider the limit $\epsilon \searrow 0$ later in this section.

All of the inequalities we prove here are of a local character. Throughout the section, we assume that $\Omega_0 \subset \Omega$ is an open, bounded set such that its closure is contained in Ω.

LEMMA 3.1. *There exists a constant C such that for any $\epsilon > 0$ and any $u \in H^2_{\mathrm{loc}}(\Omega)$, the set*

$$\Omega_1 = \{x \in \Omega_0 \colon\ |u(x)| \geq 1\}$$

satisfies

$$\int_{\Omega_1} \left(\frac{\epsilon}{2}|\nabla u|^2 + \frac{1}{4\epsilon}(1-u^2)^2\right) dx \leq C\epsilon^2(T_\epsilon(u) + E_\epsilon(u)).$$

Proof. Choose a cutoff function $\eta \in C_0^\infty(\Omega)$ with $0 \leq \eta \leq 1$ and $\eta \equiv 1$ in Ω_0. Define

$$\Omega_1^+ = \{x \in \Omega \colon\ u(x) \geq 1\}$$

and

$$\Omega_1^- = \{x \in \Omega \colon\ u(x) \leq -1\}.$$

Then

$$\begin{aligned}
\epsilon \int_{\Omega_1^+} \eta^2 |\nabla u|^2\, dx &= \epsilon \int_{\Omega_1^+} \eta^2 \nabla(u-1)\cdot\nabla u\, dx \\
&= \frac{1}{\epsilon}\int_{\Omega_1^+} \eta^2 (u-1)(1-u^2)u\, dx - \int_{\Omega_1^+} \eta^2(u-1)\tau_\epsilon(u)\, dx \\
&\quad - 2\epsilon \int_{\Omega_1^+} \eta(u-1)\nabla\eta\cdot\nabla u\, dx \\
&\leq \epsilon^2 T_\epsilon(u) + \frac{1}{\epsilon}\int_{\Omega_1^+} \eta^2[(u-1)^2 + (u-1)(1-u^2)u]\, dx \\
&\quad + \frac{\epsilon}{2}\int_{\Omega_1^+} \eta^2|\nabla u|^2\, dx + 2\epsilon\|\nabla\eta\|_{L^\infty(\Omega)} \int_{\Omega_1^+} (u-1)^2\, dx.
\end{aligned}$$

In Ω_1^+, we have $(u-1)^2 \le \frac{1}{4}(1-u^2)^2$ and

$$(u-1)^2 + (u-1)(1-u^2)u = (1-u^2)^2 \left[\frac{1}{(u+1)^2} - \frac{u}{u+1}\right] \le -\frac{1}{4}(1-u^2)^2.$$

Thus

$$\int_{\Omega_1^+} \eta^2 \left(\frac{\epsilon}{2}|\nabla u|^2 + \frac{1}{4\epsilon}(1-u^2)^2\right) dx \le \epsilon^2 T_\epsilon(u) + 2\epsilon^2 \|\nabla\eta\|_{L^\infty(\Omega)} E_\epsilon(u).$$

The same computations work for Ω_1^- instead of Ω_1^+, and the claim follows. ∎

LEMMA 3.2. *There exists a constant C such that for every $x_0 \in \Omega_0$ and $\epsilon > 0$ with $B_{2\epsilon}(x_0) \subset \Omega_0$, and for every $u \in H^2_{\mathrm{loc}}(\Omega)$,*

$$\int_{B_\epsilon(x_0)} \left(\frac{\epsilon}{2}|\nabla u|^2 + \frac{1}{4\epsilon}(1-u^2)^2\right) dx \le C\epsilon^2 (E_\epsilon(u) + T_\epsilon(u) + 1).$$

Proof. Set

$$\tilde{u}(x) = \begin{cases} 1 & \text{if } u(x) \ge 1, \\ u(x) & \text{if } -1 < u(x) < 1, \\ -1 & \text{if } u(x) \le -1. \end{cases}$$

In view of Lemma 3.1, it suffices to show that

$$\int_{B_\epsilon(x_0)} \left(\frac{\epsilon}{2}|\nabla \tilde{u}|^2 + \frac{1}{4\epsilon}(1-\tilde{u}^2)^2\right) dx \le C\epsilon^2 (E_\epsilon(u) + T_\epsilon(u) + 1).$$

Choose a cutoff function $\eta \in C_0^\infty(B_{2\epsilon}(x_0))$ with $0 \le \eta \le 1$ and $\eta \equiv 1$ in $B_\epsilon(x_0)$, such that $|\nabla\eta| \le 2/\epsilon$. Then

$$\begin{aligned} \epsilon \int_\Omega \eta^2 |\nabla\tilde{u}|^2\, dx &= \epsilon \int_\Omega \eta^2 \nabla\tilde{u} \cdot \nabla u\, dx \\ &= \frac{1}{\epsilon}\int_\Omega \eta^2 \tilde{u}(1-u^2)u\, dx - \int_\Omega \eta^2 \tilde{u}\tau_\epsilon(u)\, dx - 2\epsilon \int_\Omega \eta\tilde{u}\nabla\eta\cdot\nabla u\, dx. \end{aligned}$$

We have

$$-\int_\Omega \eta^2 \tilde{u}\tau_\epsilon(u)\, dx \le \epsilon^2 T_\epsilon(u) + \frac{1}{\epsilon}\int_{B_{2\epsilon}(x_0)} \tilde{u}^2\, dx \le \epsilon^2 T_\epsilon(u) + \frac{32}{3}\pi\epsilon^2$$

and

$$-2\epsilon \int_\Omega \eta\tilde{u}\nabla\eta\cdot\nabla u\, dx = -2\epsilon\int_\Omega \eta\tilde{u}\nabla\eta\cdot\nabla\tilde{u}\, dx - 2\epsilon\int_{\Omega_1} \eta\tilde{u}\nabla\eta\cdot\nabla u\, dx,$$

where

$$\Omega_1 = \{x \in \Omega\colon\ |u(x)| \ge 1\}.$$

Moreover,

$$\begin{aligned} -2\epsilon\int_\Omega \eta\tilde{u}\nabla\eta\cdot\nabla\tilde{u}\, dx &\le \frac{\epsilon}{2}\int_\Omega \eta^2|\nabla\tilde{u}|^2\, dx + 2\epsilon\int_{B_{2\epsilon}(x_0)} |\nabla\eta|^2\, dx \\ &\le \frac{\epsilon}{2}\int_\Omega \eta^2|\nabla\tilde{u}|^2\, dx + \frac{256}{3}\pi\epsilon^2 \end{aligned}$$

and

$$-2\epsilon \int_{\Omega_1} \eta \tilde{u} \nabla \eta \cdot \nabla u \, dx \leq \epsilon \int_{\Omega_1} \eta^2 |\nabla u|^2 \, dx + \epsilon \int_{B_{2\epsilon}(x_0)} |\nabla \eta|^2 \, dx$$
$$\leq C_1 \epsilon^2 (E_\epsilon(u) + T_\epsilon(u) + 1)$$

by Lemma 3.1 for a constant C_1 that depends only on Ω and Ω_0. Finally, we have $\tilde{u}(1-u^2)u \leq 1$ in Ω; hence

$$\frac{1}{\epsilon} \int_\Omega \eta^2 \tilde{u}(1-u^2)u \, dx \leq \frac{32}{3} \pi \epsilon^2.$$

We also have

$$\frac{1}{\epsilon} \int_{B_\epsilon(x_0)} (1-\tilde{u}^2)^2 \, dx \leq \frac{4}{3} \pi \epsilon^2.$$

Combining all the estimates, we obtain the desired inequality. ∎

LEMMA 3.3. *For any $K > 0$ there exist two constants $C, r > 0$ with the following properties. Suppose $u \in H^2_{\mathrm{loc}}(\Omega)$ and $0 < \epsilon \leq \frac{1}{2} \operatorname{dist}(\Omega_0, \partial\Omega)$ such that*

$$E_\epsilon(u) + T_\epsilon(u) \leq K. \tag{3.1}$$

Then

$$\sup_{\Omega_0} |u| \leq C.$$

Moreover,

$$\operatorname*{osc}_{B_{r\epsilon}(x_0)} u \leq \frac{1}{4}$$

for any $x_0 \in \Omega_0$.

Proof. Suppose $x_0 \in \Omega$ and $0 < \epsilon \leq \frac{1}{2} \operatorname{dist}(\Omega_0, \partial\Omega)$. If (3.1) is satisfied, then by Lemma 3.2 (applied to a different Ω_0),

$$\int_{B_\epsilon(x_0)} \left(\frac{\epsilon}{2} |\nabla u|^2 + \frac{1}{4\epsilon}(1-u^2)^2 \right) dx \leq C_1 \epsilon^2$$

for a constant C_1 that depends only on Ω, Ω_0, and K. Define

$$v(x) = u(\epsilon x + x_0).$$

Then

$$\frac{1}{2} \int_{B_1(0)} \left(|\nabla v|^2 + \frac{1}{2}(1-v^2)^2 \right) dx \leq C_1 \tag{3.2}$$

and

$$\frac{1}{4} \int_{B_1(0)} (\Delta v + (1-v^2)v)^2 \, dx \leq T_\epsilon(u) \leq K. \tag{3.3}$$

It follows from (3.2) that

$$\|v\|_{H^1(B_1(0))} \leq C_2 = C_2(\Omega, \Omega_0, K).$$

According to the Sobolev embedding theorem,

$$\|v\|_{L^6(B_1(0))} \leq C_3 = C_3(\Omega, \Omega_0, K).$$

Then (3.3) implies

$$\|\Delta v\|_{L^2(B_1(0))} \leq C_4 = C_4(\Omega, \Omega_0, K).$$

Hence

$$\|v\|_{H^2(B_{1/2}(0))} \leq C_5 = C_5(\Omega, \Omega_0, K),$$

and therefore

$$\|v\|_{C^{0,1/2}(\overline{B_{1/2}(0)})} \leq C_6 = C_6(\Omega, \Omega_0, K)$$

by the Sobolev embedding theorem. Now both inequalities in the claim of the lemma follow from the last inequality. □

The next lemma provides two of the key estimates for the proof of Theorem 1.1. The first inequality of the lemma is related to estimates of what is called the "discrepancy function" by Hutchinson and Tonegawa [9], Bellettini and Mugnai [5], and others. The second one will help to exploit the condition $\frac{\partial u_\epsilon}{\partial x^3} \geq 0$ in Theorem 1.1.

Lemma 3.4. *For any $\delta > 0$ and any $K > 0$ there exist three numbers $\epsilon_0 > 0$, $c > 0$, and $R > 0$ with the following property. Suppose $0 < \epsilon \leq \epsilon_0$ and $\eta \in C_0^1(\Omega_0)$ with $0 \leq \eta \leq K$ and*

$$\sup_{\Omega_0} |\nabla \eta| \leq \frac{c}{\epsilon}.$$

Suppose further $u \in H^2_{\mathrm{loc}}(\Omega)$ such that

$$E_\epsilon(u) + T_\epsilon(u) \leq K$$

and

$$\sup_{x_0 \in \operatorname{supp} \eta} \left(\frac{1}{\epsilon} \int_{B_{R\epsilon}(x_0) \cap \Omega} (\tau_\epsilon(u))^2 \, dx \right) \leq c.$$

Then

$$\int_{\Omega_0} \eta \left[\frac{\epsilon}{2} |\nabla u|^2 - \frac{1}{4\epsilon}(1 - u^2)^2 \right] dx \leq \delta. \tag{3.4}$$

Furthermore, there exists a universal constant C such that

$$\int_{\Omega_0} \eta(\epsilon |\nabla u| - C) \left| \frac{\partial u}{\partial x^3} \right| dx \leq \delta. \tag{3.5}$$

Proof. We argue by contradiction. Suppose that the first claim (about inequality (3.4)) were false for two given numbers $\delta > 0$ and $K > 0$. Then there would exist three sequences $\epsilon_k \searrow 0$, $c_k \searrow 0$, and $R_k \to \infty$, as well as corresponding functions $\eta_k \in C_0^1(\Omega_0)$ with $0 \leq \eta_k \leq K$ and

$$\sup_{\Omega_0} |\nabla \eta_k| \leq \frac{c_k}{\epsilon_k},$$

and $u_k \in H^2_{\mathrm{loc}}(\Omega)$ with

$$E_{\epsilon_k}(u_k) + T_{\epsilon_k}(u_k) \leq K$$

and

$$\sup_{x_0 \in \operatorname{supp} \eta_k} \left(\frac{1}{\epsilon_k} \int_{B_{R_k \epsilon_k}(x_0) \cap \Omega} (\tau_{\epsilon_k}(u_k))^2 \, dx \right) \leq c_k,$$

such that

$$\limsup_{k \to \infty} \int_{\Omega_0} \eta_k \left[\frac{\epsilon_k}{2} |\nabla u_k|^2 - \frac{1}{4\epsilon_k}(1 - u_k^2)^2 \right] dx > \delta.$$

Define

$$\Omega_k^+ = \left\{ x \in \Omega \colon\ u_k(x) > \frac{1}{2} \right\}, \quad \Omega_k^- = \left\{ x \in \Omega \colon\ u_k(x) < -\frac{1}{2} \right\},$$

and

$$\Omega_k^0 = \left\{ x \in \Omega \colon\ -\frac{1}{2} \leq u_k(x) \leq \frac{1}{2} \right\}.$$

Suppose $x_0 \in \Omega_k^0$. By Lemma 3.3, there exists some $r > 0$, depending only on Ω, Ω_0, and K, such that $|u_k| \leq 3/4$ in $B_{r\epsilon_k}(x_0)$. Thus

$$\frac{1}{4\epsilon_k} \int_{B_{\epsilon_k}(x_0)} (1 - u_k^2)^2 \, dx \geq C_1 \epsilon_k^2 \tag{3.6}$$

for a constant $C_1 > 0$ that depends only on Ω, Ω_0, and K. By Vitali's covering lemma, there exists a finite set of points $x_1^k, \ldots, x_{I_k}^k \in \Omega_k^0 \cap \operatorname{supp} \eta_k$, such that

$$\Omega_k^0 \cap \operatorname{supp} \eta_k \subset \bigcup_{i=1}^{I_k} B_{5\epsilon_k}(x_i^k),$$

but

$$B_{\epsilon_k}(x_i^k) \cap B_{\epsilon_k}(x_j^k) = \emptyset \quad \text{for } i \neq j.$$

By (3.6), we have

$$I_k \leq \frac{C_2}{\epsilon_k^2}$$

for a constant C_2 that depends only on Ω, Ω_0, and K.

Fix now some $R \geq 5$. For $k \in \mathbb{N}$ and $1 \leq i \leq I_k$, choose two cutoff functions $\xi_i^k \in C_0^\infty(B_{2R\epsilon_k}(x_i^k))$ with $0 \leq \xi_i^k \leq 1$ and $\xi_i^k \equiv 1$ in $B_{R\epsilon_k}(x_i^k)$, and $\zeta_i^k \in C_0^\infty(B_{4R\epsilon_k}(x_i^k))$ with $0 \leq \zeta_i^k \leq 1$ and $\zeta_i^k \equiv 1$ in $B_{2R\epsilon_k}(x_i^k)$, such that

$$|\nabla \xi_i^k| \leq \frac{2}{R\epsilon_k} \quad \text{and} \quad |\nabla \zeta_i^k| \leq \frac{1}{R\epsilon_k}.$$

Set

$$\xi^k = \max_{1\le i\le I_k} \xi_i^k.$$

This is a Lipschitz function with $0 \le \xi^k \le 1$ and $\xi^k \equiv 1$ on $\Omega_k^0 \cap \operatorname{supp} \eta_k$. Moreover, $\operatorname{supp} \xi^k$ is contained in the set

$$U_k = \big\{x \in \Omega\colon\ \operatorname{dist}(x, \Omega_k^0 \cap \operatorname{supp} \eta_k) < 2R\epsilon_k\big\}.$$

We also have

$$|\nabla \xi^k| \le \frac{2}{R\epsilon_k}.$$

Next, we define

$$\psi_i^k = \frac{\zeta_i^k[1-(1-\xi^k)^2]}{\sum_{j=1}^{I_k} \zeta_j^k}$$

in U_k and $\psi_i^k = 0$ in $\Omega \backslash U_k$. Then we have

$$\sum_{i=1}^{I_k} \psi_i^k = 1 - (1-\xi^k)^2.$$

Moreover,

$$|\nabla \psi_i^k| \le \frac{C_3}{\epsilon_k}$$

for a constant C_3 that depends only on R.

We have

$$\begin{aligned}\epsilon_k \int_{\Omega_k^+} \eta_k (1-\xi^k)^2 |\nabla u_k|^2\, dx = {} & \frac{1}{\epsilon_k} \int_{\Omega_k^+} \eta_k (1-\xi^k)^2 (u_k - 1)(1-u_k^2) u_k\, dx \\ & - \int_{\Omega_k^+} \eta_k (1-\xi^k)^2 (u_k-1) \tau_{\epsilon_k}(u_k)\, dx \\ & - \epsilon_k \int_{\Omega_k^+} (1-\xi^k)^2 (u_k - 1) \nabla \eta_k \cdot \nabla u_k\, dx \\ & + 2\epsilon_k \int_{\Omega_k^+} \eta_k (1-\xi^k)(u_k-1) \nabla \xi^k \cdot \nabla u_k\, dx.\end{aligned}$$

It is easy to see that

$$\lim_{k\to\infty} \left(\epsilon_k \int_{\Omega_k^+} (1-\xi^k)^2 (u_k-1) \nabla \eta_k \cdot \nabla u_k\, dx \right) = 0$$

and

$$\limsup_{k\to\infty} \left(2\epsilon_k \int_{\Omega_k^+} \eta_k (1-\xi^k)(u_k-1) \nabla \xi^k \cdot \nabla u_k\, dx \right) \le \frac{C_4}{R}$$

for a constant C_4 that depends only on K. Moreover,

$$\lim_{k\to\infty}\int_{\Omega_k^+}\eta_k(1-\xi^k)^2(u_k-1)\tau_{\epsilon_k}(u_k)\,dx=0.$$

We also have

$$\frac{1}{\epsilon_k}\int_{\Omega_k^+}\eta_k(1-\xi^k)^2(u_k-1)(1-u_k^2)u_k\,dx\le-\frac{1}{3\epsilon_k}\int_{\Omega_k^+}\eta_k(1-\xi^k)^2(1-u_k^2)^2\,dx$$

for every k. Similar estimates hold for Ω_k^- instead of Ω_k^+. It follows that

$$\limsup_{k\to\infty}\int_{\Omega_0}\eta_k(1-\xi^k)^2\left[\frac{\epsilon_k}{2}|\nabla u_k|^2+\frac{1}{4\epsilon_k}(1-u_k^2)^2\right]dx\le\frac{C_5}{R}$$

for a constant C_5 that depends only on K. If we choose R such that

$$R\ge\frac{2C_5}{\delta},$$

then we have

$$\limsup_{k\to\infty}\sum_{i=1}^{I_k}\int_{\Omega_0}\eta_k\psi_i^k\left[\frac{\epsilon_k}{2}|\nabla u_k|^2-\frac{1}{4\epsilon_k}(1-u_k^2)^2\right]dx\ge\frac{\delta}{2}.$$

Hence for infinitely many values of k (we may assume without loss of generality: for any k) there is an index $i_0(k)\in\{1,\dots,I_k\}$ such that

$$\int_{\Omega_0}\eta_k\psi_{i_0(k)}^k\left[\frac{\epsilon_k}{2}|\nabla u_k|^2-\frac{1}{4\epsilon_k}(1-u_k^2)^2\right]dx\ge\frac{\delta\epsilon_k^2}{2C_2}.$$

Define $\tilde{x}_k=x_{i_0(k)}^k$ and

$$v_k(x)=u_k(\epsilon_k x+\tilde{x}_k).$$

Moreover, define

$$\tilde{\psi}_k(x)=\psi_{i_0(k)}^k(\epsilon_k x+\tilde{x}_k).$$

Let $r>0$. By Lemma 3.2, we have

$$\frac{1}{2}\int_{B_r(0)}\left(|\nabla v_k|^2+\frac{1}{2}(1-v_k^2)^2\right)dx\le C_6r^3$$

for a constant C_6 that depends only on Ω, Ω_0, and K, whenever k is large enough. Furthermore,

$$\int_{B_r(0)}(\Delta v_k+(1-v_k^2)v_k)^2\,dx\le\frac{1}{\epsilon_k}\int_{B_{R_k\epsilon_k}(\tilde{x}_k)}(\tau_{\epsilon_k}(u_k))^2\,dx\le c_k\to 0$$

as $k\to\infty$. By Lemma 3.3, there exists a number C_7, depending only on K, such that $|v_k|\le C_7$ in $B_r(0)$ for any sufficiently large k. The number $r>0$ is arbitrary. Thus a subsequence of $\{v_k\}$ converges to a bounded solution of

$$\Delta v+(1-v^2)v=0\quad\text{in }\mathbb{R}^3. \tag{3.7}$$

The convergence is strong in $H^1(B_r(0))$ for every $r > 0$. Because $\nabla\tilde{\psi}_k$ is uniformly bounded, we may assume that $\tilde{\psi}_k \to \psi$ uniformly, where $\psi : B_{4R}(0) \to [0,1]$ is a continuous function. Moreover, we may assume $\tilde{x}_k \to x_0 \in \Omega_0$. Because of the gradient bound for η_k, the functions $\eta_k(\epsilon_k x + \tilde{x}_k)$ converge uniformly to a constant $a \in [0, K]$ after we have picked another subsequence. It follows that

$$\frac{a}{2}\int_Q \psi\left(|\nabla v|^2 - \frac{1}{2}(1-v^2)^2\right) dx \geq \frac{\delta}{2C_2} > 0.$$

However, it was proved by Modica [11] that a bounded solution of (3.7) satisfies

$$|\nabla v|^2 \leq \frac{1}{2}(1-v^2)^2$$

pointwise. This gives a contradiction, and (3.4) follows.

To prove (3.5), we use the same method. There exists a constant C such that any bounded solution of (3.7) satisfies

$$|\nabla v| \leq C.$$

(To prove this, one can first estimate the energy in balls of radius 1 with the methods from the proof of Lemma 3.2 and then apply standard regularity results for elliptic equations.) In particular

$$(|\nabla v| - C)\left|\frac{\partial v}{\partial x^3}\right| \leq 0.$$

The rest of the proof is similar to the above. We omit the details. ∎

4. Proof of Theorem 1.1. In this section we prove Theorem 1.1. Some of the facts we establish are also true without the assumption $\frac{\partial u_\epsilon}{\partial x^3} \geq 0$; therefore we work first without this condition and impose it only later.

We assume that $u_\epsilon \in H^2_{\text{loc}}(\Omega)$ are functions such that

$$\liminf_{\epsilon \searrow 0}(E_\epsilon(u_\epsilon) + T_\epsilon(u_\epsilon)) < \infty.$$

We can choose a sequence $\epsilon_k \searrow 0$ such that

$$\limsup_{k\to\infty}(E_\epsilon(u_\epsilon) + T_\epsilon(u_\epsilon)) < \infty.$$

We will choose a subsequence of $\{\epsilon_k\}$ several times in the proof. We will always denote the subsequence by $\{\epsilon_k\}$ again for convenience.

We consider the measure pairs $M^{(k)} = (\mu^{(k)}, \nu^{(k)}) \in \mathcal{W}(\Omega)$ defined by

$$\mu^{(k)} = \mathcal{L}^3 \llcorner \left(\frac{\epsilon_k}{2}|\nabla u_{\epsilon_k}|^2 + \frac{1}{4\epsilon_k}(1-u_{\epsilon_k}^2)^2\right)$$

and

$$\nu^{(k)}_{\alpha\beta} = \epsilon_k \mathcal{L}^3 \llcorner \left(\frac{\partial u_{\epsilon_k}}{\partial x^\alpha}\frac{\partial u_{\epsilon_k}}{\partial x^\beta}\right).$$

We have

$$\limsup_{k\to\infty}(\mu^{(k)}(\Omega) + W(M^{(k)})) < \infty;$$

thus we may assume that (after passing to a subsequence)

$$M^{(k)} \rightharpoonup M$$

for some $M = (\mu, \nu) \in \mathcal{W}(\Omega)$ with

$$W(M) \leq \liminf_{k\to\infty} W(M^{(k)}) \leq \liminf_{k\to\infty} T_{\epsilon_k}(u_{\epsilon_k}).$$

(The second inequality follows from the same computations as in the examples of section 2.) We also consider the Radon measures

$$h_k = \frac{1}{\epsilon_k}\mathcal{L}^3 \llcorner (\tau_{\epsilon_k}(u_{\epsilon_k}))^2$$

on Ω. We have

$$\limsup_{k\to\infty} h_k(\Omega) < \infty;$$

hence we may assume that $h_k \to h$ for a Radon measure h on Ω.

Lemma 4.1. *Every $x_0 \in \Omega$ satisfies $\mu(\{x_0\}) = 0$.*

Proof. Suppose we have $\mu(\{x_0\}) > 0$ for some $x_0 \in \Omega$. For $\phi \in C_0^1(B_1(0), \mathbb{R}^3)$ and $r > 0$, set

$$\phi_r(x) = \phi((x - x_0)/r).$$

We have

$$\int_\Omega \left(\operatorname{div}\phi_r\, d\mu - \frac{\partial\phi_r^\alpha}{\partial x^\beta}\, d\nu_{\alpha\beta} + \phi_r^\alpha H^\beta\, d\nu_{\alpha\beta}\right) = 0,$$

where H is the generalized mean curvature of M. Hence

$$\int_\Omega \left(\operatorname{div}\phi((x-x_0)/r)\, d\mu(x) - \frac{\partial\phi^\alpha}{\partial x^\beta}((x-x_0)/r)\, d\nu_{\alpha\beta}(x)\right) = -r\int_\Omega \phi^\alpha((x-x_0)/r)H^\beta(x)\, d\nu_{\alpha\beta}(x) \to 0$$

as $r \searrow 0$. On the other hand, the left-hand side converges to

$$\operatorname{div}\phi(0)\,\mu(\{x_0\}) - \frac{\partial\phi^\alpha}{\partial x^\beta}(0)\,\nu_{\alpha\beta}(\{x_0\}).$$

It follows that $\operatorname{trace}\nu(\{x_0\}) = 3\mu(\{x_0\})$. But this is impossible because

$$\operatorname{trace}\nu^{(k)} \leq 2\mu^{(k)},$$

and this contradiction completes the proof. □

Lemma 4.2. $M \in \mathcal{W}_2(\Omega)$.

Proof. We need to prove that

$$\operatorname{trace}\nu \leq \mu. \tag{4.1}$$

To this end, we choose a function $\phi \in C_0^1(\Omega)$ with $\phi \geq 0$. For a fixed $\delta > 0$, let ϵ_0, c, and R be the numbers from Lemma 3.4 belonging to δ and

$$K = \sup_{k\in\mathbb{N}}(E_{\epsilon_k}(u_{\epsilon_k}) + T_{\epsilon_k}(u_{\epsilon_k})) + \sup_\Omega \phi.$$

Define

$$V_0^{c/2} = \{x \in \Omega : \ h(\{x\}) \geq c/2\}$$

and

$$V_r^{c/2} = \bigcup_{x \in V_0^{c/2}} B_r(x)$$

for $r > 0$. The set $V_0^{c/2}$ is finite. Because of Lemma 4.1, there exists a number $r_0 > 0$ such that $\mu(V_{r_0}^{c/2}) \leq \delta$. Moreover, we have

$$\frac{1}{\epsilon_k} \int_{B_{R\epsilon_k}(x_0)} (\tau_{\epsilon_k}(u_{\epsilon_k}))^2 \, dx < c$$

for any $x_0 \in \operatorname{supp} \phi \backslash V_{r_0/2}^{c/2}$ and any sufficiently large k. If $\psi \in C^1(\Omega)$ is a function with $0 \leq \psi \leq 1$, $\psi \equiv 0$ in $V_{r_0/2}^{c/2}$, and $\psi \equiv 1$ outside of $V_{r_0}^{c/2}$, we can apply Lemma 3.4 to ϵ_k, u_k, and $\eta = \phi\psi$ for every sufficiently large k. It follows that

$$\int_\Omega \phi\psi \operatorname{trace} d\nu \leq \int_\Omega \phi\psi \, d\mu + \delta.$$

Thus

$$\int_\Omega \phi \operatorname{trace} d\nu \leq \int_\Omega \phi \, d\mu + \delta(1 + K).$$

Since δ is arbitrary, we conclude that (4.1) holds. $\square$

We now consider the functions

$$v_\epsilon = \frac{1}{\sqrt{2}} \left(u_\epsilon - \frac{u_\epsilon^3}{3} \right).$$

We have

$$\nabla v_\epsilon = \frac{1}{\sqrt{2}} (1 - u_\epsilon^2) \nabla u_\epsilon;$$

thus

$$\int_\Omega |\nabla v_\epsilon| \, dx \leq \int_\Omega \left(\frac{\epsilon}{2} |\nabla u_\epsilon|^2 + \frac{1}{4\epsilon} (1 - u_\epsilon^2)^2 \right) \, dx = E_\epsilon(u_\epsilon)$$

by Young's inequality. From the theory of functions of bounded variation and sets of bounded perimeter (see, e.g., Giusti [8] or Ambrosio, Fusco, and Pallara [3]), it follows that there exists a subsequence of $\{\epsilon_k\}$ such that $\mathcal{L}^3$-almost everywhere in Ω,

$$v_{\epsilon_k} \to v,$$

where $v : \Omega \to \{-\sqrt{2}/3, \sqrt{2}/3\}$ is a function such that the set

$$U = \left\{ x \in \Omega : \ v(x) = -\frac{\sqrt{2}}{3} \right\}$$

is of finite perimeter (in Ω). Moreover, the reduced boundary $\partial^* U$ satisfies

$$\frac{2\sqrt{2}}{3}\mathcal{H}^2 \llcorner \partial^* U \leq \mu.$$

For every point $x \in \partial^* U$, we have

$$\lim_{r \searrow 0} r^{-2}\mathcal{H}^2(\partial^* U \cap B_r(x_0)) = \pi.$$

Hence

$$\Theta_M(x) \geq \frac{2\sqrt{2}}{3}\pi.$$

By Lemma 2.7, the same inequality holds for any point in the relative closure of $\partial^* U$ in Ω, which we denote by $\overline{\partial^* U}$. Thus

$$\overline{\partial^* U} \subset \Sigma = \left\{ x \in \Omega \colon \Theta_M(x) \geq \frac{2\sqrt{2}}{3}\pi \right\}.$$

The set Σ is relatively closed. According to Proposition 2.11, it is countably 2-rectifiable. If $\nu = \mu \llcorner \sigma$, then

$$\sigma(x) = \mathrm{proj}^{\perp}_{T_x\Sigma}$$

μ-almost everywhere on Σ (which is also $\mathcal{H}^2$-almost everywhere on Σ). In particular,

$$\mathrm{trace}\, \nu \llcorner \Sigma = \mu \llcorner \Sigma.$$

The function v is locally constant on $\Omega \backslash \Sigma$ after we have changed it on an $\mathcal{L}^3$-null set.

The most difficult part of the proof of Theorem 1.1 is to show that $\mu(\Omega \backslash \Sigma) = 0$. To this end, we now use the additional assumption $\Omega = \Omega' \times \mathbb{R}$ and

$$\frac{\partial u_\epsilon}{\partial x^3} \geq 0 \tag{4.2}$$

$\mathcal{L}^3$-almost everywhere in Ω. Then we use arguments based on an idea of Ambrosio and Cabré [2].

Lemma 4.3. $\mu(\Omega \backslash \Sigma) = 0$.

Proof. Suppose $\eta \in C^2_0(\Omega \backslash \Sigma)$ is a nonnegative function, the support of which is contained in one of the connected components of $\Omega \backslash \Sigma$. Then we have either $u_{\epsilon_k}(x) \to -1$ for $\mathcal{L}^3$-almost every $x \in \mathrm{supp}\, \eta$ or $u_{\epsilon_k}(x) \to 1$ for $\mathcal{L}^3$-almost every $x \in \mathrm{supp}\, \eta$. We may assume without loss of generality that the limit is 1. Thus

$$\int_{\mathrm{supp}\, \eta} u_{\epsilon_k}\, dx \to \mathcal{L}^3(\mathrm{supp}\, \eta)$$

by Lebesgue's convergence theorem.

For $t \geq 0$, define

$$\eta_t(x^1, x^2, x^3) = \eta(x^1, x^2, x^3 - t).$$

We calculate

$$\begin{aligned}\frac{d}{dt}\int_\Omega \eta_t\,d\mu &= -\int_\Omega \frac{\partial\eta}{\partial x^3}(x^1,x^2,x^3-t)\,d\mu(x)\\ &= -\int_\Omega \frac{\partial\eta}{\partial x^\beta}(x^1,x^2,x^3-t)\,d\nu_{3\beta}(x)+\int_\Omega \eta_t H^\beta\,d\nu_{3\beta}.\end{aligned}$$

We have

$$\left|\int_\Omega \frac{\partial\eta}{\partial x^\beta}(x^1,x^2,x^3-t)\,d\nu_{3\beta}(x)\right| \le \lim_{k\to\infty}\left(\epsilon_k\int_\Omega |\nabla\eta(x^1,x^2,x^3-t)||\nabla u_{\epsilon_k}|\frac{\partial u_{\epsilon_k}}{\partial x^3}\,dx\right).$$

Using Lemma 3.4 and arguments similar to those in the proof of Lemma 4.2, we find that the right-hand side is at most

$$C\|\nabla\eta\|_{L^\infty(\Omega)}\lim_{k\to\infty}\int_{\operatorname{supp}\eta_t}\frac{\partial u_{\epsilon_k}}{\partial x^3}\,dx$$

for a universal constant C. Moreover, for any $\lambda>0$, we have

$$\begin{aligned}\left|\int_\Omega \eta_t H^\beta\,d\nu_{3\beta}\right| &\le \left(\int_\Omega \eta_t^2\,d\nu_{33}\right)^{1/2}\left(\int_\Omega H^\alpha H^\beta\,d\nu_{\alpha\beta}\right)^{1/2}\\ &\le \lambda W(M)+\frac{1}{\lambda}\lim_{k\to\infty}\left(\epsilon_k\int_\Omega \eta_t^2\left(\frac{\partial u_{\epsilon_k}}{\partial x^3}\right)^2\,dx\right)\\ &\le \lambda W(M)+\frac{C}{\lambda}\|\eta\|^2_{L^\infty(\Omega)}\lim_{k\to\infty}\int_{\operatorname{supp}\eta_t}\frac{\partial u_{\epsilon_k}}{\partial x^3}\,dx.\end{aligned}$$

Because of (4.2),

$$\int_0^\infty\int_{\operatorname{supp}\eta_t}\frac{\partial u_{\epsilon_k}}{\partial x^3}\,dx\,dt \le \int_{\operatorname{supp}\eta}(1-u_{\epsilon_k})\,dx.$$

Thus for $T>0$, we have a number C_1 such that

$$\begin{aligned}\int_\Omega \eta\,d\mu &\le \int_\Omega \eta_T\,d\mu+\lambda TW(M)+C_1\left(1+\frac{1}{\lambda}\right)\lim_{k\to\infty}\int_{\operatorname{supp}\eta}(1-u_{\epsilon_k})\,dx\\ &= \int_\Omega \eta_T\,d\mu+\lambda TW(M).\end{aligned}$$

Because λ is arbitrary and because $\mu(\Omega)<\infty$, it follows that

$$\int_\Omega \eta\,d\mu \le \liminf_{T\to\infty}\int_\Omega \eta_T\,d\mu = 0.$$

We conclude that $\mu(\Omega\backslash\Sigma)=0$. □

It follows from Lemma 4.3 and the existence of the 2-density Θ_M that μ is absolutely continuous with respect to $\mathcal{H}^2\llcorner\Sigma$. More precisely,

$$\mu=\pi^{-1}(\mathcal{H}^2\llcorner\Sigma)\llcorner\Theta_M.$$

Together with $\operatorname{trace}\nu=\mu$, this implies part (i) of Theorem 1.1 for $\theta=\Theta_M$. We have already proved (ii). Part (iii) follows from Proposition 2.2 and the facts we know about the structure of M. This concludes the proof of Theorem 1.1.

REFERENCES

[1] W. K. Allard, *On the first variation of a varifold*, Ann. of Math. (2), 95 (1972), pp. 417–491.

[2] L. Ambrosio and X. Cabré, *Entire solutions of semilinear elliptic equations in $\mathbb{R}^3$ and a conjecture of De Giorgi*, J. Amer. Math. Soc., 13 (2000), pp. 725–739.

[3] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Math. Monogr., The Clarendon Press, Oxford University Press, New York, 2000.

[4] L. Ambrosio and H. M. Soner, *A measure-theoretic approach to higher codimension mean curvature flows*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 25 (1997), pp. 27–49.

[5] G. Bellettini and L. Mugnai, *On the approximation of the elastica functional in radial symmetry*, Calc. Var. Partial Differential Equations, 24 (2005), pp. 1–20.

[6] E. De Giorgi, *Some remarks on Γ-convergence and least squares method*, in Composite Media and Homogenization Theory, Trieste, 1990, Progr. Nonlinear Differential Equations Appl. 5, Birkhäuser Boston, Boston, 1991, pp. 135–142.

[7] H. Federer, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[8] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Monogr. Math. 80, Birkhäuser-Verlag, Basel, 1984.

[9] J. E. Hutchinson and Y. Tonegawa, *Convergence of phase interfaces in the van der Waals-Cahn-Hilliard theory*, Calc. Var. Partial Differential Equations, 10 (2000), pp. 49–84.

[10] F.-H. Lin, *Varifold type theory for Sobolev mappings*, in First International Congress of Chinese Mathematicians (Beijing, 1998), AMS, Providence, RI, 2001, pp. 423–430.

[11] L. Modica, *A gradient bound and a Liouville theorem for nonlinear Poisson equations*, Comm. Pure Appl. Math., 38 (1985), pp. 679–684.

[12] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Ration. Mech. Anal., 98 (1987), pp. 123–142.

[13] L. Modica, *Gradient theory of phase transitions with boundary contact energy*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 4 (1987), pp. 487–512.

[14] L. Modica and S. Mortola, *Il limite nella Γ-convergenza di una famiglia di funzionali ellittici*, Boll. Un. Mat. Ital. A (5), 14 (1977), pp. 526–529.

[15] L. Modica and S. Mortola, *Un esempio di Γ^--convergenza*, Boll. Un. Mat. Ital. B (5), 14 (1977), pp. 285–299.

[16] R. Moser, *Stationary measures and rectifiability*, Calc. Var. Partial Differential Equations, 17 (2003), pp. 357–368.

[17] R. Moser, *Energy concentration for almost harmonic maps and the Willmore functional*, Math. Z., 251 (2005), pp. 293–311.

[18] R. Moser, *On the structure of energy concentration measures*, Calc. Var. Partial Differential Equations, to appear.

[19] P. Price, *A monotonicity formula for Yang-Mills fields*, Manuscripta Math., 43 (1983), pp. 131–166.

[20] L. Simon, *Lectures on Geometric Measure Theory*, Australian National University Centre for Mathematical Analysis, Canberra, 1983.

[21] L. Simon, *Existence of surfaces minimizing the Willmore functional*, Comm. Anal. Geom., 1 (1993), pp. 281–326.

[22] P. Sternberg, *The effect of a singular perturbation on nonconvex variational problems*, Arch. Ration. Mech. Anal., 101 (1988), pp. 209–260.

[23] Y. Tonegawa, *Phase field model with a variable chemical potential*, Proc. Roy. Soc. Edinburgh Sect. A, 132 (2002), pp. 993–1019.

[24] T. J. Willmore, *Riemannian Geometry*, Oxford Sci. Pub., The Clarendon Press, Oxford University Press, New York, 1993.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 737–751

DIFFEOMORPHISMS AND NONLINEAR HEAT FLOWS*

L. C. EVANS†, O. SAVIN†, AND W. GANGBO‡

Abstract. We show that the gradient flow $\mathbf{u}$ on L^2 generated by the energy functional $I[\mathbf{u}] := \int_U \Phi(\det D\mathbf{u})\,dx$ for vector-valued mappings is in some sense "integrable," meaning that (i) the inverse Jacobian $\beta := (\det D\mathbf{u})^{-1}$ satisfies a scalar nonlinear diffusion equation, and (ii) we can recover $\mathbf{u}$ by solving an ODE determined by β.

Key words. quasi convexity, gradient flow, Jacobian

AMS subject classifications. 35K45, 49J40

DOI. 10.1137/04061386X

1. Introduction.

1.1. Gradient flows for quasi-convex energies. This paper is a contribution to the mostly unsolved problem of understanding the gradient flow dynamics on L^2 generated by integral functionals having the form

$$I[\mathbf{v}] := \int_U F(D\mathbf{v})\,dx, \tag{1.1}$$

defined for functions $\mathbf{v} : U \to \mathbb{R}^m$, where U is an open subset of $\mathbb{R}^n$. The gradient $D\mathbf{v}$ belongs to $\mathbb{M}^{m\times n}$, the space of $m \times n$ matrices, and we are given the nonlinearity $F : \mathbb{M}^{m\times n} \to (-\infty, +\infty]$.

Quasi convexity. As is well known, the critical assumption for the existence of minimizers of $I[\cdot]$, subject to appropriate boundary conditions, is that F be *quasi-convex* in the sense of C. B. Morrey, Jr. This is the condition that

$$\int_U F(A)\,dx \le \int_U F(A + D\mathbf{v})\,dx \tag{1.2}$$

for all matrices $A \in \mathbb{M}^{m\times n}$ and all C^1 functions $\mathbf{v} : U \to \mathbb{R}^m$ vanishing on ∂U.

Dynamics. As the existence and (partial) regularity theories for minimizers are fairly well understood, it has long seemed natural to turn attention to related dynamical problems. The corresponding flow on L^2 generated by $I[\cdot]$ is the initial-value problem for the system of PDEs

$$\begin{cases} \mathbf{u}_t = \operatorname{div}(DF(D\mathbf{u})) & (t > 0), \\ \mathbf{u} = \mathbf{u}^0 & (t = 0), \end{cases} \tag{1.3}$$

*Received by the editors August 25, 2004; accepted for publication February 1, 2005; published electronically November 9, 2005.
http://www.siam.org/journals/sima/37-3/61386.html
†Department of Mathematics, UC Berkeley, Berkeley, CA 94720 (evans@math.berkeley.edu, savin@math.berkeley.edu). The first author's research was supported in part by NSF grant DMS-0070480. The second author's research was supported by the Miller Institute for Basic Research in Science, UC Berkeley.
‡School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (gangbo@math.gatech.edu). This author's research was supported in part by NSF grants DMS-00-74037 and DMS-02-00267.

with appropriate boundary conditions.

Given the quasi-convexity hypothesis (1.2), the system (1.3) is parabolic, at least in some weak sense. However, it is extremely nonlinear, so much so that it remains to date a challenging open problem to prove existence of even weak solutions, to understand uniqueness issues, and/or to show partial regularity.

Time-step approximations. One obvious approach is to approximate by an implicit *time-step approximation*. For this, we fix a step size $h > 0$ and recursively find $\mathbf{u}_{k+1}$ to minimize

$$I_k[\mathbf{v}] := \frac{1}{2}\int_U |\mathbf{v} - \mathbf{u}_k|^2\,dx + h\int_U F(D\mathbf{v})\,dx, \tag{1.4}$$

with appropriate boundary conditions, given $\mathbf{u}_k$. The Euler–Lagrange equations read

$$\begin{cases} \frac{\mathbf{u}_{k+1}-\mathbf{u}_k}{h} = \operatorname{div}(DF(D\mathbf{u}_{k+1})) \qquad (k = 0, 1, \dots), \\ \qquad \mathbf{u}_0 = \mathbf{u}^0. \end{cases} \tag{1.5}$$

This procedure generates a strong candidate for an approximation to the full dynamics (1.3). The fundamental point is that under our quasi-convexity assumption we can in fact iteratively find minimizers of (1.4).

The really hard task is passing to limits as $h \to 0$. Since our approximations $\mathbf{u}_k$ are minimizers, and not just critical points, of $I_k[\cdot]$, the expectation and hope is that we obtain in the limit some sort of reasonable weak solution of (1.3). It has, however, proved in practice impossible to carry out this program in general, owing to the usual problem in nonlinear PDE that we do not have very good uniform estimates on the approximate solutions $\mathbf{u}_k$. (The paper [E] demonstrates a completely different minimization principle, but we have not been able to exploit this usefully.)

1.2. Nonlinearities depending only on the determinant. This paper documents some progress in this matter for the case $m = n$ and nonlinearities F with the special structure

$$F(P) = \Phi(\det P) \qquad (P \in \mathbb{M}^{n\times n}), \tag{1.6}$$

where Φ is a convex function and "det" means determinant. Such a nonlinearity is quasi-convex, and it has long been known that for the static calculus of variations the particular hypothesis (1.6) has strong implications; see, for instance, Dacorogna [D].

We begin by reviewing the issue of minimizing the functional

$$I[\mathbf{v}] := \int_U F(D\mathbf{v})\,dx = \int_U \Phi(\det D\mathbf{v})\,dx \tag{1.7}$$

among mappings $\mathbf{v} = (v^1, \dots, v^n)$ from a connected, open set $U \subset \mathbb{R}^n$ into $\mathbb{R}^n$. We write the gradient matrix of $\mathbf{v}$ as

$$D\mathbf{v} = \begin{pmatrix} v^1_{x_1} & \dots & v^1_{x_n} \\ \vdots & \ddots & \vdots \\ v^n_{x_1} & \dots & v^n_{x_n} \end{pmatrix}.$$

If $\mathbf{u} = (u^1, \dots, u^n)$ is a smooth minimizer of $I[\cdot]$, subject to boundary conditions which for the moment we do not specify, then $\mathbf{u}$ solves the Euler–Lagrange system of PDEs

$$\operatorname{div}(DF(D\mathbf{u})) = \operatorname{div}(\Phi'(\det D\mathbf{u})(\operatorname{cof} D\mathbf{u})^T) = 0, \tag{1.8}$$

where $\operatorname{cof} D\mathbf{u}$ is the cofactor matrix formed from $D\mathbf{u}$. To derive (1.8) we employed the formula

$$\frac{\partial \det P}{\partial p_i^k} = (\operatorname{cof} P)_i^k \qquad (1 \le i, k \le n) \tag{1.9}$$

for the $n \times n$ matrix P, whose (i,k) entry is denoted p_i^k. Likewise, $(\operatorname{cof} P)_i^k$ means the (i,k) entry of $\operatorname{cof} P$. Formula (1.9) is a consequence of the matrix identity

$$(\operatorname{cof} P)^T P = I \det P, \tag{1.10}$$

but for any C^2 function $\mathbf{w} = (w^1, \dots, w^n)$ we have

$$\operatorname{div}((\operatorname{cof} D\mathbf{w})^T) \equiv 0; \tag{1.11}$$

that is,

$$(\operatorname{cof} D\mathbf{w})^k_{i,x_i} = 0 \qquad (k = 1, \dots, n).$$

Therefore (1.8) implies

$$0 = \Phi''(\det D\mathbf{u}) D(\det D\mathbf{u})(\operatorname{cof} D\mathbf{u})^T. \tag{1.12}$$

In view of (1.10), our multiplying (1.10) by $D\mathbf{u}$ gives

$$0 = \Phi''(\det D\mathbf{u}) D(\det D\mathbf{u})(\det D\mathbf{u}) = \frac{1}{2}\Phi''(\det D\mathbf{u}) D(\det D\mathbf{u})^2.$$

Assuming next the strict convexity condition that $\Phi'' > 0$, we deduce that $(\det D\mathbf{u})^2$ is constant within U. Thus, if $\mathbf{u}$ is smooth, we conclude that

$$\det D\mathbf{u} \equiv C \quad \text{within} \quad U \tag{1.13}$$

for some constant C.

1.3. A gradient flow. We study in this paper the corresponding "heat flow" governed by the function $I[\cdot]$, that is, the system of PDEs

$$\mathbf{u}_t = \operatorname{div}(DF(D\mathbf{u})) = \operatorname{div}(\Phi'(\det D\mathbf{u})(\operatorname{cof} D\mathbf{u})^T), \tag{1.14}$$

plus appropriate initial and boundary conditions, detailed later.

We are especially interested in the case that $\Phi(d) < \infty$ for $d > 0$, $\Phi(d) = \infty$ for $d < 0$, and $\lim_{d\to 0^+} \Phi(d) = +\infty$. Then (1.14) enforces the constraint

$$\det D\mathbf{u} > 0.$$

We can hope therefore that for each time t the mapping $x \mapsto y = \mathbf{u}(x,t)$ is a diffeomorphism, with inverse $y \mapsto x = \mathbf{v}(y,t)$. And since the static problem, recalled in section 1.1, is so simple, we hope as well that the analysis of the system (1.14) may not be so complicated.

This is in fact so, for as we will see in section 2, the quantity

$$\beta := (\det D\mathbf{u})^{-1} > 0, \tag{1.15}$$

regarded as a function of y and t, solves the nonlinear parabolic PDE

$$\beta_t = \operatorname{div}\left(\Phi''\left(\frac{1}{\beta}\right)\frac{D\beta}{\beta^2}\right) = \operatorname{div}\left(\beta\Psi'(\beta)D\beta\right) \tag{1.16}$$

with Neumann boundary conditions, where

$$\Psi(d) := d\Phi\left(\frac{1}{d}\right) \quad \text{for } d > 0.$$

Now (1.16) is singular in regimes where $\beta \to 0$ or ∞, but the maximum principle implies that if the initial data β^0 is bounded away from 0 and ∞, then so is the solution.

We will show furthermore that given β, the solution of (1.16) with appropriate initial conditions, we can then recover the mappings $\mathbf{u}$ by solving a system of ODEs governed by β and proving then that the PDE (1.14) holds. In this sense, we can regard the parabolic system of PDEs (1.14) as being somehow "integrable."

1.4. Outline. Our paper introduces in section 2 the formal computations showing how (1.16) results from (1.14). Section 3 then reverses this process to provide careful proofs: we start with the solution β of the nonlinear diffusion equation and build from it the mappings $\mathbf{u}(\cdot, t)$ for $t > 0$.

Section 4 introduces some interesting variants of our construction, the first for more general integrands than in (1.7). We discuss also a situation when the range of the initial mapping $\mathbf{u}^0$ is a proper subset W_0 of the target V. In this case we can design Φ so that the flow "fills up" V in finite time. Interesting complications occur if U and V are not in fact diffeomorphic.

The concluding section 5 introduces and analyzes a related "time-stepping" dynamic variational principle. This discussion will make much clearer the connections between our PDE (1.16) and (1.14).

2. Calculations for smooth solutions.

Suppose now U is a smooth, open, bounded, connected subset of $\mathbb{R}^n$, and

$$\mathbf{u} : \bar{U} \times [0, \infty) \to \mathbb{R}^n$$

is smooth, $\mathbf{u} = (u^1, \dots, u^n)$. In this section we suppose as well that $\mathbf{u}$ solves the system (1.14). Let $\mathbf{u}^0 = \mathbf{u}(\cdot, 0)$ denote the initial mapping.

2.1. Changing variables.

Suppose that for each time $t \geq 0$, the mapping

$$\mathbf{u}(\cdot, t) : \bar{U} \to \bar{V}$$

is a diffeomorphism, where $V \subset \mathbb{R}^n$ is a fixed open subset of $\mathbb{R}^n$. We can then invert the relationship

$$y = \mathbf{u}(x, t) \qquad (x \in \bar{U}, y \in \bar{V}) \tag{2.1}$$

to give

$$x = \mathbf{v}(y, t) \qquad \text{for } \mathbf{v} := \mathbf{u}^{-1}. \tag{2.2}$$

Set

$$\beta(y, t) := \det D\mathbf{v}(y, t) = (\det D\mathbf{u}(x, t))^{-1}. \tag{2.3}$$

2.2. A PDE for . Our main observation is that β solves a scalar, nonlinear diffusion equation.

THEOREM 2.1. *We have*

$$\begin{cases} \beta_t = \operatorname{div}\left(\Phi''\left(\frac{1}{\beta}\right)\frac{1}{\beta^2}D\beta\right) & \textit{in } V\times(0,\infty), \\ \frac{\partial\beta}{\partial\nu} = 0 & \textit{on } \partial V\times(0,\infty), \end{cases} \tag{2.4}$$

ν denoting the unit outward-pointing normal vectorfield to ∂V.

Proof. 1. Fix any time $T>0$ and select a smooth function $\zeta : \bar{V}\times[0,T]\to\mathbb{R}$ such that

$$\zeta(\cdot,0)\equiv\zeta(\cdot,T)\equiv 0. \tag{2.5}$$

Then, employing (2.1), we compute

$$\begin{aligned} &\int_0^T\int_V \beta\zeta_t + D_y\left(\Phi'\left(\frac{1}{\beta}\right)\right)\cdot D_y\zeta\, dydt \\ &= \int_0^T\int_U \left[\beta(\mathbf{u},t)\zeta_t(\mathbf{u},t) + D_x\left(\Phi'\left(\frac{1}{\beta}\right)\right)(D\mathbf{u})^{-1}\cdot D_y\zeta\right]\frac{dx}{\beta(\mathbf{u},t)}\,dt \\ &= \int_0^T\int_U \frac{\partial}{\partial t}(\zeta(\mathbf{u},t)) - D_y\zeta\cdot\mathbf{u}_t + D_x\left(\Phi'\left(\frac{1}{\beta}\right)\right)\frac{(D\mathbf{u})^{-1}}{\beta}\cdot D_y\zeta\, dxdt \\ &= -\int_0^T\int_U D_y\zeta\cdot\left[\mathbf{u}_t - D_x\left(\Phi'\left(\frac{1}{\beta}\right)\right)\frac{(D\mathbf{u})^{-1}}{\beta}\right]\, dxdt. \end{aligned} \tag{2.6}$$

Now our PDE (1.14) reads

$$\mathbf{u}_t = \operatorname{div}_x(\Phi'(\det D\mathbf{u})\det D\mathbf{u}(D\mathbf{u})^{-1}) = D_x\left(\Phi'\left(\frac{1}{\beta}\right)\right)\frac{(D\mathbf{u})^{-1}}{\beta},$$

since $\operatorname{div}((\det D\mathbf{u})(D\mathbf{u})^{-1}) = \operatorname{div}(\operatorname{cof} D\mathbf{u}^T)\equiv 0$. Consequently the expression within the square brackets in the last term of (2.6) vanishes. So

$$\int_0^T\int_V \beta\zeta_t + D_y\left(\Phi'\left(\frac{1}{\beta}\right)\right)\cdot D_y\zeta\, dydt = 0$$

for all test functions ζ as above.

2. If also $\zeta\equiv 0$ on $\partial V\times[0,T]$, we may integrate by parts to deduce

$$\beta_t + \operatorname{div}_y\left(D_y\Phi'\left(\frac{1}{\beta}\right)\right)\equiv 0, \tag{2.7}$$

and this is the PDE in (2.4). Now drop the assumption that $\zeta = 0$ on the boundary and again integrate by parts:

$$\int_0^T\int_{\partial V}\frac{\partial}{\partial\nu}\left(\Phi'\left(\frac{1}{\beta}\right)\right)\zeta\, d\mathcal{H}^{n-1}\, dt = 0.$$

It follows that

$$\frac{\partial}{\partial\nu}\left(\Phi'\left(\frac{1}{\beta}\right)\right) = -\Phi''\left(\frac{1}{\beta}\right)\frac{\partial\beta}{\partial\nu}\equiv 0 \quad \text{on} \quad \partial V\times(0,T).$$

Since $\Phi''>0$, the proof is done. $\square$

2.3. Recovering the mapping u from . We next address the question of how to recover the mapping $\mathbf{u}$ from knowledge of β. One possibility is for each time t to try to find $x \mapsto \mathbf{u}(x,t)$ solving

$$\begin{cases} \beta(\mathbf{u}(x,t),t)\det D\mathbf{u}(x,t) \equiv 1 & \text{in } \bar{U}, \\ \mathbf{u}(\cdot,t) \in \operatorname{Diff}(\bar{U},\bar{V}), \end{cases} \tag{2.8}$$

where $\operatorname{Diff}(\bar{U},\bar{V})$ denotes the set of all diffeomorphisms of $\bar{U}$ onto $\bar{V}$. As we will discuss later in section 5, this approach works, provided U and V are convex sets.

However, there is a simpler construction available. First, define the new nonlinearity

$$\Psi(d) := d\Phi\left(\frac{1}{d}\right) \qquad (d > 0). \tag{2.9}$$

Then

$$\Psi'(d) = \Phi\left(\frac{1}{d}\right) - \frac{1}{d}\Phi'\left(\frac{1}{d}\right), \quad \Psi''(d) = \frac{1}{d^3}\Phi''\left(\frac{1}{d}\right); \tag{2.10}$$

and so $\Psi : (0,\infty) \to \mathbb{R}$ is convex.

Next, perform these calculations:

$$\begin{aligned} \mathbf{u}_t &= \operatorname{div}_x(\Phi'(\det D\mathbf{u})\det D\mathbf{u}(D\mathbf{u})^{-1}) \\ &= D_x(\Phi'(\det D\mathbf{u}))\cdot(\det D\mathbf{u}(D\mathbf{u})^{-1}) \\ &= \Phi''(\det D\mathbf{u})D_x(\det D\mathbf{u})\cdot(\det D\mathbf{u}(D\mathbf{u})^{-1}) \\ &= \Phi''\left(\frac{1}{\beta}\right)\frac{1}{\beta}D_x\left(\frac{1}{\beta}\right)(D\mathbf{u})^{-1} \\ &= -\Phi''\left(\frac{1}{\beta}\right)\frac{1}{\beta^3}D_x\beta(D\mathbf{u})^{-1} \\ &= -\Psi''(\beta)D_y\beta = -D_y\Psi'(\beta). \end{aligned} \tag{2.11}$$

This computation suggests that we fix a point $x \in \bar{U}$ and then solve the ODE

$$\begin{cases} \dot{\mathbf{y}}(t) = -\Psi''(\beta(\mathbf{y}(t),t))D\beta(\mathbf{y}(t),t) & \text{for } t > 0, \\ \mathbf{y}(0) = y = \mathbf{u}^0(x), \end{cases} \tag{2.12}$$

where $\dot{} = \frac{d}{dt}$. Then by uniqueness of solutions we have $\mathbf{u}(x,t) = \mathbf{y}(t)$ for $t \geq 0$.

3. Building diffeomorphisms. The formal calculations from the previous section done with, we turn now to building rigorously a smooth solution

$$\mathbf{u} : \bar{U} \times [0,\infty) \to \bar{V}$$

of our system

$$\begin{cases} \mathbf{u}_t = \operatorname{div}(\Phi'(\det D\mathbf{u})(\operatorname{cof} D\mathbf{u})^T) & \text{in } \bar{U} \times (0,\infty), \\ \mathbf{u} = \mathbf{u}^0 \quad \text{on } \bar{U} \times \{t = 0\}, \\ \mathbf{u}(\cdot,t) \in \operatorname{Diff}(\bar{U},\bar{V}) \end{cases} \tag{3.1}$$

under some additional assumptions.

3.1. Hypotheses. We require that the initial mapping $\mathbf{u}^0 : \bar{U} \to \bar{V}$ be a diffeomorphism, mapping ∂U onto ∂V. We write

$$\beta^0 := \det D\mathbf{v}^0 \tag{3.2}$$

for $\mathbf{v}^0 := (\mathbf{u}^0)^{-1}$ and assume that there exist positive constants $0 < C_1 \leq C_2$ such that

$$C_1 \leq \beta^0 \leq C_2 \quad \text{on } \bar{V}. \tag{H1}$$

We ask also that the following *compatibility condition* hold:

$$\frac{\partial \beta^0}{\partial \nu} = 0 \quad \text{on} \quad \partial V. \tag{H2}$$

Finally we require that Φ be smooth and convex on $(0, \infty)$, with the lower bound

$$\Phi''\left(\frac{1}{\beta}\right) > 0 \quad \text{for } C_1 \leq \beta \leq C_2. \tag{H3}$$

3.2. Solving PDE and ODE. In view of (H1), (H2), the initial/boundary-value problem

$$\begin{cases} \beta_t = \operatorname{div}\left(\Phi''\left(\frac{1}{\beta}\right)\frac{D\beta}{\beta^2}\right) & \text{in } V \times (0,\infty), \\ \frac{\partial \beta}{\partial \nu} = 0 & \text{on } \partial V \times [0,\infty), \\ \beta = \beta^0 & \text{on } \bar{V} \times \{t = 0\} \end{cases} \tag{3.3}$$

has a unique, smooth solution β, with

$$0 < C_1 \leq \beta \leq C_2 \qquad \text{in } \bar{V} \times [0, \infty). \tag{3.4}$$

Next, for each $y \in \bar{V}$, solve the ODE (2.12):

$$\begin{cases} \dot{\mathbf{y}}(t) = -\Psi''(\beta(\mathbf{y}(t),t))D\beta(\mathbf{y}(t),t) & \text{for } t > 0, \\ \mathbf{y}(0) = y. \end{cases} \tag{3.5}$$

We write $\mathbf{y}(t) = \mathbf{y}(t, y)$ to display dependence on the initial point y.

THEOREM 3.1. (i) *For each given $x \in \bar{U}$, the* ODE (3.5) *has a unique solution $\mathbf{y} : [0, \infty) \to \bar{V}$, existing for all times $t \geq 0$.*

(ii) *If $y \in \partial V$, then $\mathbf{y}(t) \in \partial V$ for all times $t \geq 0$.*

(iii) *For each $t \geq 0$, the mapping*

$$\mathbf{u}(x,t) := \mathbf{y}(t, \mathbf{u}^0(x)) \qquad (x \in \bar{U},\ t \geq 0) \tag{3.6}$$

is a smooth diffeomorphism from $\bar{U}$ to $\bar{V}$, mapping ∂U onto ∂V.

Proof. Since $\frac{\partial \beta}{\partial \nu} = 0$ on ∂V, $D\beta$ is tangent to ∂V and consequently the flow does not leave $\bar{V}$. In particular, if $\mathbf{u}^0(x) \in \partial V$, then $\mathbf{x}(t) \in \partial V$ for times $t \geq 0$.

Assertion (iii) is standard. ∎

Define $\mathbf{u} : \bar{U} \times [0, \infty) \to \bar{V}$ by (3.6) and set $\mathbf{v}(\cdot, t) := \mathbf{u}^{-1}(\cdot, t)$ for each time $t \geq 0$.

THEOREM 3.2. (i) *We have*

$$\beta \equiv \det D\mathbf{v}. \tag{3.7}$$

(ii) *Furthermore,* $\mathbf{u}$ *solves the system of* PDEs (2.1)*, and the mapping*

$$t \mapsto \int_U \Phi(\det D\mathbf{u})(x,t)\,dx$$

is nonincreasing.

Proof. 1. As before, set $\alpha = \det D\mathbf{u}$, $\alpha = \alpha(x,t)$. Then

$$\alpha_t = \alpha D_x \mathbf{u}_t (D\mathbf{u})^{-1}. \tag{3.8}$$

Now

$$\mathbf{u}_t = -D_y \Psi'(\beta)$$

and so

$$D_x \mathbf{u}_t = -D_y^2 \Psi'(\beta)(D_x \mathbf{u}).$$

Hence

$$\alpha_t = -\alpha \Delta_y \Psi'(\beta). \tag{3.9}$$

Next, regarding $\beta = \beta(\mathbf{u},t)$ as a function of (x,t), we have

$$\begin{aligned}
(\alpha\beta)_t &= \alpha_t \beta + \alpha\beta_t + \alpha D_y\beta \cdot \mathbf{u}_t \\
&= -\alpha\beta\Delta_y\Psi'(\beta) + \alpha \operatorname{div}(\Psi''(\beta)\beta D_y\beta) - \alpha D_y\beta \cdot (\Psi''(\beta) D_y \beta) \\
&= 0.
\end{aligned}$$

Since $\alpha\beta \equiv 1$ at $t = 0$, we deduce that

$$\beta = \alpha^{-1} = \det D\mathbf{v}.$$

2. We have shown that $\beta \equiv \det D\mathbf{v}$, where $\mathbf{v} = \mathbf{u}^{-1}$ and $\mathbf{u}$ is defined by (3.6). We then return to the computation (2.11) to deduce that

$$\mathbf{u}_t = \dot{\mathbf{x}} = -\Psi''(\beta) D\beta = \operatorname{div}(\Phi'(\det D\mathbf{u})(\operatorname{cof} D\mathbf{u})^T). \tag{3.10}$$

Finally let us calculate

$$\begin{aligned}
\frac{d}{dt}\int_U \Phi(\det D\mathbf{u})\,dx &= \frac{d}{dt}\int_V \Phi\left(\frac{1}{\beta}\right)\beta\,dy \\
&= \int_V \left(\Phi\left(\frac{1}{\beta}\right) - \frac{1}{\beta}\Phi'\left(\frac{1}{\beta}\right)\right)\beta_t\,dy \\
&= \int_V \Psi'(\beta)\operatorname{div}(\beta\Psi''(\beta)D\beta)\,dy \\
&= -\int_V \Psi''(\beta)^2\beta|D\beta|^2\,dy \le 0. \qquad \square
\end{aligned}$$

4. Some variants.

4.1. More general nonlinearities. Our methods extend with little difficulty to the functional

$$I[\mathbf{v}] := \int_U \Phi(f(\mathbf{v})\det D\mathbf{v})\,dx \tag{4.1}$$

for Φ as before and $f : \bar{V} \to (0,\infty)$.

Euler–Lagrange equation. The corresponding Euler–Lagrange equation is

$$-\operatorname{div}(\Phi'(f \det D\mathbf{u}) f (\operatorname{cof} D\mathbf{u})^T) + \Phi'(f \det D\mathbf{u})(\det D\mathbf{u}) Df = 0,$$

which simplifies to read

$$\Phi''(f \det D\mathbf{u}) D(f \det D\mathbf{u}) f (\operatorname{cof} D\mathbf{u})^T = 0. \tag{4.2}$$

As in section 1.1 this implies that

$$f(\mathbf{u}) \det D\mathbf{u} \equiv C \quad \text{within} \quad U$$

for some constant C.

A gradient flow. The evolution associated with (4.1) is

$$\mathbf{u}_t - \operatorname{div}(\Phi'(f \det D\mathbf{u}) f (\operatorname{cof} D\mathbf{u})^T) + \Phi'(f \det D\mathbf{u})(\det D\mathbf{u}) Df = 0, \tag{4.3}$$

plus initial and boundary conditions.

As before, assume $\mathbf{v} := \mathbf{u}^{-1}$ exists and write

$$\beta := \det D\mathbf{v}.$$

THEOREM 4.1. *We have*

$$\begin{cases} \beta_t = -\operatorname{div}\left(\Phi''\left(\frac{f}{\beta}\right) f D\left(\frac{f}{\beta}\right)\right) & \text{in } V \times (0,\infty), \\ \frac{\partial}{\partial \nu}\left(\frac{f}{\beta}\right) = 0 & \text{on } \partial V \times (0,\infty). \end{cases} \tag{4.4}$$

Proof. 1. Fix any time $T > 0$ and select a smooth function $\zeta : \bar{V} \times [0,T] \to \mathbb{R}$ satisfying (2.5).

Then

$$\begin{aligned} &\int_0^T \int_V \beta\zeta_t + D_y\left(\Phi'\left(\frac{f}{\beta}\right)\right) \cdot D_y\zeta f \, dydt \\ &= \int_0^T \int_U \left[\beta(\mathbf{u},t)\zeta_t(\mathbf{u},t) + D_x\left(\Phi'\left(\frac{f}{\beta}\right)\right)(D\mathbf{u})^{-1} \cdot D_y\zeta f\right] \frac{dx}{\beta(\mathbf{u},t)} \, dt \\ &= \int_0^T \int_U \frac{\partial}{\partial t}(\zeta(\mathbf{u},t)) - D_y\zeta \cdot \mathbf{u}_t + D_x\left(\Phi'\left(\frac{f}{\beta}\right)\right)\frac{(D\mathbf{u})^{-1}}{\beta} \cdot D_y\zeta f \, dxdt \\ &= -\int_0^T \int_U D_y\zeta \cdot \left[\mathbf{u}_t - D_x\left(\Phi'\left(\frac{f}{\beta}\right)\right)\frac{(D\mathbf{u})^{-1}}{\beta} f\right] \, dxdt. \end{aligned} \tag{4.5}$$

But according to (4.3), we have

$$\begin{aligned} \mathbf{u}_t &= D(\Phi'(f \det D\mathbf{u})) f (\det D\mathbf{u})(D\mathbf{u})^{-1} \\ &= D\left(\Phi'\left(\frac{f}{\beta}\right)\right)\frac{f}{\beta}(D\mathbf{u})^{-1}. \end{aligned} \tag{4.6}$$

Consequently the expression within the square brackets in the last term of (4.5) vanishes. ∎

4.2. "Filling up" the target domain. An interesting variant of our construction is as follows. Select $\mathbf{u}^0 : \bar{U} \to \bar{W}_0$ to be a diffeomorphism, where $W_0 \subset\subset V$ is given. We will choose Φ and $\mathbf{u}$ so that

$$\begin{cases} W(t) := \mathbf{u}(U,t) & (t \geq 0), \\ W(0) = W_0 \end{cases} \tag{4.7}$$

expands to "fill up" the target V in finite time.

For this, let us take $m > 0$ and

$$\Phi(d) := \begin{cases} \frac{1}{m} d^{-m} & (d > 0), \\ +\infty & (d \leq 0). \end{cases}$$

Therefore

$$\Psi(d) = d\Phi\left(\frac{1}{d}\right) = \frac{1}{m} d^{m+1}$$

for $d > 0$. Then β solves the porous medium equation

$$\beta_t = \operatorname{div}(\Psi''(\beta)\beta D\beta) = \Delta(\beta^{m+1}).$$

5. Connections with optimal mass transfer problems. As noted in the introduction, the time-step minimization method (1.4) and (1.5) provides an extremely natural approximation method, but one which we have not been able to prove converges. This section recalls more about this procedure, to highlight the connections with Monge–Kantorovich mass transfer theory.

We are primarily motivated by Otto [O] and Jordan, Kinderlehrer, and Otto [J-K-O]. The novelty of Otto's paper [O] was to interpret (5.8) as a gradient flux of the "entropy" $S(\beta) := \int_V \Psi(\beta) dy$ with respect to the Wasserstein distance.

5.1. Time-step approximations. Assume for this section that U and V are two bounded, open, *convex* sets with smooth boundaries.

We discuss a time-discrete algorithm for the flow

$$\begin{cases} \mathbf{u}_t = \operatorname{div}(DF(D\mathbf{u})), \\ \mathbf{u}(\cdot,0) = \mathbf{u}_0, \\ \mathbf{u}(\cdot,t) \in \operatorname{Diff}(\bar{U},\bar{V}), \end{cases} \tag{5.1}$$

where, as before,

$$F(P) = \begin{cases} \Phi(\det P), & \det P > 0, \\ +\infty, & \det P \leq 0. \end{cases}$$

The system (5.1) is a gradient flux of the functional $I[\cdot]$ with respect to the L^2-metric. In section 2 we have shown that (5.1) is related to (1.16), which, as we will recall below, is the gradient flow governed by $\int_V \Psi(\beta)\, dy$ with respect to the Wasserstein distance. The algorithm which we discuss is another way to view that relation.

A discrete-time approximation. First, let us fix a time-step size $h > 0$. We introduce the implicit scheme of recursively finding $\mathbf{u}_{k+1}$ to solve

$$\begin{cases} \frac{\mathbf{u}_{k+1}-\mathbf{u}_k}{h} = \operatorname{div}(DF(D\mathbf{u}_{k+1})), \\ \mathbf{u}_{k+1} \in \operatorname{Diff}(\bar{U},\bar{V}), \end{cases} \tag{5.2}$$

given $\mathbf{u}_k$. More precisely, set

$$I_k[\mathbf{v}] := \frac{1}{2}\int_U |\mathbf{v}-\mathbf{u}_k|^2\,dx + h\int_U F(D\mathbf{v})\,dx. \tag{5.3}$$

We intend to find $\mathbf{u}_{k+1}$ to be the unique minimizer of

$$\min_{\mathbf{v}}\{I_k[\mathbf{v}] \mid \mathbf{v}\in \mathrm{Diff}(\bar{U},\bar{V})\}. \tag{5.4}$$

Changing variables. Since our nonlinearity F is neither coercive nor convex, standard calculus of variations methods do not apply. However, recent papers by Gangbo and Van der Putten [G-VP] and Maroofi [Ma] demonstrate how to exploit the special structure of $F(P)=\Phi(\det P)$ to find minimizers.

Indeed, if we apply a change of variables $y=\mathbf{u}(x)$ and set $\beta := \det(D\mathbf{u}^{-1})$, $\beta_k := \det(D\mathbf{u}_k^{-1})$, we discover that

$$I_k[\mathbf{v}] = \frac{1}{2}\int_V |y-\mathbf{u}_k(\mathbf{v}^{-1}(y))|^2 dy + h\int_V \Psi(\beta)\,dy.$$

Consequently

$$\begin{aligned}
&\min_{\mathbf{v}\in\mathrm{Diff}(\bar{U},\bar{V})} I_k[\mathbf{v}]\\
&= \inf_\beta\left\{h\int_V \Psi(\beta)\,dy + \inf_{\mathbf{v}}\left\{\frac{1}{2}\int_V |y-\mathbf{u}_k(\mathbf{v}^{-1}(y))|^2 dy \mid \beta=\det(D\mathbf{v}^{-1})\right\}\right\}\\
&= \inf_\beta\left\{h\int_V \Psi(\beta)\,dy + \inf_{\mathbf{w}}\left\{\frac{1}{2}\int_V |y-\mathbf{w}(y)|^2 dy \mid \beta_k=\beta(\mathbf{w})\det D\mathbf{w}\right\}\right\}\\
&= \inf_\beta\left\{h\int_V \Psi(\beta)\,dy + W_2^2(\beta_k,\beta)\right\},
\end{aligned} \tag{5.5}$$

where W_2, the *Wasserstein distance* between two Borel measures μ and ν, is defined as

$$W_2^2(\mu,\nu) := \frac{1}{2}\inf_{\gamma\in\Gamma(\mu,\nu)}\iint |x-y|^2 d\gamma(x,y).$$

Here $\Gamma(\mu,\nu)$ is the set of Borel measures γ on $\mathbb{R}^{2n}$ that have μ and ν as their marginals. The notation $W_2^2(\beta_k,\beta)$ means that we have identified β with the measure whose density is β.

We assume for $k=0$ that

$$\int_V \beta_0\,dy = 1,$$

where $\beta_0=\det D\mathbf{u}_0^{-1}$. This reduces the last three problems in (5.5) to minimization problems over $\mathcal{P}_a(V)$, the set of probability densities supported in V.

Define the new functional

$$J_k(\beta) := W_2^2(\beta,\beta_k) + h\int_V \Psi(\beta)\,dy. \tag{5.6}$$

Now $W_2^2(\beta_k,\cdot)$ is convex and is weakly-$*$ lower semicontinuous. Since Ψ is strictly convex, we see also that $\beta\to\int_V\Psi(\beta)\,dy$ is a strictly convex functional of β and is weakly-$*$ lower semicontinuous on subsets of L^1 that are weakly-$*$ compact. Consequently, the minimization problem

$$\inf_{\beta \in \mathcal{P}_a(V)} J_k(\beta) \tag{5.7}$$

has a unique solution β_{k+1}.

5.2. Time-step approximations for . This subsection quickly reviews a time-discrete algorithm based on the Wasserstein distance for solving

$$\begin{cases} \beta_t = \operatorname{div}\Big(\beta D[\Psi'(\beta)]\Big), \\ \beta(\cdot, 0) = \beta_0. \end{cases} \tag{5.8}$$

Let us now deal with the following nonlinear problem appearing in (5.5), where we replace β by β_{k+1}. We study the minimization problem

$$\inf_{\mathbf{v}} \left\{ \int_V |y - \mathbf{v}(y)|^2 dy \mid \beta_k = \beta_{k+1}(\mathbf{v}) \det D\mathbf{v} \right\}, \tag{5.9}$$

which, thanks to the Monge–Kantorovich theory, is known to admit a unique minimizer $\mathbf{v}_{k+1}$ (see Brenier [B]). Furthermore, $\mathbf{v}_{k+1}$ is the gradient of a convex function $\psi_{k+1} : \bar{V} \to \mathbb{R}$, satisfying the Monge–Ampere problem

$$\beta_k = \beta_{k+1}(D\psi_{k+1}) \det D^2\psi_{k+1}, \quad D\psi_{k+1}(\bar{V}) = \bar{V} \tag{5.10}$$

in the sense that

$$D\psi_{k+1} : \bar{V} \to \bar{V} \text{ a.e.} \quad \text{and} \quad \int_V f(D\psi_{k+1})\beta_k \, dx = \int_V f\beta_{k+1} \, dy \tag{5.11}$$

for all $f \in C(\mathbb{R}^n)$. Equivalently, if ϕ_{k+1} is the Legendre transform of ψ_{k+1}, then

$$D\phi_{k+1} : \bar{V} \to \bar{V} \text{ a.e.} \qquad \text{and} \qquad \int_V g(D\phi_{k+1})\beta_{k+1} \, dy = \int_V g\beta_k \, dx \tag{5.12}$$

for all $g \in C(\mathbb{R}^n)$. We write that

$$(D\psi_{k+1})_{\#}\beta_k = \beta_{k+1}, \quad (D\phi_{k+1})_{\#}\beta_{k+1} = \beta_k,$$

the symbol # denoting push-forward. Agueh [A] has shown that

$$C_1 \le \beta_{k+1} \le C_2, \tag{5.13}$$

provided

$$C_1 \le \beta_k \le C_2 \tag{5.14}$$

for constants $0 < C_1 \le C_2$

The Euler–Lagrange equations of (5.7) read

$$D\phi_{k+1}(y) = y + hD[\Psi'(\beta_{k+1}(y))], \tag{5.15}$$

and we conclude from (5.15) that

$$\beta_{k+1}(y) = (\Psi^*)'\left(\left(\phi_{k+1}(y) - \frac{|y|^2}{2}\right)/h\right), \tag{5.16}$$

where Ψ^* is the Legendre transform of Ψ.

Assume that $\beta_k \in C^{l,\alpha}(\bar{V})$ for some $\alpha > 0$ and some integer $l \geq 0$. By (5.16), $\beta_{k+1} \in W^{1,\infty}(V) \subset C^{0,\alpha}(\bar{V})$. Regularity theory for the Monge–Ampere equations (see [C1], [C2], [C3], [C4]) and (5.10) imply that $\psi_{k+1}, \phi_{k+1} \in C^{2,\alpha}(\bar{V})$. This and (5.16) demonstrate that $\beta_{k+1} \in C^{2,\alpha}(\bar{V})$. Thus

$$\gamma_{k+1} := D\phi_{k+1} \circ D\phi_k \circ \cdots \circ D\phi_1 \in C^{l+1,\alpha}(\bar{V}, \bar{V}).$$

The map

$$\mathbf{u}_{k+1} = \gamma_{k+1} \circ \mathbf{u}_0$$

is then the unique solution to (5.2), and $\mathbf{u}_{k+1} \in C^{l+1,\alpha}(\bar{U}, \bar{V})$ if $\mathbf{u}_0 \in C^{l+1,\alpha}(\bar{U}, \bar{V})$.

We record next that the time-step approximations converge as $h \to 0$.

THEOREM 5.1. *For $h > 0$, inductively define β_{k+1} to be the unique minimizer of $J_k[\cdot]$ over $\mathcal{P}_a(V)$. Set*

$$\beta^h(y,t) = \begin{cases} \beta_0(y) & \text{if} \quad t = 0, \\ \beta_k(y) & \text{if} \quad t \in ((k-1)h, kh]. \end{cases}$$

Fix $T > 0$ and assume that $T = Mh$ for an integer $M > 0$.

Then the following hold:

(i) *For each test function $\eta \in C_c^2$, we have*

$$\left| \int_{V_T} \partial_t^h \eta (\beta^h - \beta_0)\, dxdt + \int_{V_T} \operatorname{div}\Big(\beta^h D[\Psi'(\beta^h)]\Big)\, dxdt \right| \leq C_\eta h,$$

where $\partial_t^h \eta(x,t) = (\eta(x,t+h) - \eta(x,t))/h$ and $V_T = V \times (0,T)$.

(ii) *There exists a subsequence $\{h_m\}_{m=1}^\infty$ converging to 0 and $\beta \in L^1(V_T)$ such that $\{\beta^{h_m}\}_{m=1}^\infty$ converges to β. Furthermore, β satisfies the parabolic equation (5.8).*

5.3. Time-step approximations for u. Finally, we return to the approximation scheme (5.2) and consider the convergence problem as $h \to 0$.

We first record some uniform estimates.

THEOREM 5.2. *Fix $h > 0$ and inductively define $\mathbf{u}_{k+1}$ to be the unique minimizer of $I_k[\cdot]$ over $\operatorname{Diff}(\bar{U}, \bar{V})$. Define*

$$\mathbf{u}^h(\cdot,t) = \begin{cases} \mathbf{u}_0(\cdot) & \text{if} \quad t = 0, \\ \mathbf{u}_k(\cdot) & \text{if} \quad t \in ((k-1)h, kh]. \end{cases}$$

Fix $T > 0$ and assume that $T = Mh$ for an integer $M > 0$. Set $U_T = U \times (0,T)$.

Then the following hold:

(i) *For each $t \in [0,T]$ we have that $\mathbf{u}^h(\cdot,t) \in \operatorname{Diff}(\bar{U}, \bar{V}) \cap C^{l+1,\alpha}(\bar{U}, \mathbb{R}^n)$ and there are constants $C_1, C_2 > 0$ depending only on $\mathbf{u}_0$ such that*

$$C_1 \leq \det D(\mathbf{u}^h)^{-1} \leq C_2.$$

(ii) *There exists a constant $C > 0$, depending only on $\mathbf{u}_0$, such that*

$$\sum_{k=0}^{M-1} \int_U |\mathbf{u}_{k+1} - \mathbf{u}_k|^2\, dx \leq Ch.$$

(iii) *For each test function* $\mathbf{v} \in C^2$, *we have*

(5.17)
$$\left| \int_{U_T} \mathbf{u}^h \cdot \mathbf{v}_t - DF(D\mathbf{u}^h) : D\mathbf{v}\, dxdt + \int_U \mathbf{u}_0 \cdot \mathbf{v}(\cdot, 0)\, dx \right| \leq \frac{h}{2} C\sqrt{T} ||\mathbf{v}_t||_{L^\infty(U_T)}.$$

Proof. 1. Set $\beta_0 = \det D\mathbf{u}_0^{-1}$. Since $\mathbf{u}_0 \in \mathrm{Diff}(\bar{U}, \bar{V})$ we have that

$$0 < C_1 := \min_{\bar{V}} \beta_0, C_2 := \max_{\bar{V}} \beta_0 < +\infty.$$

According to the discussion above, we can choose inductively $\mathbf{u}_{k+1}$ to be the unique minimizer of I_k over $\mathrm{Diff}(\bar{U}, \bar{V})$.

2. The inequality $I_k(\mathbf{u}_{k+1}) \leq I_k(\mathbf{u}_k)$ implies that

$$\sum_{k=0}^{M-1} I_k(\mathbf{u}_{k+1}) \leq \sum_{k=0}^{M-1} I_k(\mathbf{u}_k).$$

Therefore

(5.18)
$$\frac{1}{2} \sum_{k=0}^{M-1} \int_V |\mathbf{u}_{k+1} - \mathbf{u}_k|^2 dx \leq h \int_U \Phi(\det D\mathbf{u}_0) - \Phi(\det D\mathbf{u}_M) dx \leq 2h|U| \max_{[\frac{1}{C_2}, \frac{1}{C_1}]} |\Phi|.$$

This proves (ii).

3. Suppose now that $\mathbf{v} \in C^2$, and set $t_k = kh$, $\mathbf{v}_k = \mathbf{v}(\cdot, kh)$, and $U_k = U \times (t_k, t_{k+1})$. Then

$$\int_{U_T} \mathbf{u}^h \cdot \mathbf{v}_t - DF(D\mathbf{u}^h) : D\mathbf{v}\, dxdt = \sum_{k=0}^{M-1} \int_{U_k} \mathbf{u}_{k+1} \cdot \mathbf{v}_t\, dxdt - DF(D\mathbf{u}_{k+1}) : D\mathbf{v}\, dxdt.$$

We recall that $(\mathbf{u}_{k+1} - \mathbf{u}_k)/h = \mathrm{div}(DF(D\mathbf{u}_{k+1}))$ and continue to calculate that

$$\begin{aligned}
\int_{U_T} \mathbf{u}^h \cdot \mathbf{v}_t - DF(D\mathbf{u}^h) : D\mathbf{v}\, dxdt &= \sum_{k=0}^{M-1} \int_{U_k} \mathbf{u}_{k+1} \cdot \mathbf{v}_t + \left(\frac{\mathbf{u}_{k+1} - \mathbf{u}_k}{h} \right) \cdot \mathbf{v}\, dxdt \\
&= \sum_{k=0}^{M-1} \int_U \mathbf{u}_{k+1} \cdot (\mathbf{v}_{k+1} - \mathbf{v}_k)\, dx \\
&\quad + \int_U (\mathbf{u}_{k+1} - \mathbf{u}_k) \cdot \mathbf{v}_k\, dx \\
&\quad + \sum_{k=0}^{M-1} \int_U \left(\frac{\mathbf{u}_{k+1} - \mathbf{u}_k}{h} \right) \cdot \left(\int_{t_k}^{t_{k+1}} \mathbf{v} - \mathbf{v}_k\, dt \right) dx \\
&= \int_U \mathbf{u}_M \cdot \mathbf{v}_M - \mathbf{u}_0 \cdot \mathbf{v}_0\, dx \\
&\quad + \sum_{k=0}^{M-1} \int_U \left(\frac{\mathbf{u}_{k+1} - \mathbf{u}_k}{h} \right) \cdot \left(\int_{t_k}^{t_{k+1}} \mathbf{v} - \mathbf{v}_k\, dt \right) dx.
\end{aligned}$$

Taking into account $\mathbf{v}_M = \mathbf{v}(T) = 0$ and

$$\left|\int_{t_k}^{t_{k+1}} \mathbf{v} - \mathbf{v}_k \, dt\right| \leq \frac{h^2}{2} \max_{U_T} |\mathbf{v}_t|,$$

we conclude that

$$\begin{aligned}
&\left|\int_{U_T} \mathbf{u}^h \cdot \mathbf{v}_t - DF(D\mathbf{u}^h) : D\mathbf{v} \, dxdt + \int_U \mathbf{u}_0 \cdot \mathbf{v}(\cdot, 0) \, dx\right| \\
&\qquad \leq \frac{h}{2} \, ||\mathbf{v}_t||_{L^\infty} \sum_{k=0}^{M-1} \int_U |\mathbf{u}_{k+1} - \mathbf{u}_k| \, dx \\
&\qquad \leq \frac{h}{2} \, ||\mathbf{v}_t||_{L^\infty(U_T)} \left(\sum_{k=0}^{M-1} \int_U |\mathbf{u}_{k+1} - \mathbf{u}_k|^2 \, dx\right)^{\frac{1}{2}} M^{\frac{1}{2}}. \qquad (5.19)
\end{aligned}$$

We combine (5.18) and (5.19) to finish up the proof of (iii). ☐

This theorem provides some uniform estimates, but it remains an unsolved problem to show that as $h \to 0$, the approximation $\mathbf{u}^h$ converges somehow to a solution $\mathbf{u}$ of (1.3). One particular issue is that we do not know if the gradients $D\mathbf{u}^h$ converge strongly in L^2.

Our belief is that although the scheme (5.2), (5.3), and (5.4) is obviously extremely natural, we do not currently know how fully to exploit the minimization structure. We have here a problem in the "time-dependent calculus of variations," but we do not have enough experience to understand, for instance, the proper choices of comparison functions to employ in our variational principles. The direct PDE and ODE methods in sections 2 and 3 provide a way around this difficulty for the special case of the nonlinearity (1.6).

REFERENCES

[A] M. Agueh, *Existence of Solutions to Degenerate Parabolic Equations via the Monge–Kantorovich Theory*, Ph.D. dissertation, School of Mathematics, Georgia Institute of Technology, Atlanta, GA, 2002.

[B] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.

[C1] L. A. Caffarelli, *Some regularity properties of solutions of Monge-Ampère equation*, Comm. Pure Appl. Math., 64 (1991), pp. 965–969.

[C2] L. A. Caffarelli, *The regularity of mappings with a convex potential*, J. Amer. Math. Soc., 5 (1992), pp. 99–104.

[C3] L. A. Caffarelli, *Boundary regularity of maps with convex potentials*, Comm. Pure Appl. Math., 44 (1992), pp. 1141–1151.

[C4] L. A. Caffarelli, *Boundary regularity of maps with convex potentials*—II, Ann. of Math. (2), 144 (1996), pp. 453–496.

[C-G] L. A. Caffarelli and C. E. Gutierrez, *The Monge-Ampere Equation*, Birkhäuser Boston, Boston, 2001.

[D] B. Dacorogna, *A relaxation theorem and its application to the equilibrium of gases*, Arch. Rational Mech. Anal., 77 (1981), pp. 359–386.

[E] L. C. Evans, *An unusual minimization principle for parabolic gradient flows*, SIAM J. Math. Anal., 27 (1996), pp. 1–4.

[G-VP] W. Gangbo and R. Van der Putten, *Uniqueness of equilibrium configurations in solid crystals*, SIAM J. Math. Anal., 32 (2000), pp. 465–492.

[J-K-O] R. Jordan, D. Kinderlehrer, and F. Otto, *The variational formulation of the Fokker–Planck equation*, SIAM J. Math. Anal., 29 (1998), pp. 1–17.

[Ma] H. Maroofi, *Applications of the Monge–Kantorovich Theory*, Ph.D. dissertation, School of Mathematics, Georgia Institute of Technology, Atlanta, GA, 2002.

[O] F. Otto, *The geometry of dissipative evolution equations: The porous medium equation*, Comm. Partial Differential Equations, 26 (2001), pp. 101–174.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 752–776

IMMEDIATE REGULARIZATION AFTER BLOW-UP*

MAREK FILA†, HIROSHI MATANO‡, AND PETER POLÁČIK§

Abstract. We study solutions of some supercritical parabolic equations which blow up in finite time but continue to exist globally in the weak sense. We show that the minimal continuation becomes regular immediately after the blow up time, and if it blows up again, it can only do so finitely many times.

Key words. nonlinear heat equation, blow up, regularity

AMS subject classifications. 35K57, 35B65

DOI. 10.1137/040613299

1. Introduction.

We consider the problem

$$\text{(P)}\qquad \begin{cases} u_t = \Delta u + f(u), & x \in B_1,\ t > 0, \\ u = 0, & x \in \partial B_1,\ t > 0, \\ u(x,0) = u_0(x)\ (= U_0(|x|)\,), & x \in B_1, \end{cases}$$

where $B_1 = \{x \in \mathbb{R}^N : |x| < 1\}$, $U_0 \in C([0,1])$, $U_0 \geq 0$ with $U_0(1) = 0$, and either

$$f(u) = \lambda e^u, \quad \lambda > 0, \quad 3 \leq N \leq 9 \tag{1.1}$$

or

$$f(u) = u^p, \quad \frac{N+2}{N-2} < p < p^*, \quad N \geq 3, \tag{1.2}$$

where

$$p^* = \begin{cases} \infty & \text{if} \quad N \leq 10, \\ 1 + \frac{4}{N-4-2\sqrt{N-1}} & \text{if} \quad N > 10. \end{cases}$$

In the case of the exponential nonlinearity (1.1), we shall further assume that $U_0(r)$ is a nonincreasing function on $[0,1]$.

We shall show that if a solution blows up in a finite time $t = T < \infty$ but continues to exist as a weak solution for $t > T$, then this extended solution becomes regular immediately after the blow up time T, that is, it possesses no singularity in the time

*Received by the editors August 11, 2004; accepted for publication (in revised form) March 18, 2005; published electronically November 9, 2005. The main results of this paper were previously announced in a contribution to Proceedings of Equadiff 2003. This research was partially supported by JSPS grant 4023.

http://www.siam.org/journals/sima/37-3/61329.html

†Institute of Applied Mathematics, Comenius University, 842 48 Bratislava, Slovakia (fila@fmph.uniba.sk). The research of this author was partially supported by the 21st Century COE Program at Graduate School of Mathematical Sciences, the University of Tokyo, as well as by VEGA grant 1/0259/03 and by the European Community's Human Potential Programme under contract HPRN-CT-2002-00274, Fronts – Singularities.

‡Graduate School of Mathematical Sciences, University of Tokyo, 3-8-1 Komaba, Tokyo 153-8914, Japan (matano@ms.u-tokyo.ac.jp).

§School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (polacik@math.umn.edu). The research of this author was supported in part by NSF grant DMS-0400702.

interval $T < t < T^*$ for some $T^* \in (T, \infty]$. Here, by an extended solution we mean the so-called minimal continuation, whose meaning will be clarified later.

Let us now give a more detailed description of the result together with some history of the problem. Global unbounded weak solutions of (P) with $f(u) = u^p$, $p \geq \frac{N+2}{N-2}$, were discovered in [16]. These solutions are characterized as the limit of an increasing sequence of global classical solutions $0 < u_1 < u_2 < u_3 < \cdots$ such that each u_k belongs to the domain of attraction of the stable stationary solution $u = 0$ (that is, $u_k \to 0$ as $t \to \infty$) and such that $\lim_{k\to\infty} u_k$ lies on the boundary of this domain of attraction.

The monotonicity of the sequences and the standard Kaplan-type estimate for the approximating classical solutions yield uniform bounds on certain integrals; hence the limit functions are indeed time-global weak solutions. Moreover, these weak solutions are necessarily unbounded on the time interval $[0, \infty)$, since otherwise they would remain classical for all $t > 0$ and converge to some positive stationary solution as $t \to \infty$, but the assumption $p \geq \frac{N+2}{N-2}$ and the Pohozaev identity imply nonexistence of positive stationary solutions. Consequently, one of the following alternatives occurs for each of these global weak solutions:

(a) the solution blows up in finite time;

(b) the solution remains smooth for all $t > 0$ and tends to infinity as $t \to \infty$.

For a long time it was not known which of the two alternatives really occurs. Much later Galaktionov and Vázquez [4] concluded that (a) is always the case, provided that

$$\frac{N+2}{N-2} < p \left(< 1 + \frac{6}{N-10} \quad \text{if} \quad N > 10 \right).$$

It follows from [4] and a recent result of Mizoguchi [15] that (a) holds for all $p > (N+2)/(N-2)$, $N > 2$. By *blow up in finite time* we mean that there is $T \in (0, \infty)$ such that

$$\lim_{t \to T} \|u(\cdot, t)\|_{L^\infty(B_1)} = \infty,$$

and we call T the *blow up time* of solution u. The above-mentioned global weak solutions thus have continuation beyond the blow up time in a certain sense.

Note that such continuation is not possible if $1 < p \leq \frac{N+2}{N-2}$. Indeed it is known that in this range of p every blow up is complete in the sense that the minimal (proper) continuation for t larger than the blow up time T is infinite everywhere in $B_1 \times (T, \infty)$. See [1, 4] for details. Note also that the same is true for any $p > 1$ if $u_t \geq 0$.

In the case of exponential nonlinearity (1.1), global unbounded weak solutions were constructed in [9] for a certain range of λ and later in [3] for a larger range of λ. It was also proved in [3] that global unbounded weak solutions blow up in finite time.

For both (1.1) and (1.2), it is still an open question whether the weak continuation beyond the blow up time is unique or not. However, one can define uniquely the so-called minimal continuation, and this is what we study in the present paper.

Now, given a global weak solution u, we define the set of *regular time moments* by

$$\mathcal{R} := \{t_0 > 0 : \ u \text{ is a classical solution on some time interval around } t = t_0\}$$

and the *set of singular time moments* by $\mathcal{S} = (0, \infty) \setminus \mathcal{R}$. By Lemma 2.6 in subsection

2.1, we see that $\mathcal{S} = \mathcal{B}$, where

$$\mathcal{B} := \left\{ t_0 > 0 \,:\, \limsup_{t \to t_0} \|u(\cdot, t)\|_{L^\infty(B_1)} = \infty \right\}.$$

Our main goal is to show that $\mathcal{B}$ is a finite set. This in particular implies that the solution recovers smoothness immediately after blow-up, and it remains smooth until $t = \infty$ or until the next blow-up occurs. We also give an example of a limit L^1-solution for which $\mathcal{B}$ is a singleton.

Previously some examples of "peaking solutions"—that is, solutions which blow up at $t = T$ and become smooth for $t > T$— have been known for the Cauchy problem in $\mathbb{R}^N$ (see [9, 4, 12]). In [9] and [4], examples of peaking solutions are constructed by simply gluing backward self-similar solutions (i.e., self-similar solutions that are defined for $-\infty < t < T$ and blow up at $t = T$) and forward self-similar solutions (i.e., self-similar solutions that are defined for $T < t < \infty$ and blow up at $t = T$) sharing the same blow up profile at the blow up time T. However, this method does not work for problems in bounded domains since self-similar solutions do not satisfy reasonable boundary conditions on a fixed boundary. In [13, 14], examples of solutions with multiple blow up time are constructed for the power case (1.2). These earlier works only give some examples of peaking solutions, and it has not been known whether or not global weak solutions in general behave like peaking solutions—more precisely, whether or not immediate regularization always occurs after blow-up. Our result in the present paper shows that every minimal continuation beyond blow-up has this property.

It should be noted that our result does not follow from standard parabolic estimates. Indeed, once a solution u blows up, say at $t = T$, then its blow up profile $u(x, T)$ may no longer belong to the function space in which problem (P) is well-posed; therefore, a standard bootstrap argument does not improve the regularity so much. See Remark 3.4 for details.

This paper is organized as follows. In section 2, we define limit L^1-solutions and derive various estimates for these solutions. Among other things we show that limit L^1-solutions belong to $C([0, \infty); H^1(B_1))$.

In section 3, we state our main result (Theorem 3.1). We then prove the main theorem in sections 4 and 5. The proof relies heavily on the zero-number properties of parabolic equations. Note that, as we are dealing with solutions with singularities, the zero-number is no longer monotone nonincreasing but a slightly weaker property holds (see Lemma 4.1 and Remark 4.3).

In section 6, we consider the case of exponential nonlinearity (1.1) and prove the existence of a peaking solution that blows up exactly once and remains smooth thereafter (Theorem 6.1). This result is proved by showing that a singular heteroclinic connection between certain equilibrium solutions of (P) has this property. As far as the authors know, no example of peaking solution was known previously for the exponential nonlinearity (1.1) in a bounded domain.

2. Preliminaries. Hereafter, for notational simplicity, we sometimes write the solution of (P) as $u(r, t)$ rather than $u(x, t)$, where r stands for $|x|$. This is the case in some part of subsections 2.3–2.5 and much of section 3.

2.1. Definition of L^1-solutions.

DEFINITION 2.1. *By an* L^1-solution *on an interval* $[0, T_0]$ *we mean a function* $u \in C([0, T_0];\, L^1(B_1))$ *such that* $f(u) \in L^1(Q_{T_0})$, $Q_{T_0} := B_1 \times (0, T_0)$ *and such that* *the equality*

$$\int_{B_1} (u(x,t)\Psi(x,t) - u_0(x)\Psi(x,0))\, dx - \int_0^t \int_{B_1} u\Psi_t \, dx\, ds$$
$$= \int_0^t \int_{B_1} (u\Delta\Psi + f(u)\Psi)\, dx\, ds$$

holds for any $0 < t \le T_0$ *and* $\Psi \in C^2(\overline{Q}_{T_0})$ *with* $\Psi = 0$ *on* $\partial\Omega \times [0, T_0]$. *By a* global L^1-solution *we mean a function* $u \in C([0,\infty);\, L^1(B_1))$ *which is an* L^1*-solution on* $[0, T_0]$ *for every* $T_0 > 0$.

Definition 2.2. *By a* limit L^1-solution *we mean a global* L^1*-solution which can be approximated by global classical solutions in the following way: There is a sequence* $\{u_{0,n}\}$ *in* $C(\overline{B}_1)$ *such that*

$$u_{0,n} \to u_0 \quad \text{in } C(\overline{B}_1) \tag{2.1}$$

and that the solution u_n *of* (P) *with* $u(\cdot,0) = u_{0,n}$ *exists globally for* $t \ge 0$ *and satisfies*

$$\begin{aligned} u_n(\cdot,t) &\to u(\cdot,t) && \text{in } L^1(B_1) \quad \text{for every } t > 0, \\ f(u_n) &\to f(u) && \text{in } L^1((B_1) \times (0,t)) \quad \text{for every } t > 0. \end{aligned} \tag{2.2}$$

We refer to any such sequence $\{u_n\}$ *as an* approximating sequence *for* u. *We call a limit* L^1*-solution a* minimal L^1-solution *if it has an approximating sequence that is pointwise nondecreasing in* n.

Remark 2.3. We do not need to assume in Definition 2.2 that $u(x,t)$ is a global L^1-solution. In fact, by condition (2.2), the limit function $u(x,t)$ automatically satisfies the integral identity in Definition 2.1. To see that u belongs to $C([0,\infty);\, L^1(B_1))$, recall that any approximating sequence of classical solutions u_n satisfies

$$u_n(\cdot,t) = e^{t\Delta}\, u_{0,n} + \int_0^t e^{(t-\tau)\Delta} f(u_n(\cdot,\tau))\, d\tau.$$

Letting $n \to \infty$ and using (2.2), we obtain

$$u(\cdot,t) = e^{t\Delta}\, u_0 + \int_0^t e^{(t-\tau)\Delta} f(u(\cdot,\tau))\, d\tau.$$

This proves the continuity $t \mapsto u(\cdot,t)$ in $L^1(B_1)$.

Remark 2.4. Condition (2.2) is automatically fulfilled if the approximating sequence u_n is monotone nondecreasing in n. In fact, the standard Kaplan estimate gives uniform bounds for $\|u_n(\cdot,t)\|_{L^1(B_1)}$ and $\|f(u_n)\|_{L^1(B_1\times[0,T_0])}$; hence (2.2) follows from the monotone convergence theorem.

Though we require only (2.1) and (2.2) for the approximating sequence u_n, the actual convergence takes place in a much stronger sense, as we shall see in Proposition 2.12.

The next lemma shows that any limit L^1-solution is a classical solution until it blows up.

Lemma 2.5. *Let* u *and* $\tilde{u}$ *be, respectively, a limit* L^1*-solution and a classical solution of* (P) *sharing the same initial data* u_0, *and denote by* $[0,T)$ *the maximal interval of existence for* $\tilde{u}$. *Then* $u = \tilde{u}$ *for* $0 \le t < T$. *Moreover, any approximating sequence* $\{u_n\}$ *for* u *satisfies*

$$u_n(\cdot,t) \to u(\cdot,t) \quad \text{in } C^2(\overline{B}_1) \quad \text{for each } t \in (0,T).$$

Proof. By the well-posedness of (P) in the space $C(\overline{B}_1)$, convergence (2.1) implies $u_n(\cdot,t) \to \tilde{u}(\cdot,t)$ in $C(\overline{B}_1)$ locally uniformly in $t \in (0,T)$. This, together with (2.2), yields $\tilde{u}(x,t) = u(x,t)$ for $0 \le t < T$. The C^2 convergence is a consequence of parabolic estimates. ◻

LEMMA 2.6. *Let u be a limit L^1-solution and suppose that*

$$\sup_{t_1<t<t_2} \|u(\cdot,t)\|_{L^\infty(B_1)} < \infty \tag{2.3}$$

for some $0 \le t_1 < t_2$. Then u is "regular" (i.e., a classical solution) in the interval $t_1 < t < t_2$.

Proof. By the same limiting argument as in Remark 2.3, we get

$$u(\cdot,t) = e^{(t-t_1)\Delta}\, u(\cdot,t_1) + \int_{t_1}^{t} e^{(t-\tau)\Delta} f(u(\cdot,\tau))d\tau$$

for $t \in [t_1,t_2)$. Since u is bounded we obtain $u \in C((t_1,t_2);\, C^1(\overline{B}_1))$; hence $f(u) \in C((t_1,t_2);\, C^1(\overline{B}_1))$. Using the parabolic estimates again, we see that u is a classical solution for $t_1 < t < t_2$. ◻

Another consequence of parabolic estimates is the following.

LEMMA 2.7. *Let u be a minimal L^1-solution satisfying* (2.3) *for some $0 \le t_1 < t_2$. Then any nondecreasing approximating sequence $\{u_n\}$ satisfies*

$$u_n(\cdot,t) \to u(\cdot,t) \quad \text{in } C^2(\overline{B}_1)$$

for each $t \in (t_1,t_2)$.

2.2. A pointwise Kaplan-type bound. It is well known that all global classical solutions of (P) satisfy certain integral estimates called Kaplan estimates (see [8]). In this subsection we introduce a multiscale version of Kaplan's technique and derive a useful pointwise bound for global weak solutions of (P). Our technique, which applies to radially decreasing solutions, is a slightly modified version of the one found in [11]. Note that, in the case of power nonlinearity (1.2), the same pointwise bound can also be derived by a totally different technique and it holds for any radially symmetric solutions that are not necessarily decreasing in r (see subsection 2.4 and also [10]).

To begin with, let us consider problem (P) where $u_0(x) \ge 0$ is a continuous function on $\overline{B}_1$ vanishing on ∂B_1 and $f : \mathbb{R} \to \mathbb{R}$ is a C^2 function satisfying

$$f'' > 0, \qquad \int_L^\infty \frac{du}{f(u)} < \infty \tag{2.4}$$

for some $L \ge 0$.

Next let μ be the second eigenvalue (the first positive eigenvalue) for the Laplacian in B_1 under the Neumann boundary conditions and under radial symmetry:

$$\begin{cases} \eta''(r) + \dfrac{N-1}{r}\eta'(r) = -\mu\eta(r), & 0 < r < 1, \\ \eta'(0) = \eta'(1) = 0. \end{cases} \tag{2.5}$$

More precisely, μ is a positive number such that there exists a function $\eta = \eta(r)$ satisfying (2.5) along with the condition

$$\eta'(r) < 0, \quad 0 < r < 1, \quad \eta(0) = 1,\ 0 > \eta(1).$$

Clearly these conditions determine both μ and η uniquely. For each $R \in (0,1]$, we set

$$\psi_R(x) = \eta(R^{-1}|x|) - \eta(1).$$

Then it is easily seen that ψ_R satisfies

$$\begin{cases} \Delta\psi_R = -\dfrac{\mu}{R^2}(\psi_R - C), & x \in B_R, \\ \psi_R > 0, & x \in B_R, \\ \psi_R = \dfrac{\partial \psi_R}{\partial n} = 0, & x \in \partial B_R. \end{cases}$$

Here the constant $C := -\eta(1)$ satisfies

$$\int_{B_R} \psi_R(x)\, dx = C|B_R|,$$

where

$$B_R := \{x \in \mathbb{R}^N \;:\; |x| < R\}, \quad |B_R| := \int_{B_R} dx.$$

Now we define a function $h_R(t)$ by

$$h_R(t) = \frac{1}{\int_{B_R} \psi_R\, dx} \int_{B_R} \psi_R(x) u(x,t)\, dx.$$

Then a simple computation shows that

$$\begin{aligned} \left(\int_{B_R} \psi_R\, dx\right) h_R'(t) &= \int \psi_R u_t\, dx \\ &= \int \psi_R \left(\Delta u + f(u)\right)\, dx \\ &= \int (\Delta \psi_R) u + \psi_R f(u)\, dx \\ &= \int_{B_R} \psi_R f(u)\, dx - \frac{\mu}{R^2} \int (\psi_R - C) u\, dx, \end{aligned}$$

provided that u is smooth. Thus, by Jensen's inequality, we get

$$h_R'(t) \geq f(h_R(t)) - \frac{\mu}{R^2} h_R(t) + \frac{\mu}{R^2} \frac{1}{|B_R|} \int_{B_R} u(x,t)\, dx. \tag{2.6}$$

If, in particular, $u \geq 0$, we have

$$h_R'(t) \geq f(h_R(t)) - \frac{\mu}{R^2}\, h_R(t).$$

The above inequality yields the following lemma.

LEMMA 2.8. *Let f satisfy* (2.4). *If u is a nonnegative classical solution of* (P) *defined for all $t \geq 0$, then*

$$h_R(t) \leq \rho_R, \quad t \geq 0,\ 0 < R \leq 1,$$

where ρ_R is the largest root of the equation

$$f(\rho) - \frac{\mu}{R^2}\, \rho = 0.$$

The well-known Kaplan's eigenfunction method (see [8]) consists in testing the equation in (P) by the first eigenfunction φ_1 of the Laplacian under the Dirichlet boundary condition where one chooses $\varphi_1 > 0$ such that $\int_{B_1} \varphi_1(x)\,dx = 1$. This yields a bound for $\int_{B_1} u(x,t)\varphi_1(x)\,dx$. In our case, testing by ψ_R with various $R > 0$ gives a localized version of such a bound.

Example 1. If $f(u) = \lambda e^u$, where $\lambda > 0$, then one easily sees

$$\rho_R \le \log \frac{\mu}{\lambda R^2} + \log\log \frac{\mu}{\lambda R^2} + \log \frac{e}{e-1}. \tag{2.7}$$

Example 2. If $f(u) = u^p$, where $p > 1$, then

$$\rho_R = \left(\frac{\mu}{R^2}\right)^{1/(p-1)}. \tag{2.8}$$

Remark 2.9. Mizoguchi [11] uses $\varphi_1(R^{-1}x)$ instead of ψ_R (for the power case (1.2)). This gives a slightly better constant μ, provided that solution u is radially decreasing. On the other hand, our choice of ψ_R has an advantage of yielding a more refined bound in subsection 2.5.

2.3. Basic estimates for the exponential. In this subsection we assume that

$$f(u) = \lambda e^u, \quad N > 2, \tag{2.9}$$

and that $u_0(x) = U_0(|x|)$ with

$$U_0 \in C^1([0,1]), \quad U_0' \le 0, \quad U_0(1) = 0. \tag{2.10}$$

Then $\partial u/\partial r \le 0$ for $r \ge 0$ and $t > 0$; hence Lemma 2.8 and (2.7) yield

$$0 \le u(r,t) \le \log \frac{\mu}{\lambda r^2} + \log\log \frac{\mu}{\lambda r^2} + \log \frac{e}{e-1}, \tag{2.11}$$

since $u(r,t) \le h_r(t)$. Consequently, for each $1 \le q < N/2$, there is a constant $M_q > 0$ such that

$$\|u\|_{L^q(B_1)} \le M_q, \quad \|\lambda e^u\|_{L^q(B_1)} \le M_q. \tag{2.12}$$

Next we recall that, for each $0 < \beta < 1$,

$$\|e^{t\Delta}\varphi\|_{X_\beta} \le C_\beta\, t^{-\beta}\|\varphi\|_{L^q}, \quad t > 0,$$

where $X_\beta = W_0^{2\beta,q}(B_1)$ and Δ denotes the Laplace operator in B_1 under the Dirichlet boundary conditions. Therefore, if v is a solution of the equation

$$v_t = \Delta v + g(x,t), \quad x \in B_1,\ t \ge 0 \tag{2.13}$$

under the Dirichlet boundary conditions, then from the expression

$$v(\cdot, t+\delta) = e^{\delta\Delta} v(\cdot,t) + \int_0^\delta e^{(\delta-\tau)\Delta} g(\cdot, t+\tau)\,d\tau$$

we get the estimate

$$\|v(\cdot,t+\delta)\|_{X_\beta} \le C_\beta\,\delta^{-\beta}\|v(\cdot,t)\|_{L^q} + \frac{C_\beta}{1-\beta}\,\delta^{1-\beta} \sup_{\tau\in(0,\delta)} \|g(\cdot,t+\tau)\|_{L^q}.$$

Combining this and (2.12), we obtain

$$\|u(\cdot, t+\delta)\|_{X_\beta} \le C_\beta M_q \left(\delta^{-\beta} + \frac{\delta^{1-\beta}}{1-\beta}\right), \quad t \ge 0,$$

for any $0 < \beta < 1$. Hence the following lemma holds.

LEMMA 2.10. *Assume* (2.9) *and* (2.10). *Let u be a classical solution of* (P) *defined for all $t \ge 0$. Then for each $\delta > 0$ there is a (universal) constant $A_{\delta,q,\beta} > 0$ independent of u such that*

$$\|u(\cdot, t)\|_{X_\beta} \le A_{\delta,q,\beta}, \quad \delta \le t < \infty.$$

Now choose q close enough to $N/2$ and β close enough to 1 so that the embedding

$$W_0^{2\beta,q}(B_1) \subset H_0^1(B_1) \tag{2.14}$$

is compact. This is possible since $N > 2$. Then Lemma 2.10 implies the following lemma.

LEMMA 2.11. *Let the assumptions of Lemma* 2.10 *be satisfied. Then for each $\delta > 0$ there exists a compact set $K_\delta \subset H_0^1(B_1)$ independent of u such that*

$$u(\cdot, t) \in K_\delta, \quad \delta \le t < \infty.$$

Consequently,

$$\|u(\cdot, t)\|_{H_1} \le C_\delta, \quad \delta \le t < \infty, \tag{2.15}$$

for some universal constant C_δ.

The next proposition shows that the convergence of approximating classical solutions to a limit L^1-solution takes place in a much stronger topology than (2.2).

PROPOSITION 2.12. *Assume* (2.9) *and* (2.10). *Let u be a limit L^1-solution and let $\{u_n\}$ be an approximating sequence for u. Then*

(i) *for each $t > 0$*

$$u_n(\cdot, t) \to u(\cdot, t) \quad \text{in } H_0^1(B_1), \qquad e^{u_n(\cdot,t)} \to e^{u(\cdot,t)} \quad \text{in } L^q(B_1)$$

for any $1 \le q < N/2$. Consequently, the estimates (2.12), (2.15) *hold for limit L^1-solutions;*

(ii) $u \in C((0,\infty); H_0^1(B_1))$;

(iii) *for each $t > 0$*

$$u_n(\cdot, t) \to u(\cdot, t) \quad \text{in } C^2_{loc}(\overline{B}_1 \setminus \{0\}).$$

Proof. By Lemma 2.5 and assumption (2.10) we have $\partial u/\partial r < 0$ for $0 < r \le 1$ and for $0 < t < T$, where T is the blow up time for u. Therefore, u_n satisfies the same inequality for n sufficiently large and for t sufficiently small (hence for all $t > 0$). Consequently the pointwise bound (2.11) holds for u_n with n sufficiently large. Combining this bound, the Lebesgue convergence theorem, and the former half of Lemma 2.11 (with u replaced by u_n), we obtain (i). Statement (ii) is a consequence of the continuity of $u(\cdot, t)$ in $L^1(B_1)$ and the fact that $u(\cdot, t)$ belongs to the compact set $K_\delta \subset H_0^1(B_1)$. Statement (iii) follows from bound (2.11) and local parabolic estimates for u_n. $\square$

COROLLARY 2.13. *Let u and u_n be as in Proposition* 2.12. *Then*

$$J[u_n] \to J[u] \quad \text{as } n \to \infty;$$

hence $J[u(\cdot,t)]$ is monotone nonincreasing in t, where

$$J[u] = \int_{B_1} \left(\frac{1}{2}|\nabla u|^2 - \lambda e^u \right) dx.$$

Proof. Since u_n is a classical solution, $J[u_n(\cdot,t)]$ is monotone nonincreasing in t. Letting $n \to \infty$ and using Proposition 2.12(i), we obtain the desired result. ◻

Next we show that (2.6) holds for any limit L^1-solution u. This will play a key role in subsection 2.5.

PROPOSITION 2.14. *Assume* (2.9) *and* (2.10). *Let u be a nonnegative limit L^1-solution of* (P). *Then*

$$h_R'(t) \geq \lambda e^{h_R(t)} - \frac{\mu}{R^2} h_R(t) + \frac{\mu}{R^2} \frac{1}{|B_R|} \int_{B_R} u(x,t)\, dx. \tag{2.16}$$

Proof. By (2.12) and the standard a priori estimate for the linear problem (2.13) with $g \in L^\infty((0,\infty); L^q)$, we have

$$u \in C^\gamma_{loc}((0,\infty); W^{2\beta,q}(B_1))$$

for any $0 < \gamma < 1 - \beta$ and any $1 < q < N/2$. Choosing q sufficiently close to $N/2$ and using the Sobolev embedding theorem, we see that

$$u \in C^\gamma_{loc}((0,\infty); L^r(B_1))$$

for any $1 < r < \infty$ and any $0 < \gamma < \min\{1, N/(2r)\}$. In view of this and the inequality

$$|e^u - e^v| \leq (e^u + e^v)|u - v|$$

along with bound (2.12), we get

$$e^u \in C^\gamma_{loc}((0,\infty); L^q(B_1))$$

for any $1 < q < N/2$ and for some appropriate exponent $0 < \gamma < 1$ depending on q. Again using the standard a priori estimates for (2.13) with Hölder continuous $g(\cdot,t)$, we obtain

$$u \in C^1((0,\infty); L^q(B_1)) \cap C((0,\infty); W^{2,q}(B_1))$$

for any $1 < q < N/2$. Therefore all the computations we used to derive (2.6) for classical solutions can be justified for limit L^1-solutions. This completes the proof. ◻

2.4. Basic estimates for the power. In this subsection we assume

$$f(u) = u^p, \quad N > 2, \quad p > \frac{N+2}{N-2}, \tag{2.17}$$

and that $U_0 \in C[0,1]$. Then Corollary 3.3 and Remark 3.5 in [10] show that any limit L^1-solution satisfies

$$|u(r,t)| \leq Cr^{-2/(p-1)} \quad \text{for } \ 0 \leq t < \infty \tag{2.18}$$

for some constant $C > 0$. In the special case where U_0 satisfies (2.10), a similar but more explicit pointwise bound

$$0 \le u(r,t) \le \left(\frac{\mu}{\lambda r^2}\right)^{1/(p-1)} \tag{2.19}$$

follows also from Lemma 2.8 and (2.8). These bounds imply that for each $1 \le q < \frac{N(p-1)}{2p}$, there is a constant $M_q > 0$ such that

$$\|u^p\|_{L^q(B_1)} \le M_q,$$

which is an analogue of (2.12) in the exponential case. Arguing as before, we see that Lemma 2.10 holds for the power case (1.2). Choosing q close enough to $N(p-1)/2p$ and β close enough to 1 we obtain the compact embedding (2.14). Arguing as before, we can derive the same results as Proposition 2.12 and Corollary 2.13. More precisely, we have the following proposition.

PROPOSITION 2.15. *Assume* (2.17) *and that $U_0(r)$ is a continuous function on* $[0,1]$. *Let u be a limit L^1-solution and let $\{u_n\}$ be an approximating sequence for u. Then for each $t > 0$,*

$$u_n(\cdot,t) \to u(\cdot,t) \quad in\ H_0^1(B_1) \cap L^q(B_1) \cap C^2_{loc}(\overline{B}_1 \setminus \{0\})$$

for any $1 \le q < \frac{N(p-1)}{2}$. Furthermore,

$$u \in C((0,\infty);\, H_0^1(B_1))$$

and $J[u(\cdot,t)]$ is monotone nonincreasing in t, where

$$J[u] = \int_{B_1} \left(\frac{1}{2}|\nabla u|^2 - \frac{1}{p+1}|u|^{p+1}\right) dx.$$

Note that the monotonicity assumption in (2.10) is not needed in the above proposition since (2.18) holds for any bounded initial data U_0.

2.5. Refined bound for the exponential. The following lemma gives an upper bound sharper than (2.11).

LEMMA 2.16. *Assume* (2.9) *and* (2.10). *Let u be a limit L^1-solution of* (P). *Suppose for some $0 < R \le 1$, $C_1 > 0$, $t_0 \in \mathbb{R}$, and $a > 0$,*

$$u(R,t) \ge \log\frac{1}{R^2} - C_1 \quad for\ \ t_0 \le t \le t_0 + aR^2. \tag{2.20}$$

Then

$$u(R,t_0) \le \log\frac{1}{R^2} + \alpha, \tag{2.21}$$

where α is a constant given by

$$\int_\alpha^\infty \frac{dk}{\lambda e^k - \mu(k + C_1)} = a. \tag{2.22}$$

Proof. By the assumptions we have

$$\frac{1}{|B_R|}\int_{B_R} u\,dx \ge \log\frac{1}{R^2} - C_1.$$

Combining this with (2.16), we obtain

$$h_R' \geq \lambda e^{h_R} - \frac{\mu}{R^2}\left(h_R - \log\frac{1}{R^2} + C_1\right).$$

Setting $k(t) := h_R(t) - \log(1/R^2)$, we get

$$k' \geq \frac{1}{R^2}\,(\lambda e^k - \mu k - \mu C_1).$$

It follows that

$$\int_{k(t_0)}^{k(t)} \frac{dk}{\lambda e^k - \mu(k + C_1)} \geq \frac{t - t_0}{R^2}, \quad t_0 \leq t \leq t_0 + aR^2.$$

Consequently, we have either

$$k(t_0) \leq k^*,$$

where k^* is the largest zero of the function

$$k \;\mapsto\; \lambda e^k - \mu(k + C_1),$$

or $k(t_0) > k^*$ and

$$\int_{k(t_0)}^{\infty} \frac{dk}{\lambda e^k - \mu(k + C_1)} \geq a.$$

Here we understand $k^* = -\infty$ if the function $\lambda e^k - \mu(k+C_1)$ has no zero. Now define $\alpha \in \mathbb{R}$ by (2.22). From the previous inequality it is clear that

$$k(t_0) \leq \alpha.$$

This proves (2.21). □

2.6. Characterization of the singular set. Now recall the definition of the sets $\mathcal{R}, \mathcal{S}, \mathcal{B}$ introduced in Introduction. Namely, $\mathcal{R}$ denotes the set of $t_0 > 0$ such that solution u stays classical around $t = t_0$, $\mathcal{S} := (0, \infty) \setminus \mathcal{R}$ denotes the set of "singular time moments," and

$$\mathcal{B} := \left\{ t_0 > 0 \,:\, \limsup_{t \to t_0} \|u(\cdot, t)\|_{L^\infty(B_1)} = \infty \right\}.$$

By Lemma 2.6, we have $\mathcal{S} = \mathcal{B}$. If $t_0 > 0$ is such that

$$\limsup_{t \nearrow t_0} \|u(\cdot, t)\|_{L^\infty(B_1)} < \infty,$$

then Lemma 2.6 and the local existence theorem for (P) imply that u can be continued as a classical solution beyond $t = t_0$; hence $t_0 \notin \mathcal{B}$. Therefore, $\mathcal{B}$ coincides with the set

$$\left\{ t_0 > 0 \,:\, \limsup_{t \nearrow t_0} \|u(\cdot, t)\|_{L^\infty(B_1)} = \infty \right\}.$$

In the case where u is a minimal L^1-solution, then it is easily seen that

$$\|u(\cdot,t_1)\|_{L^\infty(B_1)} < \infty$$

implies u is regular in some interval $t_1 \le t < t_1+\delta$, where $\delta > 0$ depends only on $\|u(\cdot,t_1)\|_{L^\infty}$. Therefore, if $t_0 > 0$ is such that

$$\liminf_{t\nearrow t_0} \|u(\cdot,t)\|_{L^\infty(B_1)} < \infty,$$

then $t_0 \notin \mathcal{B}$. Consequently the following lemma holds.

LEMMA 2.17. *If u is a minimal L^1-solution of* (P)*, then*

$$\mathcal{B} = \left\{ t_0 > 0 \;:\; \lim_{t\nearrow t_0} \|u(\cdot,t)\|_{L^\infty(B_1)} = \infty \right\}. \tag{2.23}$$

Finally, note that the pointwise bounds (2.11) and (2.18) show that singularity can occur only near the origin, both in the exponential case (1.1) and the power case (1.2). These pointwise bounds and local parabolic estimates for the approximating sequence $\{u_n\}$ imply that $u(x,t)$ is smooth in $(B_1 \setminus \{0\}) \times (0,\infty)$.

3. Main result.

3.1. Statement and remarks. Let I be an interval (open, half-open, or closed) with endpoints a, b, $-\infty \le a < b \le \infty$, and let f be a continuous function on I. We define the *zero-number* of f by

$$\begin{aligned}\mathcal{Z}_I(f) = \sup\{n \in \mathbb{N} \;:\; &\text{there are} \quad a < x_0 < x_1 < \cdots < x_n < b \\ &\text{such that} \quad f(x_i)f(x_{i+1}) < 0 \quad \text{for} \quad 0 \le i < n\}\end{aligned}$$

if f changes sign in I and $\mathcal{Z}_I(f) = 0$ otherwise.

THEOREM 3.1. *Let u be a minimal L^1-solution of problem* (P) *which blows up in a finite time T, and let either* (1.1) *or* (1.2) *be satisfied. Assume that the initial data $U_0(|x|)$ satisfy*

$$U_0 \in C([0,1]), \quad U_0(r) \ge 0 \ (0 \le r \le 1).$$

In case (1.1)*, assume further that $U_0(r)$ is nonincreasing in $0 \le r \le 1$. Then, there exists a positive integer k such that*

$$\mathcal{B} = \{t_i\}_{i=1}^k,$$

$$t_1 = T < t_2 < \cdots < t_k < \infty.$$

Consequently, the solution is regular except at $t = t_i$ $(i = 1, 2, \ldots, k)$. Moreover, the following estimate holds:

$$2k - 1 \le j := \min_{0<t<T} \mathcal{Z}_{[0,1]}\big(u_t(\cdot,t)\big). \tag{3.1}$$

In particular, if $U_0 \in C^2([0,1])$, then

$$2k - 1 \le j_0 := \mathcal{Z}_{[0,1]}\left(U_0'' + \frac{N-1}{r}U_0' + f(U_0)\right). \tag{3.2}$$

Remark 3.2. Since $w := u_t$ satisfies $w(1,t) = 0$ and a parabolic equation of the form

$$w_t = w_{rr} + \frac{N-1}{r} w_r + a(r,t)w,$$

the zero-number $\mathcal{Z}_{[0,1]}(u_t(\cdot,t))$ is nonincreasing in t and is finite for every $0 < t < T$. See [2] for details. Consequently, the minimum on the right-hand side of (3.1) is well-defined and is a finite integer. It is also clear that $j \le j_0$.

Remark 3.3. In order for the solution u to be a global L^1-solution, it should necessarily hold that $j > 0$. Indeed, if $j = 0$, this means that the solution satisfies $u_t(r,t) \ge 0$ $(0 \le r \le 1)$ for t close to T. By the result of [1], this means a complete blow-up; therefore, the solution cannot be continued as an L^1-solution beyond the blow up time T.

Remark 3.4. The above theorem means that the solution recovers smoothness immediately after the blow up time. Note that this result does not follow from standard parabolic estimates. Indeed, at the time of blow-up, some of the solutions may have a singularity of the form $\log(1/|x|^2) + C$ (in case (1.1)) or of the form $C|x|^{-2/(p-1)}$ (in case (1.2)), as exemplified by certain self-similar solutions. When such singularities occur, the solution profile $u(x,T)$ no longer belongs to the space where (P) is well-posed (for example, $L^q(B_1)$ with $q > \frac{(N-1)p}{2}$ in case (1.2)); therefore, parabolic regularization alone cannot bring the solution back to the space where (P) is well-posed. Thus smoothness does not follow automatically. Indeed the singular stationary solution φ^* defined below is an example of a weak solution that never becomes regular. (Since φ^* is not a minimal L^1-solution, there is no contradiction with the above theorem.)

Remark 3.5. By Lemma 2.17, we have $\lim_{t \nearrow t_0} \|u(\cdot,t)\|_{L^\infty} = \infty$ for every $t_0 \in \mathcal{B}$. However, we do not know whether or not this always implies that $\|u(\cdot,t_0)\|_{L^\infty} = \infty$. Since our equation has a supercritical nonlinearity, some subtle behavior may occur near the origin at the time of blow-up.

Incidentally, for an equation similar to (P), Pierre [17] found explicit examples of peaking solutions that blow up arbitrarily many times. He observed that

$$u(x,t) := \frac{1}{|x|^2 + \psi(t)}$$

is an L^1-solution of the equation

$$u_t = \Delta u + g(|x|,t)u^2, \quad g(r,t) := 2N - \psi'(t) - \frac{8r^2}{r^2 + \psi(t)} ,$$

provided ψ is a nonnegative C^1-function and $N > 4$. Obviously, u blows up at each time t such that $\psi(t) = 0$.

4. Further preliminaries.

4.1. Singular stationary solutions. In the proof of Theorem 3.1, the so-called singular stationary solution plays an important role. The equation

$$u_{rr} + \frac{N-1}{r} u_r + f(u) = 0, \quad r > 0, \tag{4.1}$$

has an explicit singular solution φ^* if either $f(u) = \lambda e^u$, $N \geq 3$ or $f(u) = u^p$, $N \geq 3$, $p > N/(N-2)$. Namely,

$$\varphi^*(r) = \log \frac{2(N-2)}{\lambda r^2} \tag{4.2}$$

in the former case, and

$$\varphi^*(r) = K r^{-\frac{2}{p-1}} \tag{4.3}$$

with

$$K = \left(\frac{2}{(p-1)^2} ((N-2)p - N) \right)^{\frac{1}{p-1}}$$

in the latter case. The assumption that $N \leq 9$ in (1.1) or $p < p^*$ in (1.2) guarantees the existence of a forward self-similar solution of the problem

$$\begin{cases} u_t = u_{rr} + \dfrac{N-1}{r} u_r + f(u), & r, t > 0, \\ u_r(0,t) = 0, & t > 0, \\ u(r,0) = \varphi^*(r), & r > 0, \end{cases} \tag{S}$$

which is regular for $r \geq 0$, $t > 0$ (cf. [4, 19, 18]). This forward self-similar solution is needed in our proof of Theorem 3.1.

Another notable feature of the critical dimension $N = 9$ for $f(u) = \lambda e^u$ and the critical power $p = p^*$ for $f(u) = u^p$ is that under assumption (1.1) or (1.2), it is well known (cf. [7]) that the graph of any smooth solution of (4.1) intersects with the graph of the singular solution φ^* infinitely many times, while this is not the case if $N > 9$ (in the exponential case) or $p > p^*$ (in the power case). This property will also be used in our proof of Theorem 3.1.

4.2. Zero-number properties for singular solutions. It is well known that if u and v are classical solutions of (P), then $\mathcal{Z}_{[0,1]}(u(\cdot,t) - v(\cdot,t))$ is a nonincreasing function of t. This is because $w := u - v$ satisfies a parabolic equation of the form

$$w_t = w_{rr} + \frac{N-1}{r} w_r + a(r,t) w.$$

Moreover, each time the function $r \mapsto u(r,t) - v(r,t)$ develops a degenerate zero somewhere in $[0,1]$, the above zero-number drops at least by 1. See [2] for details. It is easily seen that the same is true if u is a classical solution and $v = \varphi^*$, since $u(r,t) - \varphi^*(r)$ always have the same sign (i.e., negative) near $r = 0$. However, if u is an L^1-solution, then both u and φ^* may have a singularity at $r = 0$, and this makes the situation a bit more complicated. Nonetheless, a slightly weaker version of the above property still holds.

LEMMA 4.1. *Let $u(r,t)$ be a limit L^1-solution of* (P) *and let $\varphi^*(r)$ be the singular stationary solution. Let $t^* > 0$ and suppose that there exists a sequence $0 < \tau_1 < \cdots < \tau_k < t^*$ such that $u(r,\tau_i) - \varphi^*(r)$ has a degenerate zero in $(0,1]$ for $i = 1, 2, \ldots, k$. Then*

$$\mathcal{Z}_{[0,1]}(u(\cdot,t^*) - \varphi^*) \leq \mathcal{Z}_{[0,1]}(u(\cdot,0) - \varphi^*) - k. \tag{4.4}$$

Here we understand that $k = 0$ if there is no such τ_i in the interval $(0, t^*)$.

Proof. Let u_n be an approximating sequence for u, and let $t_0 \in (0,T)$ be such that $t_0 < \tau_1$ and that $u(r,t_0) - \varphi^*(r)$ has no degenerate zero in the interval $[0,1]$. Such t_0 exists since u is a classical solution for $0 \le t < T$; therefore, the function $r \mapsto u(r,t) - \varphi^*(r)$ can have a degenerate zero at most for a discrete set of values of t (see [2]). Then, since we have

$$u_n(\cdot,t_0) \to u(\cdot,t_0) \quad \text{in } C^2(\overline{B}_1)$$

by Lemma 2.5, the simplicity of the zeros of $u(r,t_0) - \varphi^*(r)$ implies

$$\mathcal{Z}_{[0,1]}(u_n(\cdot,t_0) - \varphi^*) = \mathcal{Z}_{[0,1]}(u(\cdot,t_0) - \varphi^*) \tag{4.5}$$

for n sufficiently large. On the other hand, Propositions 2.12 and 2.15 yield the convergence

$$u_n(\cdot,t^*) \to u(\cdot,t^*) \quad \text{in } C^2_{loc}(\overline{B}_1 \setminus \{0\}).$$

Consequently, we obtain

$$\mathcal{Z}_{[0,1]}(u_n(\cdot,t^*) - \varphi^*) \ge \mathcal{Z}_{[0,1]}(u(\cdot,t^*) - \varphi^*) \tag{4.6}$$

for n sufficiently large if the right-hand side is finite and

$$\lim_{n\to\infty} \mathcal{Z}_{[0,1]}(u_n(\cdot,t^*) - \varphi^*) = \infty \tag{4.7}$$

if the right-hand side of (4.6) is infinite. However, (4.7) is ruled out by (4.5) and the nonincrease of $\mathcal{Z}_{[0,1]}(u_n(\cdot,t^*) - \varphi^*)$, so we have (4.6).

Now let $r_i \in (0,1]$ $(i = 1,2,\ldots,k)$ be such that the function $r \mapsto u(r,\tau_i) - \varphi^*(r)$ has a degenerate zero at $r = r_i$. Suppose first that $r_i < 1$ $(i = 1,2,\ldots,k)$. (This is always true if $f(u) = u^p$, $p > 1$ or if $f(u) = \lambda e^u$, $\lambda > 0$, $\lambda \ne 2(N-2)$.) Choose a_i, b_i such that $0 < a_i < r_i < b_i < 1$ and that

$$u(a_i,\tau_i) - \varphi^*(a_i) \ne 0, \quad u(b_i,\tau_i) - \varphi^*(b_i) \ne 0 \quad \text{for} \quad i = 1,\ldots,k.$$

Next choose $\epsilon > 0$ sufficiently small so that $u(r,t) - \varphi^*(r) \ne 0$ for $r = a_i$, $\tau_i - \epsilon \le t \le \tau_i + \epsilon$ and $r = b_i$, $\tau_i - \epsilon \le t \le \tau_i + \epsilon$. We denote these two line segments by γ_i, $\tilde{\gamma}_i$. Since $u(r,t) - \varphi^*(r)$ has a degenerate zero at $(r,t) = (r_i,\tau_i)$ and since this function does not vanish on γ_i, $\tilde{\gamma}_i$, we have

$$\mathcal{Z}_{[a_i,b_i]}(u(\cdot,\tau_i - \epsilon) - \varphi^*) > \mathcal{Z}_{[a_i,b_i]}(u(\cdot,\tau_i + \epsilon) - \varphi^*). \tag{4.8}$$

Here we can choose $\epsilon > 0$ in such a way that the functions $u(\cdot,t_i \pm \epsilon) - \varphi^*$ have only simple zeros in the interval $[a_i,b_i]$ and that the intervals $[\tau_i - \epsilon, \tau_i + \epsilon]$, $i = 1,\ldots,k$, are mutually disjoint. This is possible since degenerate zeros can occur only at a discrete set of time t. Then,

$$\mathcal{Z}_{[a_i,b_i]}(u_n(\cdot,\tau_i \pm \epsilon) - \varphi^*) = \mathcal{Z}_{[a_i,b_i]}(u(\cdot,\tau_i \pm \epsilon) - \varphi^*)$$

for n sufficiently large; hence,

$$\mathcal{Z}_{[a_i,b_i]}(u_n(\cdot,\tau_i - \epsilon) - \varphi^*) > \mathcal{Z}_{[a_i,b_i]}(u_n(\cdot,\tau_i + \epsilon) - \varphi^*).$$

Moreover, since $u(r,t) - \varphi^*(r)$ does not vanish on γ_i, $\tilde{\gamma}_i$, the same is true for $u_n(r,t) - \varphi^*(r)$ for n sufficiently large. This and the above inequality imply that the function

$r \mapsto u_n(r,t) - \varphi^*(r)$ has a degenerate zero in $[a_i, b_i]$ for some $t \in (\tau_i - \epsilon, \tau_i + \epsilon)$. Consequently, at least k degenerate zeros occur in the time interval $[t_0, t^*]$; hence

$$\mathcal{Z}_{[0,1]}(u_n(\cdot, t^*) - \varphi^*) \le \mathcal{Z}_{[0,1]}(u_n(\cdot, t_0) - \varphi^*) - k.$$

Combining this inequality with (4.5) and (4.6), we obtain

$$\mathcal{Z}_{[0,1]}(u(\cdot, t^*) - \varphi^*) \le \mathcal{Z}_{[0,1]}(u(\cdot, t_0) - \varphi^*) - k.$$

Since u is a classical solution for $0 \le t < T$, we have

$$\mathcal{Z}_{[0,1]}(u(\cdot, t_0) - \varphi^*) \le \mathcal{Z}_{[0,1]}(u(\cdot, 0) - \varphi^*).$$

This and the previous inequality prove the lemma if $r_i < 1$ $(i = 1, 2, \ldots, k)$.

Consider the possibility $r_i = 1$ for some $i \in \{1, 2, \ldots, k\}$, which can occur only if $f(u) = \lambda e^u$ and $\lambda = 2(N-2)$. In this case, $u(t, \cdot) - \varphi^*$ satisfies the Dirichlet condition $u(1,t) - \varphi^*(1) = 0$. Therefore, choosing a_i as above (so that $u(t, a_i) - \varphi^*(a_i) \ne 0$ for $t \approx \tau_i$) we still obtain (4.8). We can now proceed similarly as before. ∎

Remark 4.2. Since $\mathcal{Z}_{[0,1]}(u(\cdot, t_0) - \varphi^*) < \infty$, the left-hand side of (4.4) is finite even if $\mathcal{Z}_{[0,1]}(u(\cdot, 0) - \varphi^*) = \infty$.

Remark 4.3. Note that the left-hand side of (4.4) is not necessarily monotone nonincreasing in t. This is because some intersection points between the graph of $r \mapsto u(r,t)$ and that of $\varphi^*(r)$ may escape to infinity (at $r = 0$) and later emerge from infinity repeatedly.

4.3. Rescaled equations. As usual, rescaling arguments provide useful information about the behavior of solutions near the blow up point. In the case of the exponential nonlinearity (1.1), we use the rescaling

$$w(y,s) = w^\theta(y,s) = u(r,t) + \log(\theta - t), \quad y = \frac{r}{\sqrt{\theta - t}}, \quad s = -\log(\theta - t), \tag{4.9}$$

where θ is any positive number. Then (P) is converted into the following problem:

$$(\mathrm{R}_e) \qquad \begin{cases} w_s = \dfrac{1}{\rho}(\rho w_y)_y + \lambda e^w - 1, & 0 < y < e^{s/2},\ s > -\log\theta, \\ w_y(0,s) = 0,\ w(e^{s/2}, s) = -s, & s > -\log\theta, \\ w(y, -\log\theta) = u_0(\sqrt{\theta}\, y) + \log\theta, & 0 \le y \le \theta^{-1/2}, \end{cases}$$

where

$$\rho(y) = y^{N-1} e^{-y^2/4}.$$

In the case of the power nonlinearity 1.2, we use the rescaling

$$w^\theta(y,s) = (\theta - t)^{1/(p-1)} u(r,t) \tag{4.10}$$

with y and s as before. Then, (P) is converted into

$$(\mathrm{R}_p) \qquad \begin{cases} w_s = \dfrac{1}{\rho}(\rho w_y)_y + w^p - \dfrac{1}{p-1} w, & 0 < y < e^{s/2},\ s > -\log\theta, \\ w_y(0,s) = 0,\ w(e^{s/2}, s) = 0, & s > -\log\theta, \\ w(y, -\log\theta) = \theta^{1/(p-1)} u_0(\sqrt{\theta}\, y), & 0 \le y \le \theta^{-1/2}. \end{cases}$$

Note that, in contrast to the usual setup, we do not assume that $\theta = T$, the blow up time of solution u. In fact, in our later argument, we shall need to consider the case $\theta > T$; thus, w may possess singularity in finite time. However, for the time being we assume that w is a classical solution of (R_e) or (R_p) and shall later use a limiting argument to deal with singular solutions.

Energy functionals corresponding to (R_e) and (R_p) are, respectively, the following:

$$E_e[w](s) := \int_0^{e^{s/2}} \left(\frac{1}{2}w_y^2 - \lambda e^w + w\right)\rho\, dy,$$
$$E_p[w](s) := \int_0^{e^{s/2}} \left(\frac{1}{2}w_y^2 - \frac{1}{p+1}w^{p+1} + \frac{1}{2(p-1)}w^2\right)\rho\, dy.$$

As is well known, E_p is a Lyapunov functional for (R_p). More precisely,

$$\frac{d}{ds}E_p[w](s) \le -\int_0^{e^{s/2}} w_s^2\,\rho\, dy, \quad s > -\log\theta; \tag{4.11}$$

hence, $E_p[w](s)$ is monotone decreasing in s. The same is true of E_e, at least for large s, as shown in the following lemma.

Lemma 4.4. *Let w be a global classical solution of* (R_e). *Then there is $s_0 < 1$ such that*

$$\frac{d}{ds}E_e[w](s) \le -\int_0^{e^{s/2}} w_s^2\rho\, dy, \quad s \ge s_0. \tag{4.12}$$

Proof. By direct computations we obtain

$$\begin{aligned}\frac{d}{ds}E_e[w](s) = &-\int_0^{e^{s/2}} w_s^2\rho\, dy\\ &-\left(w_y + \frac{1}{4}e^{s/2}w_y^2 + \frac{\lambda}{2}e^{-s/2} + \frac{1}{2}se^{s/2}\right)\rho\,\Big|_{y=e^{s/2}}.\end{aligned}$$

Since

$$w_y + \frac{1}{4}e^{s/2}w_y^2 \ge -e^{-s/2},$$

it suffices to choose s_0 such that $s_0 e^{s_0} > 2 - \lambda$. ∎

In the case of the power nonlinearity (1.2), the global existence of w^θ for all large s implies that $E_p[w^\theta](s) \ge 0$ (see [6]). Integrating (4.11) with respect to s, we obtain

$$\int_{s_1}^{s_2}\int_0^{e^{s/2}} (w_s^\theta)^2\rho\, dy\, ds \le E_p[w^\theta](s_1) - E_p[w^\theta](s_2).$$

Letting $s_2 \to \infty$ and using the boundedness of $E_p[w^\theta]$, we obtain

$$\int_{s_1}^{\infty}\int_0^{e^{s/2}} (w_s^\theta)^2\rho\, dy\, ds < \infty \tag{4.13}$$

(cf. Proposition 7.1 of [10]). As a matter of fact, the same estimate holds for any limit L^1-solutions. More precisely, we have the following lemma.

LEMMA 4.5. *Let u be a limit L^1-solution of* (P) *for the power nonlinearity* (1.2) *and let w^θ be as in* (4.10) *for some $\theta > 0$. Then* (4.13) *holds for any $s_1 > -\log\theta$.*

Proof. Let u_n be an approximating sequence for u. Then w_n^θ satisfies the same estimate as (4.13) for $n = 1, 2, 3, \dots$, where the bound does not depend on n. Letting $n \to \infty$ and using Fatou's lemma, we obtain (4.13). ∎

COROLLARY 4.6. *Let u and w^θ be as in Lemma* 4.5. *Then $w^\theta(y,s)$ approaches stationary solutions of* (R_p) *as $s \to \infty$ locally uniformly in $y > 0$. More precisely, the ω limit set of w^θ is contained in the set of solutions of*

$$\frac{1}{\rho}(\rho\, w_y)_y + w^p - \frac{1}{p-1} w = 0, \quad y > 0. \tag{4.14}$$

Proof. Estimate (2.18) yields

$$w^\theta(y,s) \le C y^{-2/(p-1)}. \tag{4.15}$$

This pointwise bound and parabolic regularization imply that, for any $M > 0$, the derivatives of w^θ are uniformly Hölder continuous in y in the region

$$y \in I_M := [M^{-1}, M], \quad s \ge s_1$$

for s_1 sufficiently large. Consequently $w^\theta(\cdot, s)$ remains in a compact set of $C^2(I_M)$ as s varies over $[s_1, \infty)$. Furthermore, the uniform Hölder continuity of w_s^θ and (4.13) imply that

$$w_s^\theta(y,s) \to 0 \quad \text{as} \quad s \to \infty$$

uniformly in $y \in I_M$. The conclusion of the lemma now follows immediately. ∎

In the case of the exponential nonlinearity (1.1), pointwise estimates of w^θ are much more difficult to obtain. However, under certain special circumstances we have an analogue of the above corollary.

LEMMA 4.7. *Let u be a limit L^1-solution of* (P) *for the exponential nonlinearity* (1.1) *and let w^θ be as in* (4.9). *Suppose that for some $0 < r_0 \le 1$, $C_1 > 0$, and $0 < t_1 < t_2$,*

$$u(r,t) \ge \log\frac{1}{r^2} - C_1 \quad \textit{for} \ \ 0 < r \le r_0,\ t_1 \le t \le t_2. \tag{4.16}$$

Then, for any $\theta \in (t_1, t_2)$, the solution $w^\theta(y,s)$ approaches stationary solutions of (R_e) *as $s \to \infty$ locally uniformly in $y > 0$. More precisely, the ω limit set of w^θ is contained in the set of solutions of*

$$\frac{1}{\rho}(\rho\, w_y)_y + \lambda e^w - 1 = 0, \quad y > 0. \tag{4.17}$$

Proof. By choosing a suitable constant $a > 0$, we see that

$$u(r,t) \ge \log\frac{1}{r^2} - C_1, \qquad t_0 \le t \le t_0 + ar^2$$

for any $0 < r \le r_0,\ t_1 \le t_0 \le \theta$. By Lemma 2.16, we have

$$u(r,t) \le \log\frac{1}{r^2} + \alpha, \qquad \text{for } 0 < r \le r_0,\ t_1 \le t \le \theta.$$

Therefore, w^θ satisfies

$$\log\frac{1}{y^2} - C_1 \le w^\theta(y,s) \le \log\frac{1}{y^2} + \alpha \tag{4.18}$$

for $0 < y \le r_0\, e^{s/2}$ and for all large s. Thus, once we have estimate (4.13), the same argument as in the proof of Corollary 4.6 will yield the conclusion of the lemma.

In order to prove (4.13), it suffices to show that $E_e[w^\theta](s)$ is bounded from below as $s \to \infty$. We have

$$\begin{aligned} E_e[w^\theta](s) &\ge \int_0^{e^{s/2}} (-\lambda e^{w^\theta} + w^\theta)\rho\, dy \\ &= \int_0^{r_0 e^{s/2}} (-\lambda e^{w^\theta} + w^\theta)\rho\, dy + \int_{r_0 e^{s/2}}^{e^{s/2}} (-\lambda e^{w^\theta} + w^\theta)\rho\, dy. \end{aligned}$$

The second integral is easily shown to converge to zero, as $s \to \infty$, thanks to the fact that $y \mapsto w(y,s)$ is decreasing (recall that in case (1.1) we assume that U_0 is nonincreasing). The first integral can be estimated using (4.18):

$$\begin{aligned} \int_0^{r_0 e^{s/2}} (-\lambda e^{w^\theta} + w^\theta)\rho\, dy &\ge \int_0^{e^{s/2}} \left(-\lambda e^{\log(1/y^2)+\alpha} + \log\frac{1}{y^2} - C_1\right)\rho\, dy \\ &= \int_0^{e^{s/2}} \left(-\lambda e^{\alpha}\frac{1}{y^2} + \log\frac{1}{y^2} - C_1\right) y^{N-1} e^{-y^2/4}\, dy. \end{aligned}$$

It is easily seen that the above integral remains bounded as $s \to \infty$. The lemma is proved. ∎

4.4. Singular stationary solutions for the rescaled equation. The singular stationary solution φ^* defined in (4.2) (resp., (4.3)) is also a singular stationary solution for the rescaled equation (4.17) (resp., (4.14)). That is, we have

$$\frac{1}{\rho}(\rho\,\varphi^*_y)_y + \lambda e^{\varphi^*} - 1 = 0, \quad 0 < y < \infty$$

$$\left(\text{ resp., } \frac{1}{\rho}(\rho\,\varphi^*_y)_y - \frac{1}{p-1}\,\varphi^* + (\varphi^*)^p = 0, \quad 0 < y < \infty \right).$$

In Corollary 4.6 and Lemma 4.7, we are not excluding the possibility that w^θ approaches a singular stationary solution as $s \to \infty$.

The following lemmas show that there is no singular stationary solution that lies above φ^*.

LEMMA 4.8. *Let φ^* be as in* (4.2) *and $\rho(y) = y^{N-1}e^{-\frac{y^2}{4}}$. If ψ is a solution of* (4.17) *satisfying $\psi \ge \varphi^*$, then $\psi = \varphi^*$.*

Proof. Suppose $\Phi := \psi - \varphi^* > 0$ for $0 < y < \infty$. Then $(\rho\,\Phi_y)_y < 0$ because Φ satisfies

$$(\rho\Phi_y)_y + \lambda\rho\left(e^{\psi} - e^{\varphi^*}\right) = 0.$$

Define a new space variable $z = z(y)$ by

$$z = \int_1^y \frac{d\xi}{\rho(\xi)}.$$

Then $(\rho\,\Phi_y)_y < 0$ $(y > 0)$ is equivalent to $\Phi_{zz} < 0$ $(-\infty < z < \infty)$. Therefore, Φ is a strictly concave function of z on $\mathbb{R}$, but this is impossible since $\Phi > 0$. ∎

LEMMA 4.9. *Let φ^* be as in* (4.3). *If ψ is a solution of* (4.14) *satisfying $\psi \geq \varphi^*$, then $\psi = \varphi^*$.*

Proof. Suppose $\Phi := \frac{\psi}{\varphi^*} > 1$ for $y \in (0,\infty)$. Then $(\sigma\,\Phi_y)_y < 0$ because Φ satisfies

$$(\sigma\Phi_y)_y + K^{p-1}\frac{\sigma}{y^2}\,(\Phi^p - \Phi) = 0,$$

where

$$\sigma(y) := y^{-\frac{4}{p-1}}\rho(y) = y^{N-1-\frac{4}{p-1}}\,e^{-\frac{y^2}{4}}.$$

The rest of the proof is the same as in the proof of Lemma 4.8 since our assumption $p > \frac{N+2}{N-2}$ implies $2 - N + \frac{4}{p-1} < 0$. ∎

5. Proof of the main theorem. In this section we prove Theorem 3.1. We begin with the following lemma.

LEMMA 5.1. *Under the assumptions of Theorem* 3.1*, the set $\mathcal{B}$ is decomposed into a disjoint union of finitely many closed intervals:*

$$\mathcal{B} = \bigcup_{i=1}^{k} A_i,$$

where $A_i = [t_i^1, t_i^2]$ or $A_i = \{t_i\}$ for $1 \leq i \leq k$, and A_k may also be of the form $[t_k, \infty)$.

Proof. By the definition of $\mathcal{B}$, it is clear that this set is closed. What we have to show is that the number of connected components of $\mathcal{B}$ does not exceed $(j+1)/2$.

Suppose $\mathcal{B}$ has at least k connected components. Then, considering that $u(x,t)$ is a classical solution for $t \notin \mathcal{B}$, and in view of (2.23), we can find a sequence of numbers

$$0 < \tau_1 < t_1 = T < \tau_2 < t_2 < \cdots < \tau_k < t_k$$

such that u is regular in the time interval $[\tau_i, t_i)$ and that

$$\lim_{t \nearrow t_i} u(0,t) = \infty.$$

Therefore, there exists $\tilde{\tau}_i \in (\tau_i, t_i)$ such that

$$u(0,\tau_1) < u(0,\tilde{\tau}_1) > u(0,\tau_2) < u(0,\tilde{\tau}_2) > \cdots < u(0,\tilde{\tau}_k).$$

Now let u_n be an approximating sequence for u. Then by Lemma 2.7 and the above inequality, we have

$$u_n(0,\tau_1) < u_n(0,\tilde{\tau}_1) > u_n(0,\tau_2) < u_n(0,\tilde{\tau}_2) > \cdots < u_n(0,\tilde{\tau}_k)$$

for n sufficiently large. It follows that $(u_n)_t(0,t)$ changes sign at least $2(k-1)$ times. Since $\mathcal{Z}_{[0,1]}((u_n)_t(\cdot,t))$ drops at least by 1 each time $(u_n)_t(0,t)$ changes sign, we have

$$\mathcal{Z}_{[0,1]}((u_n)_t(\cdot,\tilde{\tau}_k)) \leq \mathcal{Z}_{[0,1]}((u_n)_t(\cdot,\tau_1)) - 2(k-1).$$

Letting $n \to \infty$, we obtain

$$\mathcal{Z}_{[0,1]}(u_t(\cdot,\tilde{\tau}_k)) \leq \mathcal{Z}_{[0,1]}(u_t(\cdot,\tau_1)) - 2(k-1); \tag{5.1}$$

see Remark 5.2. Since τ_1 can be chosen arbitrarily close to $t_1 := T$, the right-hand side of (5.1) can be replaced by $j - 2(k-1)$. Moreover, since u does not blow up completely at $t = t_k$, we have

$$\mathcal{Z}_{[0,1]}(u_t(\cdot, \tilde{\tau}_k)) \geq 1$$

(cf. Remark 3.3). Combining this and (5.1), we obtain

$$1 \leq j - 2(k-1).$$

The lemma is proved. □

Remark 5.2. In obtaining (5.1) above, we have used the fact that the pointwise convergence $v_n(r) \to v(r)$ implies

$$\liminf_{n\to\infty} \mathcal{Z}_{(0,1]}(v_n) \geq \mathcal{Z}_{(0,1]}(v),$$

and that the equality holds if $v(r)$ has only simple zeros in the interval $0 \leq r \leq 1$. Thus (5.1) holds if $u_t(r, \tau_1)$ has only simple zeros in $0 \leq r \leq 1$. Since $u_t(r,t)$ can have a degenerate zero only for a discrete set of values of $t \in (0,T)$, we can always assume without loss of generality that $\tau_1 \in (0,T)$ has the above property.

Now we are ready to prove our main theorem.

Proof of Theorem 3.1. We show by contradiction that $A_1 = \{T\}$. Suppose $A_1 \supset [T, T+\delta]$ for some $\delta > 0$. By Lemma 4.1, $u(r,t) - \varphi^*(r)$ can have a degenerate zero in $0 < r \leq 1$ only at finitely many values of t. Therefore, we can find an interval $[\tau_1, \tau_2] \subset [T, T+\delta]$ (with $\tau_1 < \tau_2$) such that $u(r,t) - \varphi^*(r)$ has no degenerate zero for any $t \in [\tau_1, \tau_2]$. This means that the graph of $r \mapsto u(r,t)$ and that of $\varphi^*(r)$ always intersect transversally as t varies over the interval $[\tau_1, \tau_2]$; hence these intersection points are smooth functions of t. The number of the intersection points is uniformly bounded, as we see in Remark 4.2, but it may not be constant since some intersection points may escape into $r = 0$ (where $\varphi^* = \infty$) or emerge from $r = 0$ as t varies (see Remark 4.3). Nonetheless, by replacing $[\tau_1, \tau_2]$ with its suitable subinterval, if necessary, we may assume without loss of generality that the number of the intersection points is constant as t varies over $[\tau_1, \tau_2]$. Consequently, there exists $r_0 > 0$ such that either

$$u < \varphi^*, \quad r \in (0, r_0],\ t \in [\tau_1, \tau_2], \tag{5.2}$$

or

$$u > \varphi^*, \quad r \in (0, r_0],\ t \in [\tau_1, \tau_2]. \tag{5.3}$$

If (5.2) holds, the approximating sequence u_n satisfies

$$\begin{aligned} &u_n(r, \tau_1) < \varphi^*(r), \quad r \in (0, r_0], \\ &u_n(r_0, t) \leq u(r_0, t), \quad t \in [\tau_1, \tau_2], \end{aligned}$$

for $n = 1, 2, 3, \ldots$ since $u_1 < u_2 < u_3 < \cdots \to u$. Let $\tilde{u}$ be the solution of the initial value problem (S) introduced in subsection 4.1. Then since both u and $\tilde{u}$ are smooth outside $r = 0$, there exists some $\delta_0 > 0$ such that $u(r_0, t) < \tilde{u}(r_0, t - \tau_1)$ for $t \in [\tau_1, \tau_1 + \delta_0]$. By the comparison principle we have

$$u_n(r,t) < \tilde{u}(r, t - \tau_1), \quad r \in (0, r_0],\ t \in [\tau_1, \tau_1 + \delta_0],$$

for $n = 1, 2, 3, \ldots$. Letting $n \to \infty$, we obtain

$$u(r,t) \le \tilde{u}(r, t - \tau_1), \quad r \in (0, r_0],\ t \in [\tau_1, \tau_1 + \delta_0].$$

This implies that u is regular for $\tau_1 < t \le \tau_1 + \delta_0$, contradicting our assumption that $[T, T+\delta] \subset \mathcal{B}$.

Next we consider the case where (5.3) holds. We fix $\theta \in (\tau_1, \tau_2)$ arbitrarily and rescale u using the backward self-similar variables as in (4.9) (for the exponential case (1.1)) or as in (4.10) (for the power case (1.2)). Then the rescaled solution $w^\theta(y,s)$ satisfies (R_e) or (R_p), depending on the nonlinearity.

By Lemma 4.7 (in the case of the exponential nonlinearity) or by Corollary 4.6 (in the case of the power nonlinearity), $w^\theta(y,s)$ must approach stationary solutions as $s \to \infty$ locally uniformly in $y > 0$. Inequality (5.3) implies that w^θ can only approach stationary solutions that lie above φ^*. However, Lemmas 4.8 and 4.9 state that there is no stationary solution strictly above φ^*. Therefore,

$$w^\theta(y,s) \to \varphi^*(y) \quad \text{as} \quad s \to \infty \tag{5.4}$$

locally uniformly in $y > 0$. We shall show that this convergence cannot occur. We consider the exponential case and the power case separately.

First, let us consider the exponential case (1.1). Fix $t_0 \in (0, T)$ such that

$$\mathcal{Z}_{(0,1)}(u(\cdot, t_0) - \varphi^*) =: m_0 < \infty \tag{5.5}$$

and such that the function $r \mapsto u(r, t_0) - \varphi^*(r)$ has no degenerate zero in $0 < r \le 1$. For each $a > 0$, let $\varphi_a(r)$ be the solution of

$$\begin{cases} \varphi'' + \dfrac{N-1}{r}\varphi' + \lambda e^\varphi = 0, \quad r > 0, \\ \varphi_a'(0) = 0, \quad \varphi_a(0) = a. \end{cases} \tag{5.6}$$

In other words, $\varphi_a(|x|)$ is a stationary solution of (1.1) in $\mathbb{R}^N$. It is known that, under the assumption $3 \le N \le 9$,

$$\varphi_a(r) \to \varphi^*(r) \quad \text{as} \quad a \to \infty \tag{5.7}$$

locally uniformly in $r > 0$ and that

$$\mathcal{Z}_{(0,\infty)}(\varphi_a - \varphi^*) = \infty; \tag{5.8}$$

see [7]. The convergences (5.7) and (5.5) yield

$$\mathcal{Z}_{[0,1)}(u(\cdot, t_0) - \varphi_a) \in \{m_0, m_0 + 1\} \tag{5.9}$$

for all large a. On the other hand, by (5.8), we can choose $s = s_1$ large enough so that

$$\mathcal{Z}_{(0,\eta(s_1))}(\varphi_1 - \varphi^*) > m_0 + 1,$$

where $\eta(s) = e^{s/2}$. Then this and convergence (5.4) imply that

$$\mathcal{Z}_{(0,\eta(s_1))}(w^\theta(\cdot, s) - \varphi_1) > m_0 + 1$$

for all large s; hence

$$\mathcal{Z}_{(0,\eta(s))}(w^{\theta}(\cdot,s)-\varphi_1) > m_0+1.$$

Fix $s_2 \geq s_1$ large enough so that the above inequality holds for $s=s_2$, (5.9) holds for $a=1+s_2$, and $t_2 := \theta - e^{-s_2} > t_0$. Then the above inequality can be rewritten as

$$\mathcal{Z}_{(0,1/\sqrt{\mu})}(u^{\mu}(\cdot,t_2)-\varphi_1) > m_0+1,$$

where

$$u^{\mu}(r,t) = u(\sqrt{\mu}r,t)+\log\mu, \quad \mu=\theta-t_2.$$

Applying the rescaling $v(r)\mapsto v(r/\sqrt{\mu})-\log\mu$ to both u^{μ} and φ_1 and using the fact that

$$\varphi_1(r/\sqrt{\mu})-\log\mu = \varphi_a(r), \quad a=1-\log\mu = 1+s_2,$$

we obtain

$$\mathcal{Z}_{[0,1)}(u(\cdot,t_2)-\varphi_a) > m_0+1.$$

This, however, contradicts (5.9) since $\mathcal{Z}_{[0,1)}(u(\cdot,t)-\varphi_a)$ is monotone nonincreasing in t. This contradiction proves the assertion $A_1=\{T\}$ for the exponential case (1.1).

In the power case (1.2), the argument goes completely parallel to the above. The only difference is that φ^* has now form (4.3) instead of (4.2), φ_a is the solution of the problem

$$\begin{cases} \varphi''+\dfrac{N-1}{r}\varphi'+\varphi^p=0, \quad r>0,\\ \varphi_a'(0)=0, \quad \varphi_a(0)=a, \end{cases}$$

and the rescaling that we use is

$$u^{\mu}(r,t) = \mu^{\frac{1}{p-1}}u(\sqrt{\mu}r,t).$$

The rest of the proof is the same as in the exponential case. Thus we have $A_1=\{T\}$ both for the power case and the exponential case.

The same argument shows that the sets $A_2,\dots,A_k$ are all singletons. $\square$

6. Example of a peaking solution. In this section we consider (P) with the exponential nonlinearity (1.1) and prove the existence of a minimal L^1-solution that blows up exactly once.

THEOREM 6.1. *There exists an initial function u_0 such that the assumptions of Theorem 3.1 are satisfied in case (1.1) with $j=1$. This means that $\mathcal{B}=\{T\}$.*

We first recall some known properties of equilibria of (P) in case (1.1). The stationary problem corresponding to (P) is equivalent to

$$\text{(E)} \qquad \begin{cases} \phi_{rr}+\dfrac{N-1}{r}\phi_r+\lambda e^{\phi}=0, \qquad r\in(0,1),\\ \phi_r(0)=0, \qquad \phi(1)=0. \end{cases}$$

PROPOSITION 6.2 ([5, 7]; see Figure 1). *Denote by S the solution set of the parameterized problem (E):*

$$S=\{(\phi,\lambda) : \lambda\in\mathbb{R}^+ \textit{ and } \phi \textit{ is a solution of } \text{(E)}\}.$$

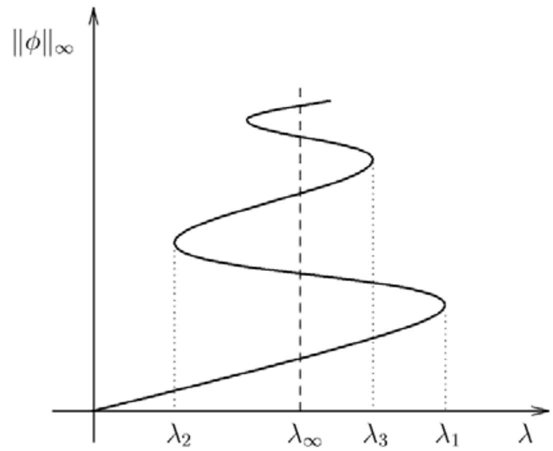


FIG. 1.

Then there exists a smooth curve

$$s \mapsto (\phi(s), \lambda(s)) : \mathbb{R}^+ \to C(\overline{B}_1(0)) \times \mathbb{R}^+$$

such that $S = \{(\phi(s), \lambda(s)) : s > 0\}$ *and that*

$$\sup_{x \in B_1(0)} \phi(s)(x) = \phi(s)(0) = s.$$

Moreover, the following holds:

(a) $\lim_{s\to 0} \lambda(s) = 0$, $\lim_{s\to\infty} \lambda(s) = \lambda_\infty := 2(N-2)$.

(b) *The set of all zeros of* $\lambda'(\cdot)$ *is given by a sequence* $0 < s_1 < s_2 < s_3 < \cdots \to \infty$ *and the critical values* $\lambda_j = \lambda(s_j)\,(j = 1, 2, 3, \dots)$ *satisfy*

$$\lambda_1 > \lambda_3 > \cdots > \lambda_{2j+1} \searrow \lambda_\infty, \quad \lambda_2 < \lambda_4 < \cdots < \lambda_{2j+2} \nearrow \lambda_\infty.$$

(c) *For each* $\lambda \le \lambda_1$ *define*

$$\phi_i^\lambda = \phi(\tilde{s}_i) \ \ (i = 0, 1, \dots),$$

where $\tilde{s}_0 < \tilde{s}_1 < \cdots$ *is the sequence of all points* s *with* $\lambda(s) = \lambda$. *This sequence is finite if* $\lambda \ne \lambda_\infty$ *and infinite if* $\lambda = \lambda_\infty$.

Next we recall the existence of a special blow up solution which can be continued globally as an L^1-solution.

PROPOSITION 6.3 (see [3]). *For any* $\lambda \in (\lambda_2, \lambda_3]$ *there is* u_0 *such that the solution* $u(\cdot, t)$ *of* (P) *has the following properties:*

(i) $u(\cdot, t)$ *blows up in finite time.*

(ii) $u(\cdot, t)$ *is a minimal* L^1*-solution.*

(iii) $u(\cdot, t)$ *is defined (as a classical solution of* (P)*) on the interval* $(-\infty, T)$ *for* $u(\cdot, t) \to \phi_2$ *in* C^1 *as* $t \to -\infty$.

(iv)

$$\frac{u_t(\cdot, t)}{\|u_t(\cdot, t)\|_{C^1(B_1)}} \to \psi_2 \quad \text{in } C^1(B_1) \quad \text{as} \quad t \to -\infty,$$

where ψ_2 is a normalized eigenfunction of

$$\begin{aligned} &\Delta\psi + \lambda e^{\phi_2(|x|)}\psi + \mu\psi = 0, \qquad x \in B_1, \\ &\psi = 0, \qquad x \in \partial B_1, \end{aligned}$$

corresponding to the second eigenvalue μ_2.

For (i)–(iii), see Theorem 3.4 in [3]. For (iv), see the proof of Lemma 3.6 (in particular, (3.14)) of the same paper.

Theorem 6.1 follows now from Proposition 6.3 because $\mathcal{Z}_{[0,1]}(\psi_2) = 1$ and ψ_2 has no degenerate zeros (thus $\mathcal{Z}_{[0,1]}(u_t(\cdot, t)) = 1$ for large negative t).

REFERENCES

[1] P. Baras and L. Cohen, *Complete blow-up after T_{max} for the solution of a semilinear heat equation*, J. Funct. Anal., 71 (1987), pp. 142–174.

[2] X.-Y. Chen and P. Poláčik, *Asymptotic periodicity of positive solutions of reaction-diffusion equations on a ball*, J. Reine Angew. Math., 472 (1996), pp. 17–51.

[3] M. Fila and P. Poláčik, *Global solutions of a semilinear parabolic equation*, Adv. Differential Equations, 4 (1999), pp. 163–196.

[4] V. Galaktionov and J.L. Vázquez, *Continuation of blow-up solutions of nonlinear heat equations in several space dimensions*, Comm. Pure Appl. Math., 50 (1997), pp. 1–67.

[5] I.M. Gelfand, *Some problems in the theory of quasilinear equations*, Amer. Math. Soc. Transl., 29 (1963), pp. 295–381.

[6] Y. Giga and R.V. Kohn, *Characterizing blowup using similarity variables*, Indiana Univ. Math. J., 36 (1987), pp. 1–40.

[7] D.D. Joseph and T.S. Lundgren, *Quasilinear Dirichlet problems driven by positive sources*, Arch. Ration. Mech. Anal., 49 (1973), pp. 241–269.

[8] S. Kaplan, *On the growth of solutions of quasi-linear parabolic equations*, Commun. Pure Appl. Math., 16 (1963), pp. 305–330.

[9] A.A. Lacey and D.E. Tzanetis, *Global, unbounded solutions to a parabolic equation*, J. Differential Equations, 101 (1993), pp. 80–102.

[10] H. Matano and F. Merle, *On non-existence of type* II *blow-up for a supercritical nonlinear heat equation*, Commun. Pure Appl. Math., 57 (2004), pp. 1494–1541.

[11] N. Mizoguchi, *On the behavior of solutions for a semilinear parabolic equation with supercritical nonlinearity*, Math. Z., 239 (2002), pp. 215–219.

[12] N. Mizoguchi, *Various behaviors of solutions for a semilinear heat equation after blowup*, J. Funct. Anal., 220 (2005), pp. 214–227.

[13] N. Mizoguchi, *Multiple blow-up of solutions for a semilinear heat equation*, Math. Ann., 331 (2005), pp. 461–473.

[14] N. Mizoguchi, *Multiple Blow-Up of Solutions for a Semilinear Heat Equation* II, preprint.

[15] N. Mizoguchi, *Boundedness of Global Solutions for a Semilinear Heat Equation with Supercritical Nonlinearity*, preprint.

[16] W.-M. Ni, P.E. Sacks, and J. Tavantzis, *On the asymptotic behavior of solutions of certain quasilinear parabolic equations*, J. Differential Equations, 54 (1984), pp. 97–120.

[17] M. Pierre, *private communication.*

[18] Ph. Souplet and F.B. Weissler, *Regular self-similar solutions of the nonlinear heat equation with initial data above the singular steady state*, Ann. Inst. H. Poincaré, Anal. Non Linéaire, 20 (2003), pp. 213–235.

[19] J.L. Vázquez, *Domain of existence and blowup for the exponential reaction diffusion equation*, Indiana Univ. Math. J., 48 (1999), pp. 677–709.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 777–796

LARGE-TIME BEHAVIOR FOR VISCOUS AND NONVISCOUS HAMILTON–JACOBI EQUATIONS FORCED BY ADDITIVE NOISE*

NICOLAS DIRR† AND PANAGIOTIS E. SOUGANIDIS‡

Abstract. We study the large-time behavior of the solutions to viscous and nonviscous Hamilton–Jacobi equations with additive noise and periodic spatial dependence. Under general structural conditions on the Hamiltonian, we show the existence of unique up to constants, global-in-time solutions, which attract any other solution.

Key words. Hamilton–Jacobi equations with additive noise, large-time behavior, Lipschitz regularity for Hamilton–Jacobi equations

AMS subject classifications. 35R60, 49L25, 35B40

DOI. 10.1137/040611896

1. Introduction.

We are interested in the long-time behavior of solutions to equations of the form

$$\mathrm{d}u - (\mathrm{tr}(A(x)D^2u) - H(Du,x))\mathrm{d}t + \mathrm{d}W(x,t) = 0 \quad \text{in } \mathbb{R}^n \times (t_0,\infty), \tag{1.1}$$

where $t_0 \in \mathbb{R}$ is arbitrary,

$$H \in C^{0,1}_{\mathrm{loc}}(\mathbb{R}^n \times \mathbb{R}^n) \quad \text{is } \mathbb{Z}^n\text{-periodic with respect to } x, \tag{1.2}$$

and, if S^n and $\mathcal{M}^{n\times m}$ are, respectively, the spaces of $n \times n$ symmetric and $n \times m$ matrices,

$$A \in C^{0,1}(\mathbb{R}^n; S^n) \text{ is } \mathbb{Z}^n\text{-periodic} \tag{1.3}$$

and

$$\text{there exists a } \mathbb{Z}^n\text{-periodic } \sigma \in C^{0,1}(\mathbb{R}^n;\mathcal{M}^{n\times m}) \text{ such that } A = \sigma\sigma^T. \tag{1.4}$$

Here we use the standard notation $C^{0,1}$ and $C^{0,1}_{\mathrm{loc}}$ for the spaces of Lipschitz continuous and locally Lipschitz continuous functions.

We note that (1.4) immediately implies that A is degenerate elliptic, i.e., for all $x, \xi \in \mathbb{R}^n \times \mathbb{R}^n$,

$$(A(x)\xi,\xi) \geq 0.$$

If A is uniformly elliptic, i.e., there exists $\nu > 0$ such that for all $x, \xi \in \mathbb{R}^n \times \mathbb{R}^n$,

$$(A(x)\xi,\xi) \geq \nu|\xi|^2,$$

*Received by the editors July 18, 2004; accepted for publication (in revised form) February 22, 2005; published electronically November 23, 2005.
http://www.siam.org/journals/sima/37-3/61189.html

†Department of Mathematics, University of Texas at Austin, 1 University Station C1200, Austin, TX 78712-0257, and MPI for Mathematics in the Sciences Leipzig, Inselstr. 22, 04103 Leipzig, Germany (ndirr@mis.mpg.de). The research of this author was supported in part by the postdoctorate program of the German Academic Exchange Service (DAAD) and the Department of Mathematics at the University of Texas at Austin.

‡Department of Mathematics, University of Texas at Austin, 1 University Station C1200, Austin, TX 78712-0257 (souganid@math.utexas.edu). The research of this author was supported in part by the National Science Foundation.

then (1.4) holds. As a matter of fact, the latter is true also if A is degenerate elliptic and $A \in C^{1,1}(\mathbb{R}^n; S^n)$.

Let $(\Omega, \mathcal{F}, P)$ be a standard probability space and

$$\Delta = \{(s,t) \in \mathbb{R}^2 : s \leqq t\}.$$

For each $(s,t) \in \Delta$, denote by $W(x,t,s,\omega)$ the increment of the random variable $W(x,\cdot,\omega)$ in the interval $[s,t]$. Then $W(x,t,s,\omega)$ has the form

$$W(x,t,s,\omega) = \sum_{i=1}^{M} F_i(x)(W_i(t,\omega) - W_i(s,\omega)), \tag{1.5}$$

where, for each $i = 1, \dots, M$,

$$W_i \text{ is a Brownian motion and } F_i \in C^2(\mathbb{R}^n) \text{ is } \mathbb{Z}^n\text{-periodic.} \tag{1.6}$$

In our analysis we do not need to assume that the Brownian motions $W_1, \dots, W_M$ are mutually independent. Indeed, throughout the paper, we use the fact that $W = (W_1, \dots, W_M)$ is continuous with respect to t almost surely in ω with increments in time which are independent and identically distributed over disjoint time intervals, and that, for all $\epsilon > 0$ and $\ell \in \mathbb{N}$,

$$\mathbb{P}\left(\sup_{t \in [0,\, l]} |W(t) - W(0)| < \epsilon \right) > 0. \tag{1.7}$$

In view of this, our analysis extends to any random forcing $\zeta(x,t,\omega)$ for which a notion of time integral $Z(x,t,s,\omega) = \int_s^t \zeta(x,\rho)\mathrm{d}\rho$ is defined in such a way that Z has the aforementioned properties. Moreover, using discontinuous viscosity solutions, it is possible to extend our analysis to equations driven by certain jump processes, such as, for example, kicking force (see [IK]). In order to keep the presentation short, we focus here on the Brownian case.

Our results hold for all initial data and initial times and for all realizations of the noise in Ω_C, the set of continuous paths of the Brownian motion, which has full measure ($\mathbb{P}(\Omega_C) = 1$), or a smaller set $\tilde{\Omega}$, also of full measure, to be defined later.

Throughout the paper we write $\mathbb{T} = [0,1]^n$, we denote by $C(\mathbb{T})$ the space of $\mathbb{Z}^n$-periodic continuous real-valued functions, and we use the seminorm $|||\cdot|||$ defined, for each $u \in C(\mathbb{T})$, by

$$|||w||| = \inf_{c \in \mathbb{R}} \|w - c\|,$$

where $\|\cdot\|$ is the usual *sup*-norm.

The deterministic version of (1.1), i.e., the equation

$$u_t - \operatorname{tr}\left(A(x)D^2u\right) + H(Du,x) = 0 \quad \text{in} \quad \mathbb{R}^n \times (t_0,\infty), \tag{1.8}$$

plays a fundamental role in our analysis.

Indeed our main result says that, under some additional assumptions on A, H and $F = (F_1, \dots, F_M)$, if (1.8) has a unique up to constants, periodic-in-space, and global-in-time attracting solution, then so does (1.1). In other words, there exists a unique up to constants, periodic with respect to x solution $u_{inv} : \mathbb{R}^n \times \mathbb{R} \times \Omega \to \mathbb{R}$ of (1.1) such that, if u is another solution of (1.1), then

$$\lim_{t\to\infty} |||u(\cdot,t) - u_{inv}(\cdot,t)||| = 0. \tag{1.9}$$

We briefly explain the strategy of the proof. The theory of a fully nonlinear stochastic PDE developed by Lions and one of the authors in [LS1], [LS2], and [LS3], which applies to more general equations, allows us to define pathwise solutions to (1.1). These can be expressed, using a simple transformation, as solutions of a PDE with random coefficients.

The comparison principle for viscosity solutions to (viscous) Hamilton–Jacobi equations implies that the distance between two solutions driven by the same noise cannot increase. Moreover, whenever the excursions of the Brownian motion remain small throughout a time interval, the solutions to (1.1) and (1.8) stay close. In view of (1.9), which holds for solutions of (1.8), the latter converge, as $t \to \infty$ to a unique up to constants attractive solution. It follows that the distance between solutions measured in the seminorm $|||\cdot|||$ decreases throughout such intervals. On the other hand, the independent increments property of W and (1.7) imply that, as $t \to \infty$, there exist enough intervals of small excursions for W. Hence the difference of any two solutions of (1.1) measured in $|||\cdot|||$ tends to 0 as $t \to \infty$. The claim then follows in a standard way.

An important step in showing that the solutions to the deterministic and stochastic equations stay close to each other in intervals of small excursions of the Brownian motion is the fact that, after times of order one, the solutions to (1.1) become Lipschitz continuous with respect to x, with a Lipschitz constant depending on the realization of the noise and not the initial datum. This fact, which is of independent interest, is the main technical result in the paper.

When the equation is of first order, i.e., $A \equiv 0$, the Lipschitz bound follows from the growth conditions on the Hamiltonian, which yield uniform L^∞-bounds on the solutions. For second-order equations, i.e., when $A \not\equiv 0$, there are two distinct cases. When H is superquadratic with respect to the gradient, it is again possible to obtain universal L^∞-bounds on the solutions. The Lipschitz estimate then follows as in the first-order case. When the Hamiltonian is superlinear but not superquadratic, the estimate is more delicate. In this case it is necessary to obtain the Lipschitz bound without using a priori L^∞-bounds for nonnegative solutions, which may not exist. Typically (see, e.g., Barles [B], Crandall, Lions, and Souganidis [CLS], and Lions [L]), the Lipschitz bounds depend on the spatial oscillations of the initial datum, a fact which is not enough for the argument here. We overcome this difficulty by obtaining uniform, after time of order one, estimates on the spacial oscillations of the solutions.

The problem under consideration in this paper is a "toy" example for far more complex models in, for example, phase transitions and growth processes (the so-called KPZ (Kadar–Parisi–Zhang) equation) and fluid mechanics (the stochastically forced Navier–Stokes equation).

The stochastic KPZ equation

$$\mathrm{d}u - \left(\epsilon \Delta u - |Du|^2\right) \mathrm{d}t - \mathrm{d}W = 0$$

is obtained by linearizing the forced mean curvature flow for small gradients and large force. Our results apply directly to this equation with additive forcing and more general operators.

Another concrete example to which our results apply is the stochastic Burgers equation with additive noise. Indeed, if $u \in C(\mathbb{R} \times (0,\infty))$ solves the stochastic Hamilton–Jacobi equation

$$\mathrm{d}u + (u_x)^2 \mathrm{d}t - \mathrm{d}W = 0,$$

then $v = u_x$ solves the Burgers equation

$$\mathrm{d}v + (v^2)_x \mathrm{d}t - \mathrm{d}W_x = 0. \tag{1.10}$$

The unique up to constants random attractor of the Hamilton–Jacobi equation yields a unique invariant measure for the Burgers equation.

Invariant measures for (1.10) and other closely related equations have been the object of extensive study. We refer to E et al. [EKMS], Iturriaga and Khanin [IK], Gomes et al. [GIKP] for the Burgers equation and Mattingly [M1], [M2] for the Navier–Stokes equation with stochastic forcing.

The large-time behavior of solutions of (1.8) depends strongly on whether $A \equiv 0$ or is uniformly elliptic, while very little is known in the degenerate case. When $A \equiv 0$, the problem was studied by Fathi [F], Roquejoffre [R], and Namah and Roquejoffre [NR1], [NR2], the most general results being the ones of Barles and Souganidis [BS2]. The behavior of (1.8) for uniformly elliptic A was studied by Barles and Souganidis [BS3].

When $A = 0$ and H is periodic in time, it was shown by Barles and Souganidis [BS1] (see also Fathi and Mather [FM]) that there are no global attracting solutions. As a matter of fact, phenomena like period doubling can occur. In the uniformly elliptic case, however, it was shown in [BS3] that there exists a unique up to constants attracting solution. Of course, the basic difference between the degenerate and uniformly elliptic settings is that, in the latter case, the equation admits a strong maximum principle.

It follows from our results that even when the equation does not have a strong maximum principle, the stochastic noise is sufficiently irregular for the solutions to lose dependence on the initial data, while this is not true in general for a deterministic time-dependent perturbation.

The proofs in our paper are based on general arguments from the theory of viscosity solutions. This allows us to consider general Hamiltonians H and matrices A. In view of the generality of our assumptions, this paper extends previous works of Iturriaga and Khanin [IK], E et al. [EKMS], and Gomes et al. [GIKP], which consider strictly convex Hamiltonians, and in [GIKP], a space independent uniformly elliptic second-order operator. If the Hamiltonian is strictly convex, the solution of (1.1) can be expressed as the value function of a control problem. The asymptotic behavior of the solutions then reduces to the study of the corresponding controlled stochastic and ordinary differential equations. Here, instead of convexity, we assume some form of asymptotic convexity of the level sets of H. Moreover, in the viscous case, the matrix A can be degenerate elliptic and may depend on space.

We remark that Gomes et al. [GIKP] show that attracting solutions for strictly convex Hamiltonians and $A = \epsilon I$ converge to attracting solutions of the first-order equation. A similar convergence result holds in our case for general A's.

The paper is organized as follows. In section 2 we introduce the notion of solution, we state all the assumptions and the main theorems of the paper, and we prove some preliminary facts. In section 3 we prove the existence of an attracting solution u_{inv} on $\mathbb{R}^n \times (-\infty, \infty)$, assuming that we have the Lipschitz regularization property discussed earlier. Section 4 is devoted to the proof of this property.

2. Assumptions, preliminaries, and results. We begin with the notion of a solution of (1.1). For this, we need the equation

$$v_t - \mathrm{tr}(A(x)D^2 v) + H(Dv + DW(x,t,t_0), x) = \mathrm{tr}(A(x)D^2 W(x,t,t_0)). \tag{2.1}$$

DEFINITION 2.1. *A function* $u : \mathbb{R}^n \times [a,b] \times \Omega \to \mathbb{R}$ *is a viscosity solution of* (1.1) *if, for all* $[t_0, t_1] \subseteq [a,b]$, *the function*

$$v(x,t,\omega) = u(x,t,\omega) - W(x,t,t_0,\omega)$$

is a viscosity solution of (2.1) *in* $\mathbb{R}^n \times [t_0, t_1]$.

This definition coincides with the more general notion of stochastic viscosity solutions in [LS1], [LS2], [LS3]. Notice that when $A \equiv 0$, for the definition we only need $F \in C^1$. When A is uniformly elliptic and sufficiently smooth—for example, when A has constant coefficients—then it is possible to give an alternative definition requiring less differentiability of the F. Indeed, consider the solution w of the linear stochastic PDE

$$\begin{cases} \mathrm{d}w(x,t,t_0) - \mathrm{tr}(A(x)D^2 w(x,t,t_0))\mathrm{d}t = \mathrm{d}W(x,t), \\ w(x,t_0,t_0) = 0. \end{cases}$$

The basic regularity theory for uniformly parabolic linear equations yields, for some $C > 0$, the estimate

$$\|w(\cdot,t,t_0)\|_{C^2(\mathbb{T})} \le C(\|W\|_{C^{0,\alpha}([t_0,t_1])} + \|F\|_{C^{2,\alpha}(\mathbb{T})}).$$

In this case we say that u is a viscosity solution of (1.1) if $v = u - w(\cdot,\cdot,t_0)$ solves

$$v_t - \mathrm{tr}(A(x)D^2 v) + H\big(Dv + Dw(x,t,t_0), x\big) = 0 \text{ in } \mathbb{R}^n \times [t_0, t_1].$$

Next we state a proposition which asserts the existence and uniqueness of pathwise solutions of (1.1). Since the result is an immediate consequence of the theory of viscosity solutions (see [CIL], [B]) and Definition 2.1, we omit the proof.

PROPOSITION 2.2. *Assume* (1.2), (1.3), (1.4), (1.5), *and* (1.6). *For all* $\omega \in \Omega_C$, $s \in \mathbb{R}$, *and* $u \in C(T)$, *there exists a unique stochastic viscosity solution* $u(\cdot,\cdot,s,\omega) \in C\big(\mathbb{R}^n \times [s,\infty)\big)$ *of* (1.1) *such that* $u(\cdot,s,s,\omega) = u$.

Throughout the paper we denote by $S^{W,A}(t,s)(u)$ the stochastic viscosity solution of (1.1) starting with initial datum u at s. The solution to (1.8) is denoted by $S^{0,A}(t,s)(u)$. When $A \equiv 0$ and the context allows it, we write $S^W(t,s)$ and $S^0(t,s)$ to denote the solution operators to (1.1) and (1.8), respectively. Finally, whenever it does not create any ambiguity, we write $S^{W,A}(t,s)$ for both $S^{W,A}(t,s)$ and $S^{0,A}(t,s)$.

Since it will be used later, we note here that, as an immediate consequence of Proposition 2.2, both $S^{0,A}(t,s)$ and $S^{W,A}(t,s)$ commute with constants, i.e., for all $c \in \mathbb{R}^n$,

$$S^{W,A}(t,s)(v+c) = S^{W,A}(t,s)(v) + c. \tag{2.2}$$

We proceed with the assumptions on the Hamiltonian H, which we will be using in this paper.

$$\begin{cases} \text{There exist } K > 0 \text{ and } q > 1 \text{ such that for all } (p,x) \in \mathbb{R}^n \times \mathbb{R}^n, \\ \qquad\qquad H(p,x) \ge K^{-1}|p|^q - K. \end{cases} \tag{2.3}$$

$$\begin{cases} \text{There exist } R_0 > 0 \text{ and a strictly increasing } \Phi \in C\big([0,\infty),[0,\infty)\big) \\ \text{with } \Phi(0) = 0, \text{ such that for all } (p,x) \in \mathbb{R}^n \times \mathbb{R}^n \text{ with } |p| \ge R_0, \\ \qquad\qquad D_pH(p,x)\cdot p - H(p,x) \ge \Phi(|p|). \end{cases} \tag{2.4}$$

$$\text{(2.5)}\qquad \begin{cases}\text{There exist } R_0 \text{ and } B > 0 \text{ such that for all } (p,x) \in \mathbb{R}^n \times \mathbb{R}^n\\ \text{with } |p| \geq R_0, -D_x H(p,x) \cdot p \leq B|p|^2(D_p H(p,x) \cdot p - H(p,x)).\end{cases}$$

(2.6)

$$\begin{cases}\text{There exist } R_0 > 0 \text{ and a strictly increasing } \Phi \in C([0,\infty);[0,\infty)) \text{ with } \Phi(0) = 0,\\ \text{ such that for some } \delta > 0, G(r) = \Phi(r)r^{-(1+\delta)} \text{ is increasing,}\\ G(r) \to \infty \ \text{ as } r \to \infty, \text{ and for all } (p,x) \in \mathbb{R}^n \times \mathbb{R}^n \text{ with } |p| \geq R_0,\\ \qquad\qquad D_p H(p,x) \cdot p - H(p,x) \geq \Phi(|p|).\end{cases}$$

$$\text{(2.7)}\qquad \begin{cases}\text{There exists } C > 0 \text{ such that for all } (p,x) \in \mathbb{R}^n \times \mathbb{R}^n \text{ with } |p| \geq R_0,\\ \qquad\qquad -D_x H(p,x) \cdot p \leq C(D_p H(p,x) \cdot p - H(p,x)).\end{cases}$$

$$\text{(2.8)}\qquad \limsup_{|p|\to\infty}(D_p H(p,x) \cdot p - H(p,x))^{-1}|D_p H(p,x)| = 0 \ \text{ uniformly in } x \in \mathbb{R}^n.$$

$$\text{(2.9)}\qquad \sup_{x\in\mathbb{R}^n} \limsup_{|p|\to\infty} \left(D_p H(p,x) \cdot p - H(p,x)\right)^{-1} |p||D_p H(p,x)| < \infty.$$

$$\text{(2.10)}\qquad \begin{cases}\text{There exist a unique } \lambda \in \mathbb{R} \text{ and a unique up to constants } U \in C(\mathbb{T}),\\ \text{both depending on } A \text{ and } H, \text{ such that for each } v \in C(\mathbb{T}) \text{ and } t_0 \in \mathbb{R},\\ \text{there exists } c \in \mathbb{R} \text{ such that}\\ \lim_{N\to\infty} \sup_{x\in\mathbb{T}} \left|S^{0,A}(t_0+N,t_0)(v) - (U+c) - \lambda N\right| = 0.\end{cases}$$

Assumptions (2.4) and (2.6) state that the level sets of H as a function of p become convex for large $|p|$. This asymptotic condition is crucial for obtaining Lipschitz bounds which do not depend on the initial data and is much weaker than requiring the Hamiltonian to be convex in p.

The sole purpose of (2.8) and (2.9) is to ensure that the Hamiltonian in (2.1), which arises after incorporating the noise, still satisfies the growth assumptions (2.3), (2.4), (2.5) in the nonviscous case and (2.3), (2.6), (2.7) in the viscous case, with constants which may depend on t_0, t, and ω.

Among all the above, the most important assumption is (2.10). It states that the corresponding deterministic equation has a global attractor, which consists—up to constants—of a single trajectory. We refer to the introduction for a discussion concerning this fact and to [BS1], [BS2], and [BS3] for results yielding (2.10) as well as an extensive list of references.

The main result of this paper is the next theorem. The strategy for the proof of the first part was outlined in the introduction. As we explain later in this section the second part is a simple consequence of the first and the stability properties of the viscosity solutions.

THEOREM 2.3. *Assume* (1.2), (1.5), (1.6), (2.3), *and* (2.10). *There exists* $\widetilde{\Omega} \subseteq \Omega$ *with* $\mathbb{P}(\widetilde{\Omega}) = 1$ *such that for every* $\omega \in \widetilde{\Omega}$, *the following hold:*

(i) *If* $A \equiv 0$ *and, in addition,* (2.4), (2.5), *and* (2.8) *hold, or if* $A \not\equiv 0$ *satisfies* (1.3), (1.4) *and, in addition,* (2.6), (2.7), (2.9) *hold and* $F_i \in C^3(\mathbb{T})$, *there exists a unique up to constants solution* $u_{inv}(\cdot,\cdot,\omega) \in C(\mathbb{R}; C^{0,1}(\mathbb{T}))$ *of* (1.1) *attracting any other solution, i.e., for any* $v \in C(\mathbb{T})$ *and* $s \in \mathbb{R}$,

$$\lim_{t\to\infty} \|\!|u_{inv}(\cdot,t,\omega) - S^W(t,s)(v)(\cdot)|\!\|= 0.$$

(ii) *Assume that* $A = \epsilon \tilde{A}$ *is uniformly elliptic and satisfies* (1.3). *If* $u^{\epsilon}_{inv}(\cdot, \cdot, \omega)$ *and* ${u_{inv}}^0(\cdot, \cdot, \omega)$ *are the unique up to constants invariant solutions of* (1.1) *corresponding to* $\epsilon > 0$ *and* $\epsilon = 0$, *respectively, then for any* $[a, b] \subset (-\infty, +\infty)$,

$$\lim_{\epsilon \to 0} \sup_{t \in [a,b]} |||u^{\epsilon}_{inv}(\cdot, t, \omega) - u^{0}_{inv}(\cdot, t, \omega)||| = 0.$$

As was already mentioned in the introduction, for $A \equiv 0$ and $H(p) = |p|^2$ this result was first proved by [EKMS] in one dimension and by [IK] in all dimensions for general strictly convex H and uniformly elliptic x-independent A. Our assumptions allow, however, to consider nonconvex Hamiltonians and degenerate elliptic A. For example, H can have the form

$$H(p, x) = |p|^2 \widehat{H}(\hat{p}, x),$$

where, for $p \in \mathbb{R}^n \setminus \{0\}$, $\hat{p} = |p|^{-1} p$, and $\widehat{H}$ is periodic in x and uniformly bounded away from 0. It is straightforward to check that all structural assumptions on H hold. Moreover, it is proved in [BS2] and [BS3] that for each $v \in C(\mathbb{T})$, $S^0(t)(v)$ has a limit as $t \to \infty$. The up to constants uniqueness of the asymptotic limit of the deterministic equation is here an assumption, which holds, for example, if $\widehat{H}$ is independent of x.

Most of the growth conditions on H are needed for the following lemma, which plays a central role in the paper. In fact, this lemma is of independent interest, as it extends known regularity results for viscous Hamilton–Jacobi equations.

For $(t_1, t_2) \in \Delta$, we write

$$C_W(t_1, t_2, \omega) = \max_i \sup_{t \in [t_1, t_2]} \left| \int_{t_1}^{t} \mathrm{d}W_i(s, \omega) \right|. \tag{2.11}$$

We have the following.

LEMMA 2.4. *Assume* (1.2), (1.3), (1.4), (1.5), (1.6), (2.3) *and either* (2.6), (2.7), (2.9), *and* $F_i \in C^3(\mathbb{T})$ *if* $A \not\equiv 0$ *is degenerate elliptic, or* (2.4), (2.5), *and* (2.8) *if* $A \equiv 0$. *For all* $\omega \in \Omega_C$ *and* $(s, t) \in \Delta$, *there exists* $L(s, t, \omega) > 0$ *such that for all* $v \in C(\mathbb{T})$,

$$\inf_{c \in \mathbb{R}} \|S^{W,A}(t, s)(v) - c\|_{C^{0,1}(\mathbb{T})} \leq L(s, t, \omega).$$

Moreover, there exists $\hat{L} : (0, \infty) \times (0, \infty) \to (0, \infty)$ *which is increasing with respect to the second argument, such that*

$$\text{if } C_W(s, t, \omega) \leq K, \quad \text{then } L(s, t, \omega) \leq \widehat{L}(t - s, K).$$

It follows from Lemma 2.4 that solutions to (1.1) are Lipschitz continuous in space with Lipschitz constant independent of the initial datum. For solutions of the deterministic time-independent equation (1.8), the lemma holds with an L which depends only on $|t - s|$.

The claim about the vanishing viscosity limit asserted in Theorem 2.3 is a simple consequence of our results and standard arguments from the theory of viscosity solutions. Indeed, Lemma 2.4 yields that the family $({u^{\epsilon}}_{inv})_{\epsilon > 0}$ is uniformly Lipschitz continuous on any given compact time interval. A simple diagonalization argument yields a subsequence which converges uniformly on compact intervals to a viscosity solution u of (1.1) with $A \equiv 0$. Lemma 3.7 below then asserts that we must have $u(x, t) = u^0_{inv}(x, t) + c(t)$. However, since both u and u^0_{inv} are solutions, the constant c cannot depend on time. Therefore the whole family $({u^{\epsilon}}_{inv})_{\epsilon > 0}$ converges up to constants to u^0_{inv}.

3. Proofs. We begin with a number of preliminary lemmas which summarize some of the key properties of the solutions of (1.1). The first lemma is an immediate consequence of the definition of a solution and the comparison principle for viscosity solutions (see [CIL]); hence we omit its proof.

LEMMA 3.1. *For all u, $v \in C(\mathbb{T})$ and $(s,t) \in \Delta$,*

$$\|S^{W,A}(t,s)(u) - S^{W,A}(t,s)(v)\|_{C(\mathbb{T})} \leq \|u - v\|_{C(\mathbb{T})}.$$

For $v_0 \in C^{0,1}(\mathbb{T})$ and $(t_1, t_2) \in \Delta$ we denote by

$$L_A(t_1,t_2) = \sup_{s\in[t_1,t_2]} \|DS^{0,A}(s,t_1)(v_0)\|$$

the uniform Lipschitz constant of the solution of the deterministic equation.

We also write C_A and C_0 for the constants

$$C_A = \max_{x\in\mathbb{T},\ |p|\leq L_A(t_1,t_2)} \big(|D_pH(p,x)|+1\big)\|F\|_{C^3(\mathbb{T})} \quad \text{if} \quad A \not\equiv 0$$

and

$$C_0 = \max_{x\in\mathbb{T},\ |p|\leq L_{A\equiv 0}(t_1,t_2)} (|D_pH(p,x)| \ +1)\|F\|_{C^2(\mathbb{T})} \quad \text{if} \quad A \equiv 0.$$

LEMMA 3.2. *Let $v_0 \in C^{0,1}(\mathbb{T})$ and $(t_1, t_2) \in \Delta$. Then*

$$\big\|S^{0,A}(t_2,t_1)(v_0) - S^{W,A}(t_2,t_1)(v_0)\big\| \leq (t_2 - t_1)\, C_A \,\|F\|\, C_W(t_1,t_2,\omega).$$

Proof. 1. To simplify the presentation we assume that $t_1 = 0,\ t_2 = T$ and we use the notation $C = C_A \,\|F\|_{L^\infty}\, C_W(t_1,t_2,\omega)$, $u = S^{W,A}(v_0)$, and $v = S^{0,A}(v_0)$.

2. Arguing by contradiction, we assume that there exists $(x_0, t_0) \in \mathbb{T}\times(0,T)$ such that, possibly after exchanging the role of u and v, $u(x_0,t_0) - v(x_0,t_0) - Ct_0 > 0$. Standard arguments from the theory of viscosity solutions (see [CIL]) then yield $\eta > 0$ and $(X_\alpha, p_\alpha, x_\alpha, t_\alpha)$, $(Y_\alpha, p_\alpha, y_\alpha, s_\alpha) \in S^n \times \mathbb{R}^n \times \mathbb{R}^n \times (0,T)$ such that, as $\alpha \to \infty$,

$$\begin{cases} |t_\alpha - s_\alpha| + \alpha|y_\alpha - x_\alpha|^2 \to 0, \quad \mathrm{tr}\,(A(y_\alpha)Y_\alpha) - \mathrm{tr}\,(A(x_\alpha)X_\alpha) \leq L\alpha|x_\alpha - y_\alpha|^2, \\ C + \eta(T-t_\alpha)^{-2} + H(p_\alpha,x_\alpha) - \mathrm{tr}(A(x_\alpha)X_\alpha) \\ \leq -\eta(T-s_\alpha)^{-2} + H(p_\alpha + DW(y_\alpha,s_\alpha),y_\alpha) - \mathrm{tr}\,(A(y_\alpha)Y_\alpha)\,. \end{cases}$$

The (degenerate) ellipticity of A, the choice of C, and the above inequalities contradict the fact that $\eta > 0$. ∎

Note that the above estimates depend on the Lipschitz constant of the deterministic equation. Hence to use this lemma, it is necessary to have a universal bound on those Lipschitz constants, like the one asserted by Lemma 2.4.

The next claim strengthens the assertion of (2.10), which asserts only pointwise convergence as $t \to \infty$ of the solution operator $S^{0,A}(t,s)$ acting on $C(\mathbb{T})$. It turns out that this convergence is uniform with respect to the initial data.

LEMMA 3.3. *Assume* (2.10) *and the hypotheses of Lemma* 2.4 *hold. There exists a unique up to constants function $U_A^* \in C^{0,1}(\mathbb{T})$ such that, for all $t \in \mathbb{R}$,*

$$\lim_{k\to\infty}\left(\sup_{v\in C^0(\mathbb{T})} |||S^{0,A}(t,-k)(v) - U_A^*|||\right) = 0.$$

Proof. 1. Since the deterministic equation does not depend on time, we may take $t = 0$. Assume that, for some $\delta > 0$, there exist $(v_k)_{k\in\mathbb{N}} \in C(\mathbb{T})$ such that

$$\||S^{0,A}(0,-k)(v_k) - U_A^*\|| \geq \delta \quad \text{for all } k \in \mathbb{N}, \tag{3.1}$$

where U_A^* is the unique (up to constants) limit which exists in view of (2.10).

2. The Lipschitz continuity asserted in Lemma 2.4 yields constants c_k such that the family $(\hat{v}_k)_{k\in\mathbb{N}}$ defined by

$$\hat{v}_k = S^{0,A}(-k+1,-k)(v_k) - c_k$$

is bounded in $C^{0,1}$ and thus compact in $C(\mathbb{T})$. Hence there exists a subsequence $k_m \to \infty$ such that $\widehat{v}_{k_m} \to \hat{v}$ in C^0.

3. Consider the family of maps $S_k : C(\mathbb{T}) \to C(\mathbb{T})$ given by

$$S_k(v) = S^{0,A}(0,-k+1)(v).$$

The contraction property yields that, as $m \to \infty$,

$$\|S_{k_m}(\hat{v}) - S_{k_m}(\hat{v}_{k_m})\| \to 0.$$

But (2.10) implies that

$$\||S_{k_m}(\hat{v}) - U_A^*\|| \to 0.$$

Hence, $S_{k_m}(\hat{v}_{k_m}) \to U_A^*$, a contradiction to (3.1). □

The next result concerns a technical property of the Brownian motion which is a consequence of the fact that the increments are independent and identically distributed. This property plays a fundamental role in our analysis as well as that of [EKMS], [IK], and [GIKP]. To state it, we need the following definition.

DEFINITION 3.4. *Fix $l, m \in \mathbb{N}$ and $k \in \mathbb{Z}$. An interval $[kl,(k+1)l]$ is called an (l,m)-small noise interval if*

$$\sup_{t\in[kl,(k+1)l]} \sup_{1\leq i\leq M} |W_i(t) - W_i(kl)| \leq \frac{1}{m}.$$

We have the following.

LEMMA 3.5. *For almost every path and for any $(l,m) \in \mathbb{N}\times\mathbb{N}$, there are two sequences of integers $(k_i^{l,m,\pm})_{i\in\mathbb{N}}$ such that $k_i^{l,m,\pm} \to \pm\infty$, as $i \to \infty$, and $[k_i^{l,m,\pm} l,(k_i^{l,m,\pm}+1)l]$ are (l,m)-small noise intervals.*

Proof. 1. We present the argument only for positive values of k.

2. Let

$$A_k^{l,m} = \left\{\omega : \sup_{kl\leq t\leq (k+1)l} \sup_{1\leq i\leq M} |W_i(t) - W_i(kl)| \leq \frac{1}{m}\right\}.$$

The increments $W(t)-W(kl)$ of the Brownian motion $W(t) = (W_1(t),\ldots,W_M(t))$ on the interval $[kl,(k+1)l]$ are independent and identically distributed. Hence the events $(A_k^{l,m})_{k\in\mathbb{N}}$ are independent and $\mathbb{P}(A_k^{l,m})$ is strictly positive and independent of k. The second Borel–Cantelli lemma then yields that

$$\mathbb{P}(\{\omega \in A_k^{l,m} \text{ for infinitely many } k\}) = 1. \qquad \square$$

The subset $\tilde{\Omega}$ of Ω of full measure in which our result holds consists of all of continuous paths which have, for each $(l,m) \in \mathbb{N} \times \mathbb{N}$, infinitely many (l,m)-small noise intervals for both positive and negative times. The precise definition of $\tilde{\Omega}$ is

$$\tilde{\Omega} = \Omega_C \cap_{(l,m)\in\mathbb{N}\times\mathbb{N}} (\cap_{j=1}^{\infty} \cup_{k=j}^{\infty} A_k^{l,m}) \cap (\cap_{j=1}^{\infty} \cup_{k=j}^{\infty} A_{-k}^{l,m}).$$

Next we use Lemmas 2.4, 3.2, 3.3, and 3.5 to establish the following.

LEMMA 3.6. *Fix $\omega \in \widetilde{\Omega}$, t_0 and $\delta > 0$. There exists $k_0 = k_0(\omega) \in \mathbb{N}$ such that for all $k \geq k_0(\omega)$ and $u,\ v \in C(\mathbb{T})$,*

$$|||S^{W,A}(t_0, t_0 - k)(u) - S^{W,A}(t_0, t_0 - k)(v)||| \leq \delta, \text{ and}$$
$$|||S^{W,A}(t_0 + k, t_0)(u) - S^{W,A}(t_0 + k, t_0)(v)||| \leq \delta.$$

Proof. 1. Since both estimates are proved similarly, here we establish only the second.

2. Lemma 3.3 yields an $M > 0$ such that for any initial datum $\hat{u}$ and any $m \in \mathbb{N}$,

$$|||S^{0,A}(m + M, m)(\widehat{v}) - U_A^*||| < \delta/4. \tag{3.2}$$

Here we use the fact that, since the deterministic equation is independent of time,

$$\sup_{\hat{v}} |||S^{0,A}(m + M, m)(\hat{v}) - U_A^*||| = \sup_{\hat{v}} |||S^{0,A}(M, 0)(\hat{v}) - U_A^*|||.$$

3. If C_A is the constant in Lemma 3.2 for the Lipschitz constant $L = \widehat{L}(1,1)$, choose $m \in \mathbb{N}$ such that $4MC_A\|F\| < \delta m$ and recall that Lemma 3.5 yields an $(M+1,m)$-small noise interval $[j(M+1),(j+1)(M+1)]$ contained in $(t_0,+\infty)$.

Fix $k_0(\omega)$ such that $t_0 + k_0(\omega) > (j+1)M$. It follows that the small noise interval is contained in $(t_0, t_0 + k_0(\omega))$].

4. Let $t_M^- = j(M+1)$ and $t_M^+ = (j+1)(M+1)$. Since $[t_M^-, t_M^- + 1]$ is contained in the small noise interval, Lemma 2.4 asserts that

$$u_0 = S^{W,A}(t_M^- + 1, t_0)(u) \quad \text{and} \quad v_0 = S^{W,A}(t_M^- + 1, t_0)(v)$$

are Lipschitz continuous with Lipschitz constant $L = \widehat{L}(1,1)$.

Applying again Lemma 2.4, we find that the last statement holds on the entire interval $[t_M^- + 1, (t_M^- + 1) + M]$, which has length M.

5. Using (3.2) and Lemma 3.2, we find

$$\begin{aligned}
&|||S^{W,A}(t_M^+, t_M^- + 1)(u_0) - S^{W,A}(t_M^+, t_M^- + 1)(v_0)||| \\
&\quad \leq |||S^{W,A}(t_M^+, t_M^- + 1)(u_0) - S^{0,A}(t_M^+, t_M^- + 1)(u_0)||| \\
&\qquad + |||S^{W,A}(t_M^+, t_M^- + 1)(v_0) - S^{0,A}(t_M^+, t_M^- + 1)(v_0)||| \\
&\qquad + |||S^{0,A}(t_M^+, t_M^- + 1)(u_0) - U_A^*||| + |||S^{0,A}(t_M^+, t_M^- + 1)(v_0) - U_A^*||| \\
&\quad \leq 4(\delta/4).
\end{aligned}$$

The contraction property guarantees now that the estimate holds for all later times $t > t_M^+$. ◻

Next we construct the global attracting solution u_{inv}^A.

LEMMA 3.7. *Fix $\omega \in \widetilde{\Omega}$. For all $u_0 \in C(T)$ and all $t \in \mathbb{R}$, the limit*

$$\widetilde{u}(\cdot, t) = \lim_{k\to\infty} S^{W,A}(t, -k)(u_0)(\cdot) \tag{3.3}$$

exists in $C(\mathbb{T})$ *and is unique up to constants. Moreover, for any* $t_1 < t_2$*, there exists* $c(t_1, t_2) \in \mathbb{R}$ *such that*

$$S^{W,A}(t_2, t_1)(\tilde{u}(t_1)) = \widetilde{u}(t_2) + c(t_2, t_1). \tag{3.4}$$

Proof. 1. Lemma 3.6 yields that the family $(u_k(\cdot, t))_{k\in\mathbb{N}}$ defined by

$$u_k(\cdot, k) = S^{W,A}(t, -k)(u_0)(\cdot)$$

is a Cauchy sequence with respect to the seminorm $|||\cdot|||$ for each fixed t. Therefore there exist constants $c_k(t)$ such that the sequence $u_k(\cdot, t) - c_k(t)$ converges in $C(\mathbb{T})$.

2. The identity (3.4) is a consequence of the C^0-continuity of the semigroup. ◻

We are now in a position to present the proof of Theorem 2.3.

Proof. In view of Lemma 3.6 and Lemma 3.7, it remains to show that there exists $c(t)$ such that the function $\tilde{v} = \tilde{u} - c$ satisfies, for all $t_1 < t_2$,

$$S^{W,A}(t_2, t_1)(\widetilde{v}(\cdot, t_1))(\cdot) = \widetilde{v}(\cdot, t_2).$$

Let $t_1 < t_2 < 0$. The semigroup property and (3.4) yield

$$S^{W,A}(0, t_1)(\widetilde{u}(\cdot, t_1)) = S^{W,A}(0, t_2)\left(\widetilde{u}(\cdot, t_2) + c(t_2, t_1)\right).$$

It follows that

$$c(t_2, t_1) = c(0, t_1) - c(0, t_2).$$

Similar expressions for $t_2 < 0 < t_1$ and $t_2,\ t_1 > 0$ yield the existence of a solution on $(-\infty, \infty)$ by setting

$$u_{inv}(x, t) = \widetilde{u}(x, t) + c(t),\ \ c(t) = c(\max\{t, 0\}, \min\{t, 0\}). \qquad ◻$$

4. The proof of the Lipschitz bounds. The proof of Lemma 2.4 is long and technical. To simplify the presentation, we divide it into a number of lemmas.

We remind the reader that the sole purpose of assumptions (2.8) and (2.9) is to ensure that the Hamiltonian in (2.1), which arises after incorporating the noise, still satisfies the growth assumptions (2.3), (2.4), (2.5) in the nonviscous case and (2.3) and (2.6), (2.7) in the viscous case, with constants depending on the noise only through the expression in (2.11). Therefore, we will usually omit the dependence of the Hamiltonian in (2.1) on t and ω, thus keeping the notation simple.

The first step towards the universal Lipschitz bound is a universal L^∞-bound for nonnegative solutions. This is the object of the following lemma.

LEMMA 4.1. *Fix* $\omega \in \Omega_C$*,* $u_0 \in C(\mathbb{T})$*, and* $s \in \mathbb{R}$ *and assume* (2.3) *and* $A \equiv 0$*. Let* u *be the solution of* (2.1) *on* $\mathbb{R} \times [s, T]$ *with* $u(\cdot, s) = u_0$*. For all* $t \geq s$*, there exists a positive constant* $C(s, t, \omega)$*, which is independent of the initial datum* u_0 *and depends on* ω *only through the expression in* (2.11)*, such that*

$$\|u(\cdot, t) - \min u_0\| \leq C(s, t, \omega).$$

Proof. 1. If H satisfies (2.3), a straightforward calculation yields that so does $\bar{H}(p, x, t) = H(p + DW(x, t, s, \omega), x)$ with a constant depending on $\|W\|_{C^\infty(\mathbb{T}\times[0,T])}$.

Without loss of generality, we may assume that $u_0(0) = \min_{\mathbb{T}} u_0 = 0$ and $s = 0$. The extension to the general case is straightforward.

2. For sufficiently large $C = C(K) > 0$, the function

$$g(x,t) = C|x|^{q/(q-1)} t^{-1/(q-1)} + Kt + 1$$

is a supersolution of (2.1).

Indeed, for C large,

$$\begin{aligned} &-C(q-1)^{-1}(|x|t^{-1})^{q/q-1} + K + H(Cq(q-1)^{-1}(|x|t^{-1})^{1/q-1}D|x|, x) \\ &\quad \geq C(q-1)^{-1}(|x|t^{-1})^{q/q-1}(K^{-1}q(qC)^{q-1}(q-1)^{1-q} - 1) - K + K \geq 0. \end{aligned}$$

For t small enough we clearly have $g(\cdot,t) \geq u(\cdot,t)$. Since the infimum of a family of supersolutions is also a supersolution, it follows that

$$\bar{g}(x,t) = \inf_{z \in \mathbb{Z}^n} g(x - z, t)$$

is a periodic supersolution of (2.1). ◻

When $A \not\equiv 0$, a universal L^∞-bound for nonnegative solutions is available only for Hamiltonians H with superquadratic growth in p. Indeed, we have the following.

Lemma 4.2. *Fix $\omega \in \Omega_C$ and $u_0 \in C(\mathbb{T})$ and assume that* (2.3) *holds with $q > 2$. Let u solve* (2.1) *on $\mathbb{R}^n \times [s,T]$ with $u(\cdot,s) = u_0$. For $(s,t) \in \Delta$ there exists a constant $C(s,t,\omega)$, independent of the initial datum and depending on ω only via* (2.11)*, such that*

$$\|u(\cdot,t) - \min u_0\| \leq C(s,t,\omega).$$

Before we present the proof, we remark that it is not expected, as follows from the discussion below, to have a universal bound on the L^∞-norm for nonnegative solutions of the viscous Hamilton–Jacobi equations with quadratic or subquadratic growth H. Indeed, for $c > 0$, consider the function $u_c : \mathbb{R}^n \times [0,T] \to \mathbb{R}$ defined by

$$u_c(x,t) = \frac{n}{2}\ln(t+c) + (4(t+c))^{-1}|x|^2 - \frac{n}{2}\ln(c),$$

which is an exact nonnegative solution to

$$u_t - \Delta u + |Du|^2 = 0.$$

It is immediate that, for each $c > 0$, $\min u_c(x,0) = 0$, $u_c(\cdot,t) \geq 0$ for all $t \geq 0$ and $\lim_{c\to 0} u_c(x,1) = +\infty$. However, the oscillation of $u_c(x,1)$ on each bounded subset of $\mathbb{R}^n$ is bounded uniformly in c.

The above solutions were obtained by applying the Hopf–Cole transform to fundamental solutions of the heat equation at time $t + c$. By applying the Hopf–Cole transformation to periodic solutions of the heat equation, it is possible to construct counterexamples in the periodic case in a similar way.

Now we prove Lemma 4.2.

Proof. 1. To simplify the presentation, we assume throughout the proof that $s = 0$ and write u_0 for $u(\cdot,0)$. Finally, as before, we assume that $\min u_0 = 0$.

2. Let

$$\beta = q - 2 > 0, \quad \gamma = (1-\theta)(q-2)(q-1)^{-1}, \quad \text{and} \quad \alpha = \gamma - 1 + 2\theta,$$

where $\theta \in (0, 2^{-1})$ is chosen so that $\alpha > 0$.

3. For $a, b > 0$ consider the function $G_{a,b} : \mathbb{R}^n \times (0,\infty) \to \mathbb{R}$ given by

$$G_{a,b}(x,t) = Kt + 2b \max_{\mathbb{T}} \operatorname{tr}(A) t^{\gamma} + at^{\alpha} + b\gamma|x|^2 t^{\gamma-1}.$$

It is immediate that for any $x \neq 0$, $\lim_{t\to 0} G_{a,b}(x,t) = +\infty$. Hence, for t small,

$$G_{a,b} \geq u_0.$$

4. The constants a, b may be chosen so that $G_{a,b}$ is a supersolution of (2.1). Indeed, since $D^2|x|^2 = 2I$, it remains to show only that

$$R_{a,b}(x,t) = a\alpha t^{\alpha-1} + |x|^2 t^{\gamma-2}\big(K^{-1}(2b\gamma)^q (|x|t^{-\theta})^{q-2} - \gamma(1-\gamma)b\big) > 0.$$

If $|x| \geqq t^{\theta}$, it is possible to find b, depending on q, θ, and K but not on a, so that $R_{a,b} > 0$.

If $|x| \leq t^{\theta}$, it is possible to choose a so that

$$R_{a,b}(x,t) \geq (a\alpha - \gamma(1-\gamma)b)t^{\alpha-1} > 0.$$

5. A periodic supersolution can be constructed as the infimum of supersolutions exactly as in the first-order case. $\square$

We remark that since

$$\inf_{\mathbb{T}} u(t,\cdot) \geq -Kt + \inf_{\mathbb{T}} u(0,\cdot)$$

and the equations commute with constants, Lemmas 4.1 and 4.2 yield automatically a bound on the oscillation

$$\operatorname{osc}(u(\cdot,t)) = \sup_{\mathbb{T}} u(\cdot,t) - \inf_{\mathbb{T}} u(\cdot,t).$$

Thus a bound on the oscillation is a weaker statement than the bounds on the L^{∞}-norm of nonnegative solutions asserted by the previous lemmas. We summarize these comments in the following corollary.

COROLLARY 4.3. *Fix $\omega \in \Omega_C$. Under the assumptions of either Lemma* 4.1 *or Lemma* 4.2*, there exists a positive constant $C(s,t,\omega)$, depending on ω only through $C_W(s,t,\omega)$ as in* (2.11)*, such that for all $(s,t) \in \Delta$ and $u_0 \in C(\mathbb{T})$,*

$$\operatorname{osc}(S^{W,A}(t,s)) \leq C(s,t,\omega).$$

The following lemma completes the proof of Lemma 2.4 in the first-order case.

LEMMA 4.4. *If* (2.4), (2.5), *and* (2.8) *hold and u solves* (1.1) *on $\mathbb{R}^n \times [s,T]$ with $A \equiv 0$, then for all $t \in [s,T]$, $u(\cdot,t)$ is Lipschitz continuous with a Lipschitz constant bounded by $L(s,t,\omega)$, which is nonincreasing for $s < t < s+1$ and depends only on* (2.11)*, H, and $\sup_{t'\in[s,T]} \|u(\cdot,t')\|$.*

Proof. 1. For almost all ω, there exists a $K(t,s,\omega) > 0$ such that if $|p| > K(t,s,\omega)$, there exist $B, R_0 > 0$ such that

$$\widetilde{H}(p,x,t,\omega) = H(p + DW(x,t,s,\omega), x)$$

satisfies (2.4) and (2.5) for fixed ω uniformly in $t \in [s,T]$. Again this is the place where (2.8) is used. In order to simplify notation, next we suppress the dependence of $\tilde{H}$ on t, s, and ω and write simply $\widetilde{H}(p,x)$. Finally, we choose $s = 0$.

2. Following [CLS] (note that (2.4) and (2.5) are (G2) and (3.2) in [CLS]), we consider the solution φ of

$$\varphi'(t) = \varphi(t)\Phi(\varphi(t)^{-1}), \tag{4.1}$$

where Φ is the increasing function in (2.4).

3. For $\lambda > 0$ let

$$z(x,t) = -\varphi(t)e^{-\lambda u(x,t)}.$$

It follows that

$$z_t - G(Dz, z, x) - \varphi'\varphi^{-1}z = 0,$$

where

$$G(p,z,x) = (\lambda z)\widetilde{H}(-(\lambda z)^{-1}p, x).$$

Note that if $q = -(\lambda z)^{-1}p$, then

$$D_z G(p,z,x) = \lambda\big(\widetilde{H}(q,x) - qD_p\widetilde{H}(q,x)\big)$$

and

$$D_x G(p,z,x) = \lambda z D_x\widetilde{H}(q,x) = -|p|\,|q|^{-1}D_x\widetilde{H}(q,x).$$

4. If, for some $C > 0$,

$$w(x,y,t) = z(x,t) - z(y,t) - C|x-y|$$

has a positive maximum M at (x_0, y_0, t_0), then in particular $x_0 \neq y_0$, so $|x-y|$ is smooth in a neighborhood of (x_0,y_0,t_0).

Using the definition of the viscosity solutions with $p = C(x_0 - y_0)|x_0 - y_0|^{-1}$ and noting that $p = C\hat{p}$, we find

$$\begin{aligned} 0 &\leq G(p, z(x_0,t_0), x_0) - G(p, z(y_0,t_0), y_0) + \varphi'(\varphi^{-1})(t_0)(z(x_0,t_0) - z(y_0,t_0)) \\ &= \int_0^1 [|x_0 - y_0|\hat{p}\cdot D_x G(p, z(r), x(r)) + D_z G(p, z(r), x(r))(z(x_0,t_0) - z(y_0,t_0))]\mathrm{d}r \\ &\quad + \varphi'(\varphi^{-1})(t_0)\big(z(x_0,t_0) - z(y_0,t_0)\big), \end{aligned}$$

where

$$q(r) = -(\lambda z(r))^{-1}p, \quad x(r) = y_0 + r(x_0 - y_0), \quad z(r) = z(y_0,t_0) + r(z(x_0,t_0) - z(y_0,t_0)).$$

Hence

$$\begin{aligned} 0 \leq &\int_0^1 \Big(-C|q(r)|^{-1}|x_0 - y_0|\hat{q}(r)\cdot D_x\widetilde{H}\big(q(r), x(r)\big)\Big)\mathrm{d}r \\ &+ \varphi'(\varphi^{-1})(t_0)\big(z(x_0,t_0) - z(y_0,t_0)\big) \\ &- \lambda\big(z(x_0,t_0) - z(y_0,t_0)\big)\int_0^1 \big(q(r)\cdot D_q\widetilde{H}(q(r), x(r)) - \widetilde{H}(q(r), x(r))\big)\mathrm{d}r. \end{aligned}$$

Assume next that C is such that

$$C \geq \lambda \sup_{\mathbb{T}} |z| R_0 \geq \lambda \varphi e^{\|u^-\|} R_0,$$

so that $|q| \geq R_0$ and, hence, $D_q\widetilde{H} \cdot q - \widetilde{H} \geqq 0$, and recall that $\varphi'(\varphi)^{-1} \geqq 0$.

Since by assumption

$$z(x_0, t_0) - z(y_0, t_0) \geq C|x_0 - y_0|,$$

there exists, in view of (2.5), a constant $B > 0$ such that

$$0 \leq \left(\int_0^1 (B - \lambda) g(r) \mathrm{d}r + \varphi'(t_0)(\varphi(t_0))^{-1} \right) \left(z(x_0, t_0) - z(y_0, t_0) \right),$$

where

$$g(r) = q(r) \cdot D_q\widetilde{H}(q(r), x(r)) - \widetilde{H}(q(r), x(r)).$$

Choosing $\lambda = B + 1$ and using (2.4) and (4.1), we find

$$0 \leq (z(x_0, t_0) - z(y_0, t_0)) \int_0^1 [\Phi(\varphi^{-1}(t_0)) - \Phi(|q(r)|)] \mathrm{d}r.$$

Recalling that $|q| = Ce^{\lambda u}(\lambda\varphi)^{-1}$ and that Φ is strictly increasing, we obtain, for $C > \lambda e^{\lambda \|u^-\|_\infty}$, the desired contradiction. ◻

We continue with the Lipschitz bound in the second-order case. Here we argue using the classical Bernstein method, which yields a universal Lipschitz bound depending only on the oscillation of the initial datum.

In the subquadratic but superlinear case, we will use this bound iteratively to obtain a bound for the oscillation which is independent of the initial datum (see Lemma 4.7). Of course, for a superquadratic Hamiltonian, the oscillation is easily bounded by Lemma 4.2, so the Lipschitz bound follows directly from Lemma 4.5.

To this end, let $\varphi : [s, T] \to [0, \infty)$ be a solution of the ordinary differential inequality

$$\varphi_t \leq \min(\varphi^{1/2}, 1), \quad \varphi(s) = 0. \tag{4.2}$$

LEMMA 4.5. *Let u solve* (2.1) *on $\mathbb{T} \times [s, T]$ and assume that*

$$\tilde{H}(p, x, t, s, \omega) = H(p + DW(x, t, s, \omega), x) - \operatorname{tr}\left(A(x) D^2 W(x, t, s, \omega)\right)$$

satisfies (2.3), (2.6), *and* (2.7) *on* $[s, T]$. *There exist $\kappa \in [0, 1)$ and $C_{R_0} > 0$, both independent of the initial datum $u(\cdot, s)$, such that for all $t \in [s, T]$,*

$$\|Du(\cdot, t)\| \leq \varphi(t)^{-1/2} C_{R_0} (1 + \operatorname{osc}(u(\cdot, s))^\kappa). \tag{4.3}$$

The fact that $\kappa < 1$ is very critical, since it implies that even if the oscillation is large initially, it will be much smaller at the end of the time interval. It follows from the proof that for δ as in (2.6), $\kappa(\delta) \to 1$ as $\delta \to 0$. Therefore the method does not apply to Hamiltonians with just linear growth.

Further, notice that the constants in (2.3), (2.6), and (2.7) depend on the realization of the noise in a given time interval, but only through (2.11), so they are bounded if the interval is a small noise interval.

Finally, we remark that it is straightforward to check that the particular equation

$$u_t - \epsilon\Delta u + |Du + DW(x,t,s,\omega)|^2 = 0$$

satisfies the conditions of Lemma 4.5.

For the proof of Lemma 4.5 we need a rough a priori bound on the oscillation. To this end, let

$$L(\omega) = \sup_{(x,y,t)\,\in\,\mathbb{R}^n\times\mathbb{R}^n\times[s,T]} |H(DW(x,t,s,\omega),x) - H(DW(y,t,s,\omega),y)|.$$

Note that the dependence on ω is through (2.11).

LEMMA 4.6. *For all $(s,t)\in\Delta$, we have*

$$\operatorname{osc}(u(\cdot,t)) \le \operatorname{osc}(u(\cdot,s)) + L|t-s|.$$

Proof. The estimate follows directly from the fact that

$$\operatorname{osc}(u(\cdot,t))_t = \left(\sup_{\mathbb{T}} u(\cdot,t) - \inf_{\mathbb{T}} u(\cdot,t)\right)_t \le L. \qquad \square$$

We continue with the proof of Lemma 4.5, which uses some of the techniques of [CLS].

Proof. 1. To simplify things we assume that $s=0$. The functions $v(\cdot,t) = u(\cdot,t)+Kt$ and $u(\cdot,t)$ have the same Lipschitz constant and v solves an equation with a nonnegative Hamiltonian. We may therefore assume that the Hamiltonian is nonnegative, i.e., $K=0$. Moreover, to simplify the presentation, we drop the dependence on ω and write $\tilde{H}(p,x,t)$ instead of $\tilde{H}(p,x,t,0,\omega)$. Finally, we write

$$O_0 = \operatorname{osc}(u(\cdot,0)).$$

2. Let $m(t)$ and $x_m(t)$ denote, respectively, the maximum of the function $u(\cdot,t)$ and the point where the maximum is assumed, i.e., for all $x\in\mathbb{T}$,

$$m(t) = u(x_m(t),t) \ge u(x,t).$$

Then

$$|u(x,t)-m(t)| \le \operatorname{osc}(u(\cdot,t)) \le \operatorname{diam}(\mathbb{T})\|Du(\cdot,t)\|.$$

Since $\tilde{H}\ge 0$, we know that

$$m_t(t) = u_t(x_m(t),t) = [\operatorname{tr}(A(x_m(t))D^2u(x_m(t),t)) - \tilde{H}(0,x_m(t),t)] \le 0.$$

3. For $\lambda>0$ consider the function

$$z(x,t) = \varphi(t)|Du(x,t)|^2 + \lambda(m(t)-u(x,t)).$$

Let (x_0,t_0) be a point where z achieves its maximum. The goal is to show that there exist $\lambda>0$ such that either $t_0=0$ or $|Du(x_0,t_0)|\le R_0$.

In order to keep the presentation simple, in what follows we assume that A is the identity matrix. The modifications needed for general A are straightforward, so we omit them.

4. If either $t_0 = 0$ or $|Du(x_0, t_0)| \leq R_0$, then

$$z(x,t) \leq R_0^2 + \lambda(O_0 + LT).$$

Hence, for all $(x,t) \in \mathbb{T} \times [0,T]$,

$$\varphi(t)|Du(x,t)|^2 \leq R_0^2 + \lambda(O_0 + LT) + \lambda(u(x,t) - m(t)) \leq R_0^2 + \lambda(O_0 + LT).$$

Assume that

$$O_0 \geq 1 + LT.$$

It then follows that for all $(x,t) \in \mathbb{T} \times [0,T]$,

$$\varphi(t)^{1/2}|Du(x,t)| \leq (R_0^2 + \lambda(O_0 + LT))^{1/2} \leq (R_0^2 + 2\lambda O_0)^{1/2}.$$

Since R_0 is given, we may assume that $\lambda \geq R_0$. The above estimate then can be simplified to read

$$\|Du(\cdot,t)\| \leq C\,\lambda\varphi(t)^{-1/2}(1 + \left(O_0\lambda^{-1}\right)^{1/2}). \tag{4.4}$$

5. Assume that $t_0 > 0$ and $|Du(t_0, x_0)| > R_0$. The classical calculations associated with Bernstein's method then yield the following sequence of inequalities, where C is the constant in (2.7) and where z and $\tilde{H}$ are evaluated at (x_0, t_0) and $(Du(x_0,t_0), x_0, t_0)$:

$$\begin{aligned}
0 &\leq z_t - \Delta z = \lambda m_t - \lambda(u_t - \Delta u) \\
&\quad + 2\varphi Du \cdot D(u_t - \Delta u) - 2\varphi|D^2u|^2 + \varphi_t|Du|^2 \\
&\leq \lambda\tilde{H} - 2\varphi Du \cdot D\tilde{H} - 2\varphi|D^2u|^2 + \varphi_t|Du|^2 \\
&\leq \lambda\tilde{H} - \lambda Du \cdot D_p\tilde{H} - 2\varphi Du \cdot D_x\tilde{H} + \varphi_t|Du|^2 \\
&\leq -(\lambda - C)\Phi(|Du|) + \varphi_t|Du|^2.
\end{aligned}$$

If $3\lambda \geq 4C$, then

$$0 \leq -\lambda\Phi(|\operatorname{grad} u|) + 4\varphi_t|\operatorname{grad} u|^2.$$

Dividing by $|Du|^{1+\delta}$, we obtain, always at (x_0, t_0),

$$0 \leq -\lambda G(|Du|) + 4\varphi_t|Du|^{1-\delta}.$$

Consider the set

$$D_{R_0} = \{(x,t) \in \mathbb{T} \times [0,T] \ : \ |\operatorname{grad} u(x,t)| \geq R_0\}$$

and let

$$\lambda_0 = \sup_{(x,t) \in D_{R_0}} 4\varphi_t(t)G(|Du(x,t)|)^{-1}|Du(x,t)|^{1-\delta}. \tag{4.5}$$

If we choose $\lambda > \lambda_0$, then it is impossible for the Bernstein function z to have an interior maximum, unless at the maximum we have $|Du| \leq R_0$, in which case (4.4) holds. It remains to show that λ_0 depends only on the data.

6. Let $(\bar{x}, \bar{t})$ be such that

$$\lambda_0 = 4\phi_t(\bar{t})(G(|Du(\bar{x}, \bar{t})|))^{-1}|Du(\bar{x}, \bar{t})|^{1-\delta}.$$

If such $(\bar{x}, \bar{t})$ does not exist, we argue using approximate maximizers—we leave the details to the reader. Moreover, since $\phi_t(0) = 0$, if $\lambda_0 > 0$, then $\bar{t} > 0$.

Choose $\lambda \in (\lambda_0, 2\lambda_0)$. Using (4.4) and (4.5), we find, for some universal constant $C > 0$, which is independent of λ and the initial datum, that

$$|Du(\bar{x}, \bar{t})| \leq C\, \varphi_t(\bar{t})(G(|Du(\bar{x}, \bar{t})|)\varphi(\bar{t}))^{-1/2}|Du(\bar{x}, \bar{t})|^{1-\delta}(1 + (O_0\lambda^{-1})^{1/2}).$$

Note that since $G(|Du(\bar{x}, \bar{t})|) \geq G(R_0)$ and $\varphi_t \leq \varphi^{1/2}$,

$$|Du(\bar{x}, \bar{t})|^{\delta} \leq C(1 + (O_0\lambda^{-1})^{1/2}).$$

Inserting the above in (4.5) and using (4.2) yield, for a different universal constant C,

$$\lambda_0 \leq C(1 + (O_0\lambda^{-1})^{1/2})^{(1-\delta)/\delta} \leq 2^{(1-\delta)/\delta}C(1 + (O_0{\lambda_0}^{-1})^{(1-\delta)/2\delta}).$$

7. We may assume that

$$2^{(1-\delta)/\delta+1}C \leq \lambda_0,$$

and hence

$$\lambda_0 \leq C(O_0{\lambda_0}^{-1})^{(1-\delta)/2\delta},$$

which implies

$$\lambda_0 \leq CO_0^{(1-\delta)(1+\delta)^{-1}}.$$

It follows that there exists $\rho \in (0, 1)$, independent of the initial condition, such that

$$\lambda_0 \leq CO_0^{1-\rho}.$$

Inserting a λ with $\lambda \in (\lambda_0, 2\lambda_0)$ in (4.4) yields (4.3). $\square$

We conclude with a lemma which provides a universal bound on the oscillation via a bootstrap procedure.

LEMMA 4.7. *Assume the hypotheses of Lemma* 4.5. *There exists a universal constant C, which is independent of the initial datum, such that, after time $T = 1$, the oscillation of u is bounded by C.*

Proof. 1. Since we may assume that $\varphi(t) \geq t^{\beta}$ for some $\beta > 0$, we find that if $\mathrm{osc}(u(\cdot, 0))$ is sufficiently large, then Lemma 4.5 asserts the existence of $\hat{\kappa} \in (0, 1)$ and $C > 1$ such that, after a time interval of length τ,

$$\mathrm{osc}(u(\cdot, t + \tau)) \leq C\tau^{-\beta}(\mathrm{osc}(u(\cdot, t)))^{\hat{\kappa}}.$$

If the oscillation at some time is already bounded by a power of the universal constant C, there is nothing to prove. Therefore we assume that

$$\text{if } \hat{C} = C^{2(1-\hat{\kappa})^{-1}}, \ \text{ then } \ \mathrm{osc}(u) \geq \widehat{C}.$$

If $2\kappa = (1+\hat{\kappa}) < 2$, we obtain the simpler recursion

$$\operatorname{osc}(u(\cdot, t+\tau)) \le \tau^{-\beta}(\operatorname{osc}(u(\cdot, t)))^{\kappa}.$$

2. Choose a sufficiently small $\beta_1 > 0$, let $\bar{\kappa} = \beta\beta_1 + \kappa < 1$, and consider the recursively defined sequences

$$O_l = O_{l-1}^{\bar{\kappa}} \quad \text{and} \quad \tau_l = O_{l-1}^{-\beta_1}.$$

If the numbers O_l are given by $O_l = O_0^{(\bar{\kappa})^l}$, it follows that

$$\operatorname{osc}\left(u\left(\cdot, \sum_{i=0}^{l} \tau_i\right)\right) \le \max(\widehat{C}, O_l).$$

3. Let l_M be the smallest integer such that $O_{l_M} \le 2\widehat{C}$. Then

$$O_{l_M-1} = O_0^{\bar{\kappa}^{(l_M-1)}} \ge 2\widehat{C} \quad \text{and} \quad O_{l_M} \ge (2\widehat{C})^{\bar{\kappa}}.$$

Recall that O_0 and l_M are sufficiently large, β_1 is sufficiently small, $0 \le l \le l_M$, and define

$$s_l = B^{(\bar{\kappa})^{-l}} \quad \text{and} \quad B = (O_{l_M})^{-\beta_1}.$$

We have

$$\sum_{l=0}^{l_M} \tau_l = \sum_{l=0}^{l_M} (O_0^{-\beta_1})^{(\bar{\kappa})^{l-l_M+l_M}} = \sum_{l=0}^{l_M} ((O_0^{-\beta_1})^{(\bar{\kappa}^{l_M})})^{(\bar{\kappa})^{l-l_M}} = \sum_{l=0}^{l_M} s_l,$$

and, since $\bar{\kappa} < 1$,

$$(\bar{\kappa})^{-l}((\bar{\kappa})^{-1} - 1) \ge r(\bar{\kappa}) = (\bar{\kappa})^{-1}((\bar{\kappa})^{-1} - 1) > 0.$$

Moreover

$$B = (O_{l_M})^{-\beta_1} \le (2\widehat{C})^{-\bar{\kappa}\beta_1} < 1.$$

Therefore

$$s_{l+1}{s_l}^{-1} = B^{(\bar{\kappa})^{-(l+1)} - (\bar{\kappa})^{-l}} \le B^{r(\bar{\kappa})} < 1.$$

Thus the series $\sum \tau_l$ converges by comparison with the geometric series. Note that the powers β_1, κ are independent of the length of the a priori chosen time interval. ∎

REFERENCES

[B] G. Barles, *Regularity results for first order Hamilton-Jacobi equations*, Differential Integral Equations, 3 (1990), pp. 103–125.

[BS1] G. Barles and P. E. Souganidis, *Some counterexamples on the asymptotic behavior of the solutions of Hamilton-Jacobi equations*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 963–968.

[BS2] G. Barles and P. E. Souganidis, *On the large time behavior of Hamilton–Jacobi equations*, SIAM J. Math. Anal., 31 (2000), pp. 925–939.

[BS3] G. Barles and P. E. Souganidis, *Space-time periodic solutions and long-time behavior of solutions to quasi-linear parabolic equations*, SIAM J. Math. Anal., 32 (2001), pp. 1311–1323.

[CLS] M. G. Crandall, P.-L. Lions, and P. E. Souganidis, *Maximal solutions and universal bounds for some partial differential equations of evolution*, Arch. Rational Mech. Anal., 105 (1989), pp. 163–190.

[CIL] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[DZ] G. Da Prato and J. Zabczyk, *Ergodicity for Infinite-Dimensional Systems*, London Math. Soc. Lecture Note Ser. 229, Cambridge University Press, Cambridge, UK, 1996.

[EKMS] W. E, K. Khanin, A. Mazel, and Ya. Sinai, *Invariant measures for Burgers equation with stochastic forcing*, Ann. of Math. (2), 151 (2000), pp. 877–960.

[F] A. Fathi, *Sur la convergence du semi-groupe de Lax-Oleinik*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 267–270.

[FM] A. Fathi and J. Mather, *Failure of convergence of the Lax-Oleinik semi-group in the time-periodic case*, Bull. Soc. Math. France, 128 (2000), pp. 473–483.

[FS] W. H. Fleming and S.-J. Sheu, *Stochastic variational formula for fundamental solutions of parabolic PDE*, Appl. Math. Optim., 13 (1985), pp. 193–204.

[GIKP] D. Gomes, R. Iturriaga, K. Khanin, and P. Padilla, *Convergence of viscosity solutions of Burgers equation with random forcing*, submitted.

[IK] R. Iturriaga and K. Khanin, *Burgers turbulence and random Lagrangian systems*, Comm. Math. Phys., 232 (2003), pp. 377–428.

[L] P.-L. Lions, *Regularizing effects for first-order Hamilton-Jacobi equations*, Applicable Anal., 20 (1985), pp. 283–307.

[LS1] P.-L. Lions and P. E. Souganidis, *Fully nonlinear stochastic partial differential equations: Non-smooth equations and applications*, C. R. Acad. Sci. Paris Sér. I Math., 327 (1998), pp. 735–741.

[LS2] P.-L. Lions and P. E. Souganidis, *Fully nonlinear stochastic pde with semilinear stochastic dependence*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 617–624.

[LS3] P.-L. Lions and P. E. Souganidis, *Uniqueness of weak solutions of fully nonlinear stochastic partial differential equations*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 783–790.

[M1] J. Mattingly, *Ergodicity of 2D Navier-Stokes equations with random forcing and large viscosity*, Comm. Math. Phys., 206 (1999), pp. 273–288.

[M2] J. Mattingly, *Exponential convergence for the stochastically forced Navier-Stokes equations and other partially dissipative dynamics*, Comm. Math. Phys., 230 (2002), pp. 421–462.

[NR1] G. Namah and J. M. Roquejoffre, *Remarks on the long time behaviour of the solutions of Hamilton-Jacobi equations*, Comm. Partial Differential Equations, 24 (1999), pp. 883–893.

[NR2] G. Namah and J. M. Roquejoffre, *Convergence to periodic fronts in a class of semilinear parabolic equations*, NoDEA Nonlinear Differential Equations Appl., 4 (1997), pp. 521–536.

[R] J. M. Roquejoffre, *Convergence to steady states or periodic solutions in a class of Hamilton-Jacobi equations*, J. Math. Pures Appl., 80 (2001), pp. 85–104.

[S] P. Souganidis, *Stochastic homogenization of Hamilton-Jacobi equations and some applications*, Asymptot. Anal., 20 (1999), pp. 1–11.

[W] W. Woyczyński, *Burgers-KPZ Turbulence*, Lecture Notes in Math. 1700, Springer-Verlag, Berlin, 1998.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 797–814

GLOBAL EXISTENCE OF CLASSICAL SOLUTIONS FOR THE TWO-DIMENSIONAL OLDROYD MODEL VIA THE INCOMPRESSIBLE LIMIT*

ZHEN LEI† AND YI ZHOU‡

Abstract. In this paper, we will study the Oldroyd model describing fluids with viscoelastic properties. Global classical solutions for the two-dimensional incompressible Oldroyd model with small initial displacements are shown to exist via the incompressible limit. The main difficulty is the lack of the damping mechanism on the deformation tensor.

Key words. incompressible limit, global existence, Oldroyd model

AMS subject classifications. 76A05, 76D03

DOI. 10.1137/040618813

1. Introduction. In the context of hydrodynamics, the motion of the fluid flow is classically described by a time-dependent family of orientation preserving diffeomorphisms $x(t, X)$, $0 \le t < T$. Material points X in the reference configuration are deformed to the spatial position $x(t, X)$ at time t. The deformation tensor F is defined as

$$F(t, X) = \frac{\partial x}{\partial X}(t, X).$$

When we work in the Eulerian coordinate, we define $H(t, x)$ such that $H(t, x(t, X)) = F(t, X)$. With no ambiguity, we will not distinguish these two notations and always use the notation F in this paper.

Applying the chain rule, we see that $F(t, x)$ satisfies the following transport equation (see [17], for example):

$$\partial_t F + u \cdot \nabla F = \nabla u F,$$

which stands for $\partial_t F_{ij} + u \cdot \nabla F_{ij} = \partial_k u_i F_{kj}$. We point out that in this paper we will use the notation $F_{ij} = \frac{\partial x_i}{\partial X_j}$, $(\nabla \cdot F)_i = \partial_j F_{ij}$, and $(\nabla u)_{ij} = \frac{\partial u_i}{\partial x_j}$, and summation over repeated indices will always be understood.

The nonlinear viscoelastic fluid system of the compressible Oldroyd model takes the following form:

$$(1.1) \qquad \begin{cases} \partial_t \rho + u \cdot \nabla \rho + \rho \nabla \cdot u = 0, \\ \partial_t u + u \cdot \nabla u + \lambda^2 \frac{p'(\rho)}{\rho} \nabla \rho = \frac{\mu}{\rho}(\Delta u + \nabla(\nabla \cdot u)) + \frac{1}{\rho} \nabla \cdot (\rho F F^T), \\ \partial_t F + u \cdot \nabla F = \nabla u F, \end{cases}$$

*Received by the editors November 12, 2004; accepted for publication (in revised form) March 29, 2005; published electronically November 23, 2005. This research was supported by the National Science Foundation of China (grant 10225102) and the Doctoral Program Foundation of the Ministry of Education of China.

http://www.siam.org/journals/sima/37-3/61881.html

†School of Mathematics, Fudan University, Shanghai 200433, People's Republic of China, and School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, People's Republic of China (leizhn@yahoo.com).

‡School of Mathematics, Fudan University, Shanghai 200433, People's Republic of China (yizhou@fudan.ac.cn).

where ρ is the density, u is the fluid velocity, $p(\rho)$ is a given equation of state independent of the large parameter λ with $p'(\rho) > 0$ for $\rho > 0$, and F is the deformation tensor introduced above. We emphasize that the solutions will depend on the value of the parameter λ; however, with the exception of the statements of the main theorems, the dependence will not be displayed for reasons of notational convenience.

On the other hand, the nonlinear viscoelastic fluid system of the incompressible Oldroyd model is a distinctly different system of the unknowns (u, F, q) given by

$$\begin{cases} \nabla \cdot u = 0, \\ \partial_t u + u \cdot \nabla u + \nabla q = \mu \Delta u + \nabla \cdot (FF^T), \\ \partial_t F + u \cdot \nabla F = \nabla u F, \end{cases} \tag{1.2}$$

where the density in the undeformed reference configuration has been set equal to one. The scalar pressure q, the deformation tensor F, and the fluid velocity u must be determined with u satisfying the constraint $\nabla \cdot u = 0$. For more details, one should see [14, 15].

One expects, under appropriate conditions on the initial data, that the solutions ρ^λ, $\lambda^2 \frac{1}{\rho^\lambda} \nabla p^\lambda$, u^λ, F^λ of the compressible system (1.1) converge to the solutions 1, ∇q, u, F of the incompressible system (1.2) as $\lambda \longrightarrow \infty$. It is well known that long time behavior of solutions to the viscoelastic equations depends on strong dispersive estimates [8, 9, 11, 20]. For the wave equation, the generalized energy method, based on the Lorentz invariance and global Sobolev inequalities, provides an elegant and efficient means of combining energy and decay estimates; see [8, 9], for example. Recently Sideris and Thomases [21] studied an elastodynamic system which is not Lorentz invariant in three space dimensions via the incompressible limit through the use of weighted Sobolev inequalities involving the smaller number of generators (also in [3, 4]). Because of the presence of the damping term in the momentum equation, the Oldroyd model we study in this paper is neither Lorentz invariant nor scaling invariant; thus their methods do not work in our situation. In fact, the main difficulty for system (1.2) is the lack of the damping mechanism on F. This is different from the cases studied in [16], where the contribution of the strain rate in the constitutive equation is ignored, and in [2, 7, 13], where a linear damping term is present.

On the other hand, one can easily check the energy law

$$\frac{1}{2} \frac{d}{dt} \int_\Omega (|u|^2 + |F|^2) \, dx = -\mu \|\nabla u\|^2, \tag{1.3}$$

which gives the dissipation to the whole incompressible system. This is the reason that Lin, Liu, and Zhang can prove the global existence for small initial displacements in [15]. With the aid of the identities $\det\ F = 1$ and $\nabla \cdot F^T = 0$, they obtained the estimates of the linear terms by transforming them into nonlinear terms. Motivated by their ideas, we believe that local solutions of the compressible system (1.1) should be uniform stable and converge to a global solution of the limiting incompressible system (1.2).

In this paper, we first prove that classical local solutions of the equations of motion exist for sufficiently small disturbances from the general incompressible initial data. This result depends on a modified method of [10], where Klainerman and Majda developed a general theory to study the incompressible limit of compressible fluids in general framework of quasi-linear hyperbolic systems depending on a large parameter. Their method can also be extended to cover the viscous equations. But since the deformation tensor F in system (1.2) is not "fast scale" (in the sense of [10]), our system does not satisfy their structural conditions.

We also prove the uniform stability of the local existence family which yields a lifespan of the compressible system (1.1) and allows for convergence to a global solution of the limiting incompressible equations by means of compactness arguments. The strength of this convergence improves with the degree of incompressibility satisfied by the initial data.

It is well known that many fluids do not satisfy the Newtonian law. There have been many attempts to capture different phenomena for non-Newtonian fluids; see [5, 6, 12, 16, 18, 19], for example. The fluid of Oldroyd type is one of the classical non-Newtonian fluids with memory. When additional damping mechanisms are added, many mathematical results concerning these systems are proved [2, 7, 13, 16]. When the damping mechanism on the deformation tensor F is lost, recently Lin, Liu, and Zhang [15] proved the global existence of classical solutions for the two-dimensional incompressible Oldroyd model by introducing the induced stress to find the dissipation of the system. But it is of interest to see the incompressible system mathematically justified as a limit of the slightly compressible system. However, the two-dimensional inviscid case is still open.

The paper is organized as follows. In section 2, we state the main results of this paper. In section 3, we get the dispersive energy estimates since the local existence and the uniform stability estimates of the solutions of the compressible system (1.1) have been established in [13]. The dispersive energy estimates will allow us to take the limit to obtain a global solution to the incompressible Oldroyd system in section 4.

2. Statements of main results. To avoid complications at the boundary, we concentrate below on the periodic case where $x \in \mathsf{T}^2$, the two-dimensional torus. In fact the whole space problem and the Dirichlet problem of smooth bounded domain can also be treated, at the expense of complicating the proofs below. In the following, $\|\cdot\|$, $\|\cdot\|_s$, and $\|\cdot\|_\infty$ will denote the norms in $L^2(\mathsf{T}^2)$, $H^s(\mathsf{T}^2)$, and $L^\infty(\mathsf{T}^2)$, respectively.

Define

$$E_s(U(t)) = \|\lambda(\rho-1)\|_s^2 + \|u\|_s^2 + \|F-\overline{F}\|_s^2,$$

$$\widetilde{E}_s(U(t)) = \sum_{|\alpha|\le s}\int_{\mathsf{T}^2}\left(\lambda^2\frac{p'(\rho)}{\rho}|\nabla^\alpha(\rho-1)|^2 + \rho|\nabla^\alpha u|^2 + \rho|\nabla^\alpha(F-\overline{F})|^2\right)\,dx, \tag{2.1}$$

where $U(t) = (\rho,\ u,\ F)$ and $\overline{F}$ is a constant 2×2 matrix with $\det(\overline{F}) = 1$. It is obvious that if $|\rho-1|$ is small, we have

$$E_s(U(t)) \sim \widetilde{E}_s(U(t)). \tag{2.2}$$

THEOREM 2.1. *Consider the compressible Oldroyd model* (1.1) *with the following initial datum in* $H^{s+1}(\mathsf{T}^2)$ *(integer* $s\ge 4$*):*

$$\rho^\lambda(0,x) = 1 + \widetilde{\rho}_0^\lambda(x),\quad u^\lambda(0,x) = u_0(x) + \widetilde{u}_0^\lambda(x),\quad F^\lambda(0,x) = F_0(x) + \widetilde{F}_0^\lambda(x), \tag{2.3}$$

where $u_0(x)$, $F_0(x)$ *satisfy the incompressible constraints*

$$\nabla\cdot u_0 = 0,\quad \nabla\cdot F_0^T = 0,\quad \det F_0 = 1, \tag{2.4}$$

(where $\nabla \cdot F_0^T = 0$ *means* $\partial_j (F_0)_{ji} = 0,\ i = 1, 2,$ *as was pointed out before) and* $\widetilde{\rho}_0^\lambda(x),\ \widetilde{u}_0^\lambda(x),\ \widetilde{F}_0^\lambda(x)$ *are assumed to satisfy*

$$\|\widetilde{\rho}_0^\lambda(x)\|_s \le \delta_0/\lambda^2, \quad \|\widetilde{u}_0^\lambda(x)\|_{s+1} \le \delta_0/\lambda, \quad \|\widetilde{F}_0^\lambda(x)\|_s \le \delta_0/\lambda, \quad \delta_0 \ \ \textit{small}. \tag{2.5}$$

Then the following statements hold.

Uniform stability: There exist fixed constants T_0 *and* κ *independent of* λ *such that a unique classical* C^2 *solution* $(\rho^\lambda, u^\lambda, F^\lambda)$ *of the compressible Oldroyd system* (1.1) *exists for all large* λ *on the time interval* $[0, T_0]$*. Furthermore, the solution family satisfies*

$$E_s(U^\lambda(t)) + E_{s-1}(\partial_t U^\lambda(t)) + \mu \int_0^{T_0} (\|\nabla u^\lambda\|_s^2 + \|\nabla \partial_t u^\lambda\|_{s-1}^2)\ dt \le \kappa \tag{2.6}$$

for all $t \in [0, T_0]$*. Moreover, we have*

$$E_s(U^\lambda(t)) + \mu \int_0^{T_0} \|\nabla u^\lambda\|_s^2\ dt \le 4(\|u_0\|_s^2 + \|F_0 - \overline{F}\|_s^2), \tag{2.7}$$

provided λ *is appropriately large and* $t \in [0, T_0]$*.*

Local existence for incompressible system. There exist functions $u,\ F$ *with* $\|u\|_s + \|F\|_s \le \kappa,\ t \in [0, T_0]$*, such that*

$$\begin{cases} \rho^\lambda \longrightarrow 1 \ \ in \ \ L^\infty(0, T_0; H^s) \cap Lip([0, T_0], H^{s-1}), \\ (u^\lambda, F^\lambda) \longrightarrow (u, F) \ \ weakly^* \ \ in \ \ L^\infty(0, T_0; H^s) \cap Lip([0, T_0], H^{s-1}), \\ (u^\lambda, F^\lambda) \longrightarrow (u, F) \ \ in \ \ C([0, T_0], H^{s-\epsilon}), \end{cases} \tag{2.8}$$

where ϵ *is an arbitrarily small positive constant. The function* (u, F) *is a* C^2 *solution of equations of the incompressible Oldroyd type*

$$\begin{cases} \nabla \cdot u = 0, \\ P\{\partial_t u + u \cdot \nabla u - \mu \Delta u - \nabla \cdot (F F^T)\} = 0, \\ \partial_t F + u \cdot \nabla F = \nabla u F, \end{cases} \tag{2.9}$$

with the initial datum

$$u(0, x) = u_0(x), \qquad F(0, x) = F_0(x), \tag{2.10}$$

which satisfy the constraints (2.4)*, where* P *is the* L^2*-projection on the divergence-free vector fields.*

Remark 2.1. If we denote

$$\partial_t u + u \cdot \nabla u - \mu \Delta u - \nabla \cdot (F F^T) = \nabla q,$$

then we have

$$\frac{1}{\rho^\lambda} \lambda^2 \nabla p(\rho^\lambda) \longrightarrow \nabla q \ \ weakly^* \ in \ \ L^\infty(0, T_0; H^{s-2}) \cap L^2(0, T_0; H^{s-1})$$

which means that $\lambda^2 (\|\nabla \rho^\lambda\|_{s-2} + \int_0^t \|\nabla \rho^\lambda\|_{s-1}^2\ dt)$ is uniformly bounded in $t \in [0, T_0]$.

We can use the method originally for general quasi-linear hyperbolic systems and viscous equations developed by Klainerman and Majda [10] and modified for the case when the “slow scale” F has no damping mechanism (see [13]).

THEOREM 2.2. *Consider the solutions of the two-dimensional compressible Oldroyd model obtained in Theorem* 2.1. *Suppose that the initial datum additionally satisfies*

$$\|u_0\|_s + \|F_0 - \overline{F}\|_s < \varepsilon, \tag{2.11}$$

where ε *is a positive constant and*

$$\overline{F} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

Suppose further that the disturbances from $(1,\ u_0,\ F_0)$ *satisfy* (2.5). *If* ε *is sufficiently small, then for every fixed* $\overline{T} > 0$, *the solution* $(\rho^\lambda,\ u^\lambda,\ F^\lambda)$ *satisfies the following estimates:*

$$\begin{aligned} &E_s(U^\lambda(t)) + \mu \int_0^t \|\nabla u^\lambda\|_s^2 \, dt \le C\varepsilon^2, \qquad t \in [0, T^\lambda), \\ &E_{s-1}(\partial_t U^\lambda(t)) + \mu \int_0^t \|\nabla \partial_t u^\lambda\|_{s-1}^2 \, dt \le C \exp Ct, \quad 0 \le t \le \overline{T}, \end{aligned} \tag{2.12}$$

where $T^\lambda > \overline{T}$ *and* $T^\lambda \longrightarrow \infty$ *as* $\lambda \longrightarrow \infty$.

We point out that the uniform bounds for the initial energy in (2.5) imply, in the limit as $\lambda \longrightarrow \infty$, that the initial deformation is driven toward incompressibility. Since the bounds on the energy from Theorem 2.2 are uniform in λ, we will be able to take the limit as λ goes to infinity to obtain a global solution to the equations of incompressible Oldroyd type (1.2).

THEOREM 2.3. *Consider the two-dimensional incompressible system of Oldroyd type* (1.2) *with the initial datum* (2.10) *which satisfies the constraints* (2.4) *and* (2.11). *Then there exists a unique global classical solution* (u, F) *which satisfies*

$$\|u\|_s + \|F - \overline{F}\|_s \le C\varepsilon, \tag{2.13}$$

provided ε *is sufficiently small.*

3. Dispersive energy estimates and proof of Theorem 2.2. In this section we will derive the dispersive energy estimates and prove Theorem 2.2. We refer the reader to [13] for the proof of Theorem 2.1. As a result, we have

$$\|\lambda(\rho^\lambda - 1)\|_s^2 + \|u^\lambda\|_s^2 + \|F^\lambda - \overline{F}\|_s^2 \le 4(\|u_0\|_s^2 + \|F_0 - \overline{F}\|_s^2) \le 4\varepsilon^2, \tag{3.1}$$

$$\begin{aligned} &\|\lambda \rho_t^\lambda\|_{s-1} + \|\lambda \nabla \cdot u^\lambda\|_{s-1} + \|\partial_t u^\lambda\|_{s-1} + \|\partial_t F^\lambda\|_{s-1} \\ &\qquad + |\lambda^2 \nabla \rho^\lambda\|_{s-2} + \int_0^t \|\lambda^2 \nabla \rho^\lambda\|_{s-1}^2 + \|\nabla \partial_t \dot{u}^\lambda\|_{s-1}^2 \, dt \le \kappa \end{aligned} \tag{3.2}$$

provided λ is appropriately large and $0 \le t \le T_0$. In fact, noting (2.6) and (2.7), we need only to check the bound of the term $\|\lambda \nabla \cdot u^\lambda\|_{s-1}$. We rewrite the first equation of (1.1) as

$$\nabla \cdot u^\lambda = \partial_t \rho^\lambda + u^\lambda \cdot \nabla(\rho^\lambda - 1) + (\rho^\lambda - 1)\nabla \cdot u^\lambda.$$

Thus by using the Sobolev imbedding and Lemma 3.2 below, we obtain

$$\begin{aligned}\|\lambda\nabla\cdot u^\lambda\|_{s-1} \le C(&\|\lambda\partial_t\rho^\lambda\|_{s-1} + \|\lambda(\rho^\lambda-1)\|_{s-1}\|\nabla\cdot u^\lambda\|_\infty\\ &+\|\nabla\cdot u^\lambda\|_{s-1}\|\lambda(\rho^\lambda-1)\|_\infty\\ &+\|\nabla(\rho^\lambda-1)\|_{s-1}\|u^\lambda\|_\infty + \|u^\lambda\|_{s-1}\|\nabla(\rho^\lambda-1)\|_\infty)\\ \le C(&\|\lambda\partial_t\rho^\lambda\|_{s-1} + \|\lambda(\rho^\lambda-1)\|_s),\end{aligned}$$

which gives the desired bound.

Before proving Theorem 2.2, we first show the following proposition, which is essential and will be used to get the dispersive energy estimates.

PROPOSITION 3.1. *If we set the density in the undeformed reference configuration equal to one, then we have*

$$\partial_j(\rho F_{ji}) = 0 \tag{3.3}$$

for $i = 1, 2$.

Proof. Noting the identity

$$\partial_{X_j}[(\det F)F^{-T}]_{ij} = 0, \qquad i = 1, 2,$$

we can use the conservation law of mass $\rho\cdot\det F = 1$ to get

$$\begin{aligned}\frac{1}{\rho}\nabla\cdot(\rho FF^T)_i &= \det F\frac{\partial X_k}{\partial x_j}\partial_{X_k}(\rho FF^T)_{ij}\\ &= \det F F^{-T}_{jk}\partial_{X_k}(\rho F_{il}F^T_{lj})\\ &= \partial_{X_k}(F^{-T}_{jk}F_{il}F^T_{lj})\\ &= \partial_{X_k}F_{ik}\\ &= F_{jk}\partial_j F_{ik},\end{aligned}$$

and hence

$$\partial_j(\rho F_{ji}) = 0$$

for $i = 1, 2$.

We are now ready to move to the proof of the main results of this section. As it is rather long, we divide it into four steps.

Proof of Theorem 2.2.

Step 1. Define

$$(\dot\rho, \dot u, \dot F) = (\rho - 1, u, F - \overline{F}). \tag{3.4}$$

One can rewrite system (1.1) as

$$\begin{cases}\partial_t\dot\rho + \dot u\cdot\nabla\dot\rho + \rho\nabla\cdot\dot u = 0,\\ \partial_t\dot u + \dot u\cdot\nabla\dot u + \lambda^2\frac{p'(\rho)}{\rho}\nabla\dot\rho = \frac{\mu}{\rho}(\Delta\dot u + \nabla(\nabla\cdot\dot u))\\ \qquad\qquad + \frac{1}{\rho}\nabla\cdot(\rho\dot F\dot F^T) + \frac{1}{\rho}\nabla\cdot(\rho\dot F\overline{F}^T + \rho\overline{F}\dot F^T),\\ \partial_t\dot F + \dot u\cdot\nabla\dot F = \nabla\dot u\dot F + \nabla\dot u\overline{F}.\end{cases} \tag{3.5}$$

We will use the so-called energy method. Let $U(t,x) = (\rho(t,x), u(t,x), F(t,x))$ be a local solution of the compressible Oldroyd model (1.1) obtained in Theorem 2.1. Start by applying the derivative D^α, $|\alpha| \le s$, to the system (3.5). We have

$$
\left\{
\begin{aligned}
&\partial_t D^\alpha \dot\rho + \dot u \cdot \nabla D^\alpha \dot\rho + \rho \nabla \cdot D^\alpha \dot u \\
&\qquad + [D^\alpha(\dot u \cdot \nabla \dot\rho) - \dot u \cdot \nabla D^\alpha \dot\rho] + [D^\alpha(\rho\nabla \cdot \dot u) - \rho \nabla \cdot D^\alpha \dot u] = 0,\\
&\partial_t D^\alpha \dot u + \dot u \cdot \nabla D^\alpha \dot u + \ \lambda^2 \tfrac{p'(\rho)}{\rho} \nabla D^\alpha \dot\rho \\
&\qquad - \ \tfrac{\mu}{\rho}(\Delta D^\alpha \dot u + \nabla\nabla\cdot D^\alpha \dot u) + [D^\alpha(\dot u \cdot \nabla \dot u) - \dot u \cdot \nabla D^\alpha \dot u]\\
&\qquad + \ \lambda^2[D^\alpha(\tfrac{p'(\rho)}{\rho}\nabla\dot\rho) - \tfrac{p'(\rho)}{\rho}\nabla D^\alpha \dot\rho]\\
&\quad = \tfrac{1}{\rho}\nabla\cdot D^\alpha(\rho \dot F \dot F^T) + \tfrac{1}{\rho}\nabla \cdot D^\alpha(\rho \dot F \overline{F}^T + \rho \overline{F}\dot F^T)\\
&\qquad + [D^\alpha(\tfrac{\mu}{\rho}(\Delta \dot u + \nabla\nabla\cdot \dot u)) - \tfrac{\mu}{\rho}(\Delta D^\alpha \dot u + \nabla\nabla\cdot D^\alpha \dot u)]\\
&\qquad + [D^\alpha(\tfrac{1}{\rho}\nabla\cdot(\rho\dot F\dot F^T)) - \ \tfrac{1}{\rho}\nabla\cdot D^\alpha(\rho \dot F \dot F^T)]\\
&\qquad + [D^\alpha(\tfrac{1}{\rho}\nabla\cdot(\rho \dot F\overline{F}^T + \rho\overline{F}\dot F^T)) - \tfrac{1}{\rho}\nabla\cdot D^\alpha(\rho\dot F\overline{F}^T + \rho \overline{F}\dot F^T)],\\
&\partial_t D^\alpha \dot F + \dot u \cdot \nabla D^\alpha \dot F + [D^\alpha(\dot u \cdot \nabla \dot F) - \dot u \cdot \nabla D^\alpha \dot F]\\
&\quad = \nabla D^\alpha \dot u \dot F + \nabla D^\alpha \dot u \overline{F} + [D^\alpha(\nabla \dot u \dot F) - \nabla D^\alpha \dot u \dot F].
\end{aligned}
\right. \tag{3.6}
$$

Next we proceed with the energy method by taking the L^2 inner product of (3.6) with $\lambda^2\frac{p'(\rho)}{\rho}D^\alpha\dot\rho$, $\rho D^\alpha \dot u$, and $\rho D^\alpha \dot F$, respectively. Then after integration by parts we obtain

$$
\frac{1}{2}\frac{d}{dt}\widetilde{E}(D^\alpha U(t)) + \mu(\|D^\alpha \nabla \dot u\|^2 + \|D^\alpha \nabla\cdot \dot u\|^2) = \sum_{1\le j\le 9} I_j, \tag{3.7}
$$

where

$$
I_1 = \frac{1}{2}\int_{\mathsf{T}^2}\left[\lambda^2 |D^\alpha\dot\rho|^2 \partial_t \frac{p'(\rho)}{\rho} + (|D^\alpha \dot u|^2 + |D^\alpha \dot F|^2)\partial_t \rho\right] dx, \tag{3.8}
$$

$$
I_2 = \frac{1}{2}\int_{\mathsf{T}^2}\left[\lambda^2 |D^\alpha\dot\rho|^2 \nabla\cdot\left(\frac{p'(\rho)}{\rho}\dot u\right) + (|D^\alpha \dot u|^2 + |D^\alpha \dot F|^2)\nabla\cdot(\rho \dot u)\right] dx, \tag{3.9}
$$

$$
I_3 = \lambda^2 \int_{\mathsf{T}^2} p''(\rho) D^\alpha \dot\rho D^\alpha \dot u \cdot \nabla \rho \, dx, \tag{3.10}
$$

$$
I_4 = \int_{\mathsf{T}^2}[D^\alpha \dot u_i \partial_j D^\alpha(\rho \dot F_{ik}\dot F_{jk}) + \rho D^\alpha \dot F_{ik}\partial_j D^\alpha u_i \dot F_{jk}] \, dx, \tag{3.11}
$$

$$
\begin{aligned}
I_5 = -\lambda^2 \int_{\mathsf{T}^2} &\frac{p'(\rho)}{\rho}\{[D^\alpha(\dot u \cdot \nabla\dot\rho) - \dot u \cdot \nabla D^\alpha \dot\rho] + [D^\alpha(\rho\nabla\cdot \dot u) - \rho\nabla\cdot D^\alpha \dot u]\}D^\alpha\dot\rho\\
&+ \rho\left[D^\alpha\left(\frac{p'(\rho)}{\rho}\nabla\dot\rho\right) - \frac{p'(\rho)}{\rho}\nabla D^\alpha\dot\rho\right]\cdot D^\alpha \dot u \, dx,
\end{aligned} \tag{3.12}
$$

$$
I_6 = -\int_{\mathsf{T}^2}\rho\{[D^\alpha(\dot u\cdot\nabla\dot u) - \dot u \cdot \nabla D^\alpha \dot u]\cdot D^\alpha \dot u + [D^\alpha(\dot u \cdot \nabla \dot F) - \dot u \cdot \nabla D^\alpha \dot F]_{ik} D^\alpha \dot F_{ik}\} \, dx, \tag{3.13}
$$

$$I_7 = \int_{\mathsf{T}^2} \rho \left[D^\alpha \left(\frac{\mu}{\rho}(\Delta \dot{u} + \nabla\nabla \cdot \dot{u}) \right) - \frac{\mu}{\rho}(\Delta D^\alpha \dot{u} + \nabla\nabla \cdot D^\alpha \dot{u}) \right] \cdot D^\alpha \dot{u}\ dx, \tag{3.14}$$

$$\begin{aligned} I_8 = \int_{\mathsf{T}^2} & \rho \left[D^\alpha \left(\frac{1}{\rho} \nabla \cdot (\rho \dot{F} \dot{F}^T) \right) - \frac{1}{\rho} \nabla \cdot D^\alpha (\rho \dot{F} \dot{F}^T) \right] \cdot D^\alpha \dot{u} \\ & + \rho \left[D^\alpha \left(\frac{1}{\rho} \nabla \cdot (\rho \dot{F} \overline{F}^T + \rho \overline{F} \dot{F}^T) \right) - \frac{1}{\rho} \nabla \cdot D^\alpha (\rho \dot{F} \overline{F}^T + \rho \overline{F} \dot{F}^T) \right] \\ & \cdot D^\alpha \dot{u} + \rho [D^\alpha (\nabla \dot{u} \dot{F}) - \nabla D^\alpha \dot{u} \dot{F}]_{ik} D^\alpha \dot{F}_{ik}\ dx, \end{aligned} \tag{3.15}$$

$$I_9 = \int_{\mathsf{T}^2} \nabla \cdot D^\alpha (\rho \dot{F} \overline{F}^T + \rho \overline{F} \dot{F}^T) D^\alpha \dot{u} + \rho \partial_j D^\alpha \dot{u}^i \overline{F}_{jk} D^\alpha \dot{F}_{ik}\ dx. \tag{3.16}$$

To estimate the quantities I_j, $1 \le j \le 9$, we need the following lemma.

LEMMA 3.2. *Assume $f,\ g \in H^s(\mathsf{T}^2)$. Then for any multi-index $\alpha = (\alpha_1, \dots, \alpha_n)$, $|\alpha| \le s$, we have*

$$\|\nabla^\alpha (fg)\| \le C(\|f\|_\infty \|\nabla^\alpha g\| + \|g\|_\infty \|\nabla^\alpha f\|),$$

$$\|\nabla^\alpha (fg) - f \nabla^\alpha g\| \le C(\|\nabla f\|_\infty \|g\|_{s-1} + \|g\|_\infty \|\nabla f\|_{s-1}).$$

LEMMA 3.3. *If $f : R^n \longrightarrow R$ is a smooth function, then for any positive integer s and constant $M > 0$, we have*

$$\|\nabla f(u)\|_{s-1} \le C \|\nabla u\|_{s-1}$$

for all $\|u\|_s \le M$, where C depends only on s, n, M, and f.

We refer the reader to [1, 10] for the proof of the above two lemmas. Now let us estimate the right side of (3.7) term by term.

By Sobolev imbedding,

$$\begin{aligned} |I_1| &\le C \|\partial_t \rho\|_\infty \int_{\mathsf{T}^2} [\lambda^2 |D^\alpha \dot{\rho}|^2 + |D^\alpha \dot{u}|^2 + |D^\alpha \dot{F}|^2]\ dx \\ &\le C\lambda^{-1} \|\lambda \partial_t \rho\|_{s-2} \int_{\mathsf{T}^2} [\lambda^2 |D^\alpha \dot{\rho}|^2 + |D^\alpha \dot{u}|^2 + |D^\alpha \dot{F}|^2]\ dx \\ &\le C\lambda^{-1} (\|\lambda D^\alpha \dot{\rho}\|^2 + \|D^\alpha \dot{u}\|^2 + \|D^\alpha \dot{F}\|^2). \end{aligned} \tag{3.17}$$

Similarly, we estimate I_2 and I_3 as follows:

$$\begin{aligned} |I_2| &\le C(\|\nabla \cdot \dot{u}\|_\infty + \|\nabla \dot{\rho}\|_\infty) \int_{\mathsf{T}^2} [\lambda^2 |D^\alpha \dot{\rho}|^2 + |D^\alpha \dot{u}|^2 + |D^\alpha \dot{F}|^2]\ dx \\ &\le C\lambda^{-1} (\|\lambda \nabla \dot{\rho}\|_{s-2} + \|\lambda \nabla \cdot \dot{u}\|_{s-2}) \int_{\mathsf{T}^2} [\lambda^2 |D^\alpha \dot{\rho}|^2 + |D^\alpha \dot{u}|^2 + |D^\alpha \dot{F}|^2]\ dx \\ &\le C\lambda^{-1} (\|\lambda D^\alpha \dot{\rho}\|^2 + \|D^\alpha \dot{u}\|^2 + \|D^\alpha \dot{F}\|^2), \end{aligned} \tag{3.18}$$

$$\begin{aligned} |I_3| &\le C\lambda^{-1} \|\lambda^2 \nabla \rho\|_{s-2} \int_{\mathsf{T}^2} |\lambda D^\alpha \dot{\rho}|^2 + |D^\alpha \dot{u}|^2\ dx \\ &\le C\lambda^{-1} (\|\lambda D^\alpha \dot{\rho}\|^2 + \|D^\alpha \dot{u}\|^2). \end{aligned} \tag{3.19}$$

Step 2. Estimates for I_k, with $k \geq 4$ when $D^\alpha = \nabla^\alpha$, $|\alpha| = s$ and $\alpha = 0$.

First, we let $D^\alpha = \nabla^\alpha,\ |\alpha| = s,\ s \geq 4$. We can use Lemma 3.2 and integration by parts to get

$$|I_4| \leq C\|\nabla \dot{u}\|_s \|\dot{F}\|_\infty \|\nabla^\alpha \dot{F}\| \leq C\varepsilon(\|\nabla \dot{u}\|_s^2 + \|\Delta \dot{F}\|_{s-2}^2). \tag{3.20}$$

We now use Lemmas 3.2 and 3.3 to estimate I_5 to I_8.

$$\begin{aligned} |I_5| \leq& C\lambda^{-1}\|\lambda\dot{\rho}\|_s[(\|\lambda^2\nabla\dot{\rho}\|_{s-1}\|\nabla\dot{u}\|_\infty + \|\lambda^2\nabla\dot{\rho}\|_\infty\|\nabla\dot{u}\|_{s-1}) \\ &\qquad\qquad + (\|\lambda\nabla\dot{\rho}\|_{s-1}\|\lambda\nabla\cdot\dot{u}\|_\infty + \|\lambda\nabla\dot{\rho}\|_\infty\|\lambda\nabla\dot{u}\|_{s-1})] \\ &+ C\lambda^{-1}\|\dot{u}\|_s\|\lambda^2\nabla\dot{\rho}\|_\infty\|\lambda\nabla\dot{\rho}\|_{s-1} \\ \leq& C\lambda^{-1}(\|\lambda^2\nabla\dot{\rho}\|_{s-1}^2 + \|\lambda\dot{\rho}\|_s^2 + \|\dot{u}\|_s^2). \end{aligned} \tag{3.21}$$

By a similar argument, we have

$$\begin{aligned} |I_6| &\leq C[\|\dot{u}\|_s\|\nabla\dot{u}\|_\infty\|\nabla\dot{u}\|_{s-1} + \|\nabla^\alpha\dot{F}\|(\|\nabla\dot{F}\|_\infty\|\nabla\dot{u}\|_{s-1} + \|\nabla\dot{F}\|_{s-1}\|\nabla\dot{u}\|_\infty)] \\ &\leq C\varepsilon(\|\nabla\dot{u}\|^2 + \|\nabla\dot{u}\|_s^2 + \|\Delta\dot{F}\|_{s-2}^2), \end{aligned} \tag{3.22}$$

$$|I_7| \leq C\|\dot{u}\|_s(\|\nabla\dot{\rho}\|_\infty\|\Delta\dot{u}\|_{s-1} + \|\nabla\dot{\rho}\|_{s-1}\|\Delta\dot{u}\|_\infty) \leq C\lambda^{-1}(\|\dot{u}\|_s^2 + \|\nabla\dot{u}\|_s^2), \tag{3.23}$$

$$\begin{aligned} |I_8| \leq& C\|\nabla^\alpha\dot{u}\|(\|\nabla\dot{\rho}\|_{s-1}\|\nabla\cdot(\rho\dot{F}\dot{F}^T)\|_\infty + \|\nabla\dot{\rho}\|_\infty\|\nabla\cdot(\rho\dot{F}\dot{F}^T)\|_{s-1} \\ &+ \|\nabla\dot{\rho}\|_{s-1}\|\nabla(\rho\dot{F})\|_\infty + \|\nabla\dot{\rho}\|_\infty\|\nabla(\rho\dot{F})\|_{s-1}) \\ &+ C\|\nabla^\alpha\dot{F}\|(\|\nabla\dot{u}\|_\infty\|\nabla\dot{F}\|_{s-1} + \|\nabla\dot{u}\|_{s-1}\|\nabla\dot{F}\|_\infty) \\ \leq& C\lambda^{-1}(\|\nabla^\alpha\dot{u}\|^2 + \|\lambda\dot{\rho}\|_s^2) + C\varepsilon(\|\nabla\dot{u}\|_{s-1}^2 + \|\Delta\dot{F}\|_{s-2}^2). \end{aligned} \tag{3.24}$$

At last, we estimate I_9 as follows:

$$\begin{aligned} I_9 &= \int_{\mathsf{T}^2} \nabla\cdot\nabla^\alpha\rho(\dot{F}\overline{F}^T + \overline{F}\dot{F}^T)\nabla^\alpha\dot{u} + \rho\partial_j\nabla^\alpha\dot{u}^i\overline{F}_{jk}\nabla^\alpha\dot{F}_{ik}\,dx \\ &= -\int_{\mathsf{T}^2} \overline{F}_{ik}\partial_j\nabla^{\alpha-1}(\rho\dot{F}_{jk})\nabla\nabla^\alpha\dot{u}_i \\ &\quad + [\partial_j\nabla^\alpha\dot{u}_i\nabla^\alpha(\rho\dot{F}_{ik}\overline{F}_{jk}) - \rho\partial_j\nabla^\alpha\dot{u}_i\nabla^\alpha\dot{F}_{ik}\overline{F}_{jk}]\,dx. \end{aligned}$$

Noting (3.3), we have

$$\begin{aligned} |I_9| &\leq C\lambda^{-1}(\|\nabla\dot{u}\|_s^2 + \|\lambda\dot{\rho}\|_s^2) + C\|\nabla\dot{u}\|_s(\|\nabla\dot{\rho}\|_{s-1}\|\dot{F}\|_\infty + \|\nabla^{\alpha-1}\dot{F}\|\|\nabla\dot{\rho}\|_\infty) \\ &\leq C\lambda^{-1}(\|\nabla\dot{u}\|_s^2 + \|\lambda\dot{\rho}\|_s^2 + \|\Delta\dot{F}\|_{s-2}^2). \end{aligned} \tag{3.25}$$

On the other hand, when $|\alpha| = 0$, we have $I_5 = I_6 = I_7 = I_8 = 0$. Furthermore, if we go back to (3.11) and (3.16), we find

$$I_4 = \int_{\mathsf{T}^2} [\dot{u}_i\partial_j(\rho\dot{F}_{ik}\dot{F}_{jk}) + \rho\dot{F}_{ik}\partial_j u_i\dot{F}_{jk}]\,dx = 0,$$

$$\begin{aligned} I_9 &= \int_{\mathsf{T}^2} \dot{u}_i\overline{F}_{ik}\partial_j(\rho\dot{F}_{jk}) - \rho\partial_j\dot{u}_i\dot{F}_{ik}\overline{F}_{jk} + \rho\partial_j\dot{u}_i\overline{F}_{jk}\dot{F}_{ik}\,dx \\ &= \int_{\mathsf{T}^2} \dot{u}_i\overline{F}_{ik}\partial_j(\rho\overline{F}_{jk})\,dx \\ &= -\int_{\mathsf{T}^2} \dot{u}_i\overline{F}_{ik}\partial_j(\dot{\rho}\overline{F}_{jk})\,dx. \end{aligned}$$

Thus, we have

$$|I_9| \le C\lambda^{-1}(\|\nabla\dot u\|^2 + \|\lambda\dot\rho\|^2). \tag{3.26}$$

By adding up all these estimates (3.17)–(3.26), we have thus far obtained

$$\begin{aligned}&\frac{d}{dt}[\widetilde{E}(U(t)) + \widetilde{E}(\nabla^s U(t))] + \mu(\|\nabla\dot u\|_s^2 + \|\nabla\cdot\dot u\|_s^2)\\ &\quad\le C\lambda^{-1}(\|\lambda^2\nabla\dot\rho\|_{s-1}^2 + E_s(U(t)) + C\varepsilon\|\Delta\dot F\|_{s-2}^2).\end{aligned} \tag{3.27}$$

At this stage, it is clear we need to estimate the second term on the right side of the above inequality.

We introduce

$$w = u - \frac{1}{\mu}\Delta^{-1}\nabla^\perp\cdot(\rho\dot F), \tag{3.28}$$

where $\nabla^\perp = (-\partial_2, \partial_1)^T$. Noting (3.2), (3.3), and

(3.29)
$$\nabla\nabla^\perp\cdot\rho\dot F = \Delta\begin{pmatrix}\rho\dot F_{21} & -\rho\dot F_{11}\\ \rho\dot F_{22} & -\rho\dot F_{12}\end{pmatrix} + \begin{pmatrix}-\partial_{2i}(\rho\dot F_{i1}) & \partial_{1i}(\rho\dot F_{i1})\\ -\partial_{2i}(\rho\dot F_{i2}) & \partial_{1i}(\rho\dot F_{i2})\end{pmatrix} + \nabla^2(\rho\dot F_{12} - \rho\dot F_{21}),$$

we find that

$$\|\Delta\dot F\| \le C(\|\nabla\Delta\dot u\| + \|\nabla\Delta w\|) + C\lambda^{-1}\|\lambda\dot\rho\|_2 + \|\nabla^2(\rho\dot F_{12} - \rho\dot F_{21})\|, \tag{3.30}$$

(3.31)
$$\|\nabla^{s-2}\Delta\dot F\| \le C(\|\nabla\Delta\dot u\|_{s-2} + \|\nabla\Delta w\|_{s-2}) + C\lambda^{-1}\|\lambda\dot\rho\|_s + \|\nabla^2(\rho\dot F_{12} - \rho\dot F_{21})\|_{s-2}.$$

From the conservation law of mass $\rho\cdot\det(\dot F + \overline{F}) = 1$ we have

$$\rho\dot F_{21} - \rho\dot F_{12} = 1 - \rho + \rho\dot F_{12}\dot F_{21} - \rho\dot F_{11}\dot F_{22}. \tag{3.32}$$

Thus, we have

$$\|\nabla^2(\rho\dot F_{12} - \rho\dot F_{21})\|_{s-2} \le C\lambda^{-1}\|\lambda\dot\rho\|_s + C\varepsilon\|\Delta\dot F\|_{s-2}. \tag{3.33}$$

Combining (3.30), (3.31), (3.32), and (3.33), we obtain

$$\|\Delta\dot F\|_{s-2} \le C\lambda^{-1}\|\lambda\dot\rho\|_s + C(\|\nabla\Delta\dot u\|_{s-2} + \|\nabla\Delta w\|_{s-2}). \tag{3.34}$$

Now, substituting (3.34) into (3.27), we finally arrive at

$$\begin{aligned}&\frac{d}{dt}[\widetilde{E}(U(t)) + \widetilde{E}(\nabla^s U(t))] + \mu(\|\nabla\dot u\|_s^2 + \|\nabla\cdot\dot u\|_s^2)\\ &\quad\le C\lambda^{-1}(\|\lambda^2\nabla\dot\rho\|_{s-1}^2 + E_s(U(t)) + C\varepsilon\|\nabla\Delta w\|_{s-2}^2).\end{aligned} \tag{3.35}$$

To get the dispersive a priori energy estimates of the solutions of the compressible Oldroyd system (1.1), it is clear we also need an equation of w.

Step 3. Estimates for $\|\Delta w\|_{s-2}$ and $E_s(U(t))$.

With the aid of (3.3), one can rewrite the momentum equation (the second equation of (3.5)) as

$$\partial_t \dot u + \dot u\cdot\nabla\dot u + \lambda^2\frac{1}{\rho}\nabla p = \frac{\mu}{\rho}\Delta w + \frac{\mu}{\rho}\nabla\nabla\cdot\dot u + \frac{1}{\rho}\nabla\cdot(\rho\dot F\dot F^T) - \frac{1}{\rho}\nabla\rho. \tag{3.36}$$

On the other hand, from the first and third equations of (3.5) we have

$$\partial_t(\rho\dot F) + \dot u\cdot\nabla(\rho\dot F) = \nabla\dot u(\rho\dot F) + \nabla\dot u\rho\overline{F} - \nabla\cdot\dot u(\rho\dot F). \tag{3.37}$$

By applying Δ to (3.36) and $-\frac{1}{\mu}\nabla^\perp\cdot$ to (3.37) and then adding up the resulting equation, we find

$$\partial_t\Delta w + \dot u\cdot\nabla\Delta w + \lambda^2\frac{1}{\rho}\nabla\Delta p - \frac{\mu}{\rho}\Delta^2 w - \frac{\mu}{\rho}\nabla\nabla\cdot\Delta\dot u = h, \tag{3.38}$$

where

$$\begin{aligned} h_i =& -[\Delta(\dot u\cdot\nabla\dot u_i) - \dot u\cdot\nabla\Delta\dot u_i] - \lambda^2\left[\Delta\left(\frac{1}{\rho}\partial_i p\right) - \frac{1}{\rho}\partial_i\Delta p\right] \\ &+ \frac{1}{\mu}[\partial_1\dot u_k\partial_k(\rho\dot F_{i2}) - \partial_2\dot u_k\partial_k(\rho\dot F_{i1})] \\ &+ \left[\Delta\left(\frac{\mu}{\rho}\Delta w_i\right) - \frac{\mu}{\rho}\Delta^2 w_i\right] + \left[\Delta\left(\frac{\mu}{\rho}\partial_i\nabla\cdot\dot u\right) - \frac{\mu}{\rho}\partial_i\nabla\cdot\Delta\dot u\right] \\ &+ \Delta\left[\frac{1}{\rho}\partial_j(\rho\dot F\dot F^T)_{ij} - \frac{1}{\rho}\partial_i\rho\right] - \frac{1}{\mu}\nabla_j^\perp[\nabla\dot u\rho\dot F + \nabla\dot u\rho\overline{F} - \nabla\cdot\dot u\rho\dot F]_{ij} \end{aligned} \tag{3.39}$$

for $i = 1, 2$.

By taking the L^2 inner product of (3.38) with $\rho\Delta w$, we can use integration by parts to obtain

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|\rho\Delta w\|^2 + \mu\|\nabla\Delta w\|^2 =& \frac{1}{2}\int_{\mathbb{T}^2}(\nabla\cdot(\rho\dot u) + \partial_t\rho)|\Delta w|^2dx - \mu(\nabla\cdot\Delta\dot u, \nabla\cdot\Delta w) \\ &+ (\lambda^2\Delta p, \Delta\nabla\cdot w) + (\rho h, \Delta w). \end{aligned} \tag{3.40}$$

We can use (3.2) to get

$$\left|\frac{1}{2}\int_{\mathbb{T}^2}(\nabla\cdot(\rho\dot u + \partial_t\rho))|\Delta w|^2dx\right| \le C\lambda^{-1}\|\Delta w\|^2. \tag{3.41}$$

On the other hand, from a direct computation we have

$$\nabla\cdot(\nabla^\perp\cdot\rho\dot F) = \partial_{1i}\dot F_{i2} - \partial_{2i}\dot F_{i1} = \Delta\rho. \tag{3.42}$$

Thus we have

$$-\mu(\nabla\cdot\Delta\dot u, \nabla\cdot\Delta w) \le -\mu\|\nabla\cdot\Delta\dot u\|^2 + C\lambda^{-1}(\|\lambda\dot\rho\|_2^2 + \|\nabla\Delta\dot u\|^2). \tag{3.43}$$

We now estimate the term $(\rho h, \Delta w)$ in (3.40) where h is given in (3.39). First, we use integration by parts to get

$$\begin{aligned} &\left|(-[\Delta(\dot u\cdot\nabla\dot u_i) - \dot u\cdot\nabla\Delta\dot u_i] - \frac{1}{\mu}\nabla_j^\perp[\nabla\dot u\rho\dot F + \nabla\dot u\rho\overline{F} - \nabla\cdot\dot u\rho\dot F]_{ij}, \rho\Delta w_i)\right| \\ &\le \|\Delta w\|_\infty\|\nabla^2\dot u\|\|\nabla\dot u\| + C\|\nabla\Delta w\|\|\nabla\dot u\|\|\dot F\|_\infty + \frac{1}{\mu}\|\nabla\Delta w\|\|\nabla\dot u\| \\ &\le C\varepsilon(\|\nabla\dot u\|^2 + \|\nabla^2\dot u\|^2) + \frac{\mu}{2}\|\nabla\Delta w\|^2 + \frac{2}{\mu^3}\|\nabla\dot u\|^2. \end{aligned} \tag{3.44}$$

We can use the integration by parts to estimate the third term of $(\rho h, \Delta w)$ as follows:

$$|([\partial_1 \dot{u}_k \partial_k \dot{F}_{2i} - \partial_2 \dot{u}_k \partial_k \dot{F}_{1i}], \rho\Delta w_i)| \leq C\varepsilon(\|\nabla \dot{u}\|^2 + \|\nabla\Delta w\|^2) + C\lambda^{-1}(\|\Delta w\|^2 + \|\dot{F}\|^2). \tag{3.45}$$

Finally, with the aid of (3.34), we estimate the rest terms of $(\rho h, \Delta w)$ as follows:

$$\begin{aligned}
&\bigg(-\lambda^2\left[\Delta\left(\frac{1}{\rho}\partial_i p\right) - \frac{1}{\rho}\partial_i\Delta p\right] + \left[\Delta\left(\frac{\mu}{\rho}\Delta w_i\right) - \frac{\mu}{\rho}\Delta^2 w_i\right]\\
&\quad + \Delta\left[\frac{1}{\rho}\partial_j(\rho\dot{F}\dot{F}^T)_{ji} + \frac{1}{\rho}\partial_i\rho\right], \rho\Delta w_i\bigg)\\
&\leq C\lambda^{-1}\|\Delta w\|(\|\lambda^2\nabla^2\rho\|\|\lambda\nabla\rho\|_\infty + \|\nabla\Delta w\|\|\lambda\nabla\rho\|_\infty + \|\Delta w\|_\infty\|\lambda\nabla^2\rho\|)\\
&\quad + C\varepsilon\|\nabla\Delta w\|\|\Delta\dot{F}\| + C\lambda^{-1}\|\nabla\Delta w\|(\|\lambda\Delta\rho\| + \|\lambda\nabla\rho\|\|\lambda\nabla\rho\|_\infty)\\
&\leq C\lambda^{-1}(\|\Delta w\|^2 + \|\lambda\dot{\rho}\|_2^2 + \|\nabla\Delta w\|^2) + C\varepsilon(\|\nabla\Delta w\|^2 + \|\nabla\Delta\dot{u}\|^2).
\end{aligned} \tag{3.46}$$

Summing up (3.44)–(3.46), we obtain

$$\begin{aligned}
|(\rho h, \Delta w)| &\leq C\varepsilon(\|\nabla\dot{u}\|^2 + \|\nabla\Delta\dot{u}\|^2 + \|\nabla\Delta w\|^2)\\
&\quad + C\lambda^{-1}(\|\Delta w\|^2 + \|\lambda\dot{\rho}\|_2^2 + \|\dot{F}\|^2) + \frac{\mu}{2}\|\nabla\Delta w\|^2 + \frac{2}{\mu^3}\|\nabla\dot{u}\|^2.
\end{aligned} \tag{3.47}$$

To estimate the term $(\lambda^2\Delta p, \Delta\nabla\cdot w)$, we use the decomposition

$$(\lambda^2\Delta p, \Delta\nabla\cdot w) = (\lambda^2\Delta p, \Delta\nabla\cdot\dot{u}) - \frac{1}{\mu}(\lambda^2\Delta p, \nabla\cdot(\nabla^\perp\cdot\rho\dot{F})). \tag{3.48}$$

By multiplying the second equation of system (3.5) by ρ and then applying the divergence operator to the resulting equation, we obtain

$$\lambda^2\Delta p = -\nabla\cdot(\rho\partial_t\dot{u}) + 2\mu\Delta\nabla\cdot\dot{u} - \nabla\cdot(\rho\dot{u}\cdot\nabla\dot{u}) + \nabla\cdot(\nabla\cdot(\rho\dot{F}\dot{F}^T)) - 2\Delta\rho, \tag{3.49}$$

where we used the equality (3.42). Substituting (3.49) into (3.48), one obtains

$$\begin{aligned}
|(\lambda^2\Delta p, \Delta\nabla\cdot w)| &= \bigg|\bigg(-\nabla\cdot(\rho\partial_t\dot{u}) + 2\mu\Delta\nabla\cdot\dot{u} - \nabla\cdot(\rho\dot{u}\cdot\nabla\dot{u})\\
&\qquad + \nabla\cdot(\nabla\cdot(\rho\dot{F}\dot{F}^T)) - 2\Delta\rho, \Delta\nabla\cdot\dot{u} - \frac{1}{\mu}\Delta\rho\bigg)\bigg|\\
&\leq 2\mu\|\Delta\nabla\cdot\dot{u}\|^2 + C\lambda^{-1}(\|\lambda\dot{\rho}\|_2^2 + \|\lambda\Delta\nabla\cdot\dot{u}\|^2 + \|\Delta\dot{F}\|^2\\
&\quad + \|\nabla\cdot\partial_t\dot{u}\|^2 + \|\nabla\nabla\cdot\dot{u}\|^2) + C\varepsilon(\|\Delta\dot{F}\|^2 + \|\nabla\dot{u}\|^2)\\
&\leq 2\mu\|\Delta\nabla\cdot\dot{u}\|^2 + C\lambda^{-1}\|\lambda\dot{\rho}\|_2^2\\
&\quad + C\varepsilon(\|\nabla\dot{u}\|^2 + \|\nabla\Delta\dot{u}\|^2 + \|\nabla\Delta w\|^2) + C\lambda^{-1},
\end{aligned} \tag{3.50}$$

where in the last inequality we used (3.2).

Finally, substituting (3.41), (3.43), (3.47), and (3.50) into (3.40), we arrive at

$$\begin{aligned}
\frac{d}{dt}\|\rho\Delta w\|^2 + \mu\|\nabla\Delta w\|^2 &\leq C\lambda^{-1}(\|\Delta w\|^2 + \|\lambda\dot{\rho}\|_2^2 + \|\dot{F}\|^2) + C\varepsilon(\|\nabla\dot{u}\|^2 + \|\nabla\Delta\dot{u}\|^2)\\
&\quad + \frac{4}{\mu^3}\|\nabla\dot{u}\|^2 + 2\mu\|\Delta\nabla\cdot\dot{u}\|^2 + C\lambda^{-1}.
\end{aligned} \tag{3.51}$$

To estimate the higher order derivatives of w, we apply ∇^{s-2} to (3.38) to obtain

$$\partial_t\nabla^{s-2}\Delta w+\nabla^{s-2}(\dot u\cdot\nabla\Delta w)+\lambda^2\nabla^{s-2}\left(\frac{1}{\rho}\nabla\Delta p\right)-\nabla^{s-2}\left(\frac{\mu}{\rho}\Delta^2 w+\frac{\mu}{\rho}\Delta\nabla\nabla\cdot\dot u\right)$$
$$=\nabla^{s-2}h.$$

Then by taking the L^2 inner product of the resulting equation with $\rho\nabla^{s-2}\Delta w$, we have

(3.52)
$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\|\rho\nabla^{s-2}\Delta w\|^2+\mu\|\nabla^{s-2}\nabla\Delta w\|^2\\
&=\mu(\nabla^{s-2}\nabla\nabla\cdot\Delta\dot u,\nabla^{s-2}\Delta w)+\frac{1}{2}\int_{\mathbb{T}^2}(\nabla\cdot(\rho\dot u)+\partial_t\rho)|\nabla^{s-2}\Delta w|^2dx\\
&+(\nabla^{s-2}h,\rho\nabla^{s-2}\Delta w)-(\nabla^{s-2}(\dot u\nabla\cdot\Delta w)-\dot u\nabla^{s-2}\nabla\cdot\Delta w,\rho\nabla^{s-2}\Delta w)\\
&+\left(\nabla^{s-2}\left(\frac{\mu}{\rho}\Delta^2w+\frac{\mu}{\rho}\Delta\nabla\nabla\cdot\dot u\right)-\left(\frac{\mu}{\rho}\nabla^{s-2}\Delta^2w+\frac{\mu}{\rho}\nabla^{s-2}\Delta\nabla\nabla\cdot\dot u\right),\rho\nabla^{s-2}\Delta w\right)\\
&+\lambda^2(\nabla^{s-2}\Delta p,\nabla^{s-2}\Delta\nabla\cdot w)-\lambda^2\left(\nabla^{s-2}\left(\frac{1}{\rho}\nabla\Delta p\right)-\frac{1}{\rho}\nabla^{s-2}\nabla\Delta p,\rho\nabla^{s-2}\Delta w\right).
\end{aligned}$$

In what follows, we will estimate the above terms separately. First, similar to getting (3.41) and (3.43), we have

$$\begin{aligned}
&\mu(\nabla^{s-2}\nabla\nabla\cdot\Delta\dot u,\nabla^{s-2}\Delta w)+\frac{1}{2}\int_{\mathbb{T}^2}(\nabla\cdot(\rho\dot u)+\partial_t\rho)|\nabla^{s-2}\Delta w|^2dx\\
&\quad\le-\mu\|\nabla^{s-2}\nabla\cdot\Delta\dot u\|^2+C\lambda^{-1}(\|\lambda\dot\rho\|_s^2+\|\nabla^{s-2}\Delta w\|^2+\|\nabla^{s-2}\nabla\Delta\dot u\|^2).
\end{aligned}\tag{3.53}$$

By using Lemma 3.2, we can estimate the second term of the third line (3.52) as

$$(\nabla^{s-2}(\dot u\nabla\cdot\Delta w)-\dot u\nabla^{s-2}\nabla\cdot\Delta w,\rho\nabla^{s-2}\Delta w)\le C\varepsilon\|\nabla\Delta w\|_{s-2}^2.\tag{3.54}$$

By integration by parts, we can estimate the fourth line of (3.52) as follows:

$$\begin{aligned}
&\left|\left(\nabla^{s-2}\left(\frac{\mu}{\rho}\Delta^2w+\frac{\mu}{\rho}\Delta\nabla\nabla\cdot\dot u\right)-\left(\frac{\mu}{\rho}\nabla^{s-2}\Delta^2w+\frac{\mu}{\rho}\nabla^{s-2}\Delta\nabla\nabla\cdot\dot u\right),\rho\nabla^{s-2}\Delta w\right)\right|\\
&=\left|\left(\nabla^{s-2}\left(\nabla\frac{\mu}{\rho}\Delta\nabla w+\nabla\frac{\mu}{\rho}\nabla\nabla\nabla\cdot\dot u\right)\right.\right.\\
&\quad\left.\left.-\left(\nabla\frac{\mu}{\rho}\nabla^{s-2}\nabla\Delta w+\nabla\frac{\mu}{\rho}\nabla^{s-2}\nabla\nabla\nabla\cdot\dot u\right),\rho\nabla^{s-2}\Delta w\right)\right|\\
&\quad+\left|\left(\nabla^{s-2}\left(\frac{\mu}{\rho}\Delta\nabla w+\frac{\mu}{\rho}\nabla\nabla\nabla\cdot\dot u\right)\right.\right.\\
&\quad\quad\left.\left.-\left(\frac{\mu}{\rho}\nabla^{s-2}\nabla\Delta w+\frac{\mu}{\rho}\nabla^{s-2}\nabla\nabla\nabla\cdot\dot u\right),\nabla\left(\rho\nabla^{s-2}\Delta w\right)\right)\right|\\
&\le C\lambda^{-1}(\|\lambda\dot\rho\|_s^2+\|\nabla\Delta w\|_{s-2}^2).
\end{aligned}$$
(3.55)

Noting that $p = p(\rho)$, by using Lemmas 3.2 and 3.3, we can similarly estimate the last line of (3.52) as follows:

$$
\begin{aligned}
&\left|-\lambda^2\left(\nabla^{s-2}\nabla\left(\frac{1}{\rho}\Delta p\right)-\frac{1}{\rho}\nabla^{s-2}\nabla\Delta p,\rho\nabla^{s-2}\Delta w\right)\right| \\
&= \lambda^2\left|\left(\nabla^{s-2}\nabla\left(\nabla\frac{1}{\rho}\nabla p\right)-\nabla\frac{1}{\rho}\nabla^{s-2}\nabla\nabla p,\rho\nabla^{s-2}\Delta w\right)\right. \\
&\quad\left.+\left(\nabla^{s-2}\nabla\left(\frac{1}{\rho}\nabla p\right)-\frac{1}{\rho}\nabla^{s-2}\nabla\nabla p,\nabla(\rho\nabla^{s-2}\Delta w)\right)\right| \\
&\le C\lambda^{-1}(\|\lambda^2\nabla\rho\|_{s-2}\|\lambda\dot\rho\|_s\|\nabla^{s-2}\Delta w\|+\|\lambda^2\nabla\rho\|_{s-2}\|\lambda\dot\rho\|_s\|\nabla^{s-2}\Delta\nabla w\|) \\
&\le C\lambda^{-1}(\|\lambda\dot\rho\|_s^2+\|\Delta\nabla w\|_{s-2}^2),
\end{aligned}
\tag{3.56}
$$

where in the last inequality we used (3.2).

In order to estimate the term $(\nabla^{s-2}h,\rho\nabla^{s-2}\Delta w)$, we go back to (3.39). It is rather easy to see that

$$
|(-\nabla^{s-2}[\Delta(\dot u\cdot\nabla\dot u_i)-\dot u\cdot\nabla\Delta\dot u_i],\rho\nabla^{s-2}\Delta w_i)|\le C\varepsilon(\|\nabla\dot u\|_s^2+\|\nabla\Delta w\|_{s-2}^2),
\tag{3.57}
$$

$$
\begin{aligned}
\left|\left(\lambda^2\nabla^{s-2}\left[\Delta\left(\frac{1}{\rho}\partial_i p\right)-\frac{1}{\rho}\partial_i\Delta p\right],\rho\nabla^{s-2}\Delta w_i\right)\right| &\le C\lambda^{-1}\|\nabla^{s-2}\nabla\Delta w\|\|\lambda^2\nabla\rho\|_{s-2}\|\lambda\dot\rho\|_s \\
&\le C\lambda^{-1}(\|\nabla^{s-2}\nabla\Delta w\|^2+\|\lambda\dot\rho\|_s^2),
\end{aligned}
\tag{3.58}
$$

$$
\left|\left(\frac{1}{\mu}\nabla^{s-2}[\partial_1\dot u_k\partial_k(\rho\dot F_{i2})-\partial_2\dot u_k\partial_k(\rho\dot F_{i1})],\rho\nabla^{s-2}\Delta w_i\right)\right|\le C\varepsilon(\|\nabla^{s-2}\nabla\Delta w\|^2+\|\nabla\dot u\|_s^2),
\tag{3.59}
$$

$$
\begin{aligned}
&\left|\left(\nabla^{s-2}\left[\Delta\left(\frac{\mu}{\rho}\Delta w_i\right)-\frac{\mu}{\rho}\Delta^2 w_i\right]+\nabla^{s-2}\left[\Delta\left(\frac{\mu}{\rho}\partial_i\nabla\cdot\dot u\right)-\frac{\mu}{\rho}\partial_i\nabla\cdot\Delta\dot u\right],\rho\nabla^{s-2}\Delta w_i\right)\right| \\
&\qquad\le C\lambda^{-1}(\|\lambda\dot\rho\|_s^2+\|\nabla\Delta w\|_{s-2}^2+\|\nabla\dot u\|_s^2),
\end{aligned}
\tag{3.60}
$$

$$
\begin{aligned}
&\left|\left(\nabla^{s-2}\Delta\left[\frac{1}{\rho}\partial_j(\rho\dot F\dot F^T)_{ij}+\frac{1}{\rho}\partial_i\rho\right],\rho\nabla^{s-2}\Delta w\right)\right| \\
&\quad\le C\varepsilon(\|\nabla\Delta w\|_{s-2}^2+\|\Delta\dot F\|_{s-2}^2)+C\lambda^{-1}(\|\lambda\dot\rho\|_s^2+\|\nabla\Delta w\|_{s-2}^2) \\
&\quad\le C\varepsilon(\|\nabla\Delta w\|_{s-2}^2+\|\nabla\dot u\|_s^2)+C\lambda^{-1}\|\lambda\dot\rho\|_s^2,
\end{aligned}
\tag{3.61}
$$

$$
\begin{aligned}
&\left|\left(\frac{1}{\mu}\nabla^{s-2}\nabla_j^\perp[\nabla\dot u\rho\dot F+\nabla\dot u\rho\overline{F}-\nabla\cdot\dot u\rho\dot F]_{ij},\rho\nabla^{s-2}\Delta w_i\right)\right| \\
&\quad\le C\varepsilon\|\nabla\dot u\|_{s-2}^2+\frac{\mu}{2}\|\nabla^{s-2}\nabla\Delta w\|^2+\frac{2}{\mu^3}\|\nabla^{s-2}\nabla\dot u\|_{s-2}^2.
\end{aligned}
\tag{3.62}
$$

Thus, combining (3.57) through (3.62), we have

$$
\begin{aligned}
|(\nabla^{s-2}h,\rho\nabla^{s-2}\Delta w)| &\le C\lambda^{-1}\|\lambda\dot\rho\|_s^2+C\varepsilon(\|\nabla\dot u\|_s^2+\|\nabla\Delta w\|_{s-2}^2) \\
&\quad+\frac{\mu}{2}\|\nabla^{s-2}\nabla\Delta w\|^2+\frac{2}{\mu^3}\|\nabla^{s-2}\nabla\dot u\|^2.
\end{aligned}
\tag{3.63}
$$

provided λ is appropriately large.

Finally we estimate the last remaining term $\lambda^2(\nabla^{s-2}\Delta p, \nabla^{s-2}\Delta\nabla\cdot w)$ in (3.52). Noting (3.2) and (3.42), we can use (3.28) and (3.49) to get

$$\begin{aligned}\lambda^2(\nabla^{s-2}\Delta p, \nabla^{s-2}\Delta\nabla\cdot w) &= \frac{1}{\mu}(\nabla^{s-2}\nabla\cdot(\rho\partial_t\dot{u}), \nabla^{s-2}\Delta\rho)\\ &\quad + (\nabla^{s-1}\nabla\cdot(\rho\partial_t\dot{u}), \nabla^{s-3}\Delta\nabla\cdot\dot{u})\\ &\quad + \Big(2\mu\nabla^{s-2}\Delta\nabla\cdot\dot{u} - \nabla^{s-2}\nabla\cdot(\rho\dot{u}\cdot\nabla\dot{u}) + 2\nabla^{s-2}\Delta\rho\\ &\qquad + \nabla^{s-2}\nabla\cdot(\nabla\cdot\rho\dot{F}\dot{F}^T), \nabla^{s-2}\Delta\nabla\cdot\dot{u} + \frac{1}{\mu}\nabla^{s-2}\Delta\rho\Big)\\ &\le C\mu\|\nabla\dot{u}\|_s^2 + C\varepsilon\|\nabla\Delta w\|_{s-2}^2\\ &\quad + C\lambda^{-1}(\|\lambda\dot{\rho}\|_s^2 + \|\nabla\cdot\partial_t\dot{u}\|_{s-1}^2) + C\lambda^{-1}.\end{aligned} \tag{3.64}$$

Combining (3.51) through (3.56), (3.63), and (3.64), we obtain

$$\begin{aligned}&\frac{d}{dt}(\|\rho\Delta w\|^2 + \|\rho\nabla^{s-2}\Delta w\|^2) + \mu\|\nabla\Delta w\|_{s-2}^2\\ &\quad\le C\lambda^{-1}(E_s(U(t) + \|\nabla\cdot\partial_t\dot{u}\|_{s-1}^2 + \|\Delta w\|_s^2)) + C\left(\mu + \frac{1}{\mu^3}\right)\|\nabla\dot{u}\|_s^2 + C\lambda^{-1}.\end{aligned} \tag{3.65}$$

Multiplying inequality (3.35) by an appropriately large constant $M(\frac{1}{\mu^3}+\mu)$ and then adding up the resulting inequality with (3.65), we obtain

$$\begin{aligned}&\frac{d}{dt}[\widetilde{E}(U(t)) + \widetilde{E}(\nabla^s U(t)) + \|\rho\Delta w\|^2 + \|\rho\nabla^{s-2}\Delta w\|^2] + (\|\nabla\dot{u}\|_s^2 + \|\nabla\Delta w\|_{s-2}^2)\\ &\quad\le C\lambda^{-1}(1 + \|\lambda^2\nabla\rho\|_{s-1}^2 + \|\nabla\cdot\partial_t\dot{u}\|_{s-1}^2) + C\lambda^{-1}[E_s(U(t)) + \|\Delta w\|_{s-2}^2].\end{aligned} \tag{3.66}$$

Noting (3.2), one can use Gronwall's inequality to obtain

$$\widetilde{E}(U(t)) + \widetilde{E}(\nabla^s U(t)) + \int_0^t \|\nabla\dot{u}\|_s^2\,dt \le 2[\widetilde{E}(U(0)) + \widetilde{E}(\nabla^s U(0))] + \lambda^{-1}\widetilde{E}(U(t)) \tag{3.67}$$

for $0 \le t \le T_\lambda$, with $T_\lambda = \lambda^{1-\delta}$ ($\delta < 1$ is a small positive constant), provided λ is sufficiently large. Noting (2.2) and

$$\widetilde{E}(U(0)) + \widetilde{E}(\nabla^s U(0)) \le 2(\|u_0\|_s^2 + \|F_0 - \overline{F}\|_s^2) + \lambda^{-2},$$

we have

$$E(U(t)) + E(\nabla^s U(t)) + \int_0^t \|\nabla\dot{u}\|_s^2\,dt \le C(\|u_0\|_s^2 + \|F_0 - \overline{F}\|_s^2) + \lambda^{-2} \le C\varepsilon^2, \tag{3.68}$$

provided λ is sufficiently large and $0 \le t \le T_\lambda$, which gives the proof of the first inequality of (2.2).

Step 4. Estimate for $E_{s-1}(\partial_t U(t))$.

Let $D^\alpha = \partial_t \nabla^{\alpha-1}$, $|\alpha| = s$ in (3.7). In this case we do not need the decay. We proceed as before. A similar process for (3.17)–(3.19) yields

$$|I_1| + |I_2| + |I_3| \le C\lambda^{-1} E(\partial_t \nabla^{s-1} U(t)). \tag{3.69}$$

To estimate I_4, we now return to (3.11) and write

$$I_4 = \int_{\mathsf{T}^2} [-\partial_t \nabla^{s-1} \partial_j \dot{u}_i \partial_t \nabla^{s-1} (\rho \dot{F}_{ik} \dot{F}_{jk}) + \rho \partial_t \nabla^{s-1} \dot{F}_{ik} \partial_t \partial_j \nabla^{s-1} u_i \dot{F}_{jk}] \, dx.$$

Thus,

$$|I_4| \le C\varepsilon E_{s-1}(\partial_t U(t)) + \frac{\mu}{8} \|\partial_t \nabla^{s-1} \nabla \dot{u}\|^2. \tag{3.70}$$

For the term I_5, one can use Lemmas 3.2 and 3.3 to estimate it as follows:

$$\begin{aligned} |I_5| \le\ & C\lambda^2 [\|\partial_t \nabla^{s-1} \dot{\rho}\| (\|\nabla \dot{u}\|_\infty \|\partial_t \nabla \dot{\rho}\|_{s-2} + \|\partial_t \nabla \dot{\rho}\|_\infty \|\nabla \dot{u}\|_{s-2} \\ & + \|\partial_t \nabla \dot{u}\|_\infty \|\nabla \dot{\rho}\|_{s-2} + \|\partial_t \nabla \dot{u}\|_{s-2} \|\nabla \dot{\rho}\|_\infty) \\ & + \|\partial_t \nabla^{s-1} \dot{u}\| (\|\nabla \dot{\rho}\|_\infty \|\partial_t \nabla \dot{\rho}\|_{s-2} + \|\partial_t \nabla \dot{\rho}\|_\infty \|\nabla \dot{\rho}\|_{s-2})] \\ \le\ & C\varepsilon (\|\lambda \partial_t \nabla^{s-1} \dot{\rho}\|^2 + \|\partial_t \nabla^{s-1} \dot{u}\|^2 + \|\lambda \partial_t \dot{\rho}\|^2 + \|\partial_t \dot{u}\|^2). \end{aligned} \tag{3.71}$$

Similarly, for the terms I_6, I_7, I_8, one can estimate them as follows:

$$|I_6| \le C\varepsilon (\|\partial_t \nabla^{s-1} \dot{u}\|^2 + \|\partial_t \nabla^{s-1} \dot{F}\|^2), \tag{3.72}$$

$$|I_7| \le C\varepsilon \lambda^{-1} (\|\nabla^{s-1} \partial_t \dot{u}\|^2 + \|\nabla^s \partial_t \dot{u}\|^2 + \|\lambda \partial_t \nabla^{s-1} \dot{\rho}\|^2), \tag{3.73}$$

$$|I_8| \le C\varepsilon (\|\partial_t \nabla^{s-1} \dot{u}\|^2 + \|\partial_t \nabla^{s-1} \dot{F}\|^2 + \|\lambda \partial_t \nabla^{s-1} \dot{\rho}\|^2), \tag{3.74}$$

$$|I_9| \le C\varepsilon (\|\partial_t \nabla^{s-1} \dot{F}\|^2 + \|\lambda \partial_t \nabla^{s-1} \dot{\rho}\|^2) + \frac{\mu}{8} \|\partial_t \nabla^s \dot{u}\|^2. \tag{3.75}$$

On the other hand, when $D^\alpha = \partial_t$, by using (3.3) one obtains

$$\begin{aligned} |I_9| &= \left| \int_{\mathsf{T}^2} \partial_{tj} (\rho \overline{F}_{ik} \dot{F}_{jk}) \partial_t \dot{u}^i - \partial_{jt} \dot{u}^i \partial_t (\rho \dot{F}_{ik} \overline{F}_{jk}) + \rho \partial_{jt} \dot{u}^i \overline{F}_{jk} \partial_t \dot{F}_{ik} \ dx \right| \\ &= \left| \int_{\mathsf{T}^2} -\partial_{tj} \rho \overline{F}_{jk} \overline{F}_{ik} \partial_t \dot{u}^i - \partial_{jt} \dot{u}^i \partial_t \rho \dot{F}_{ik} \overline{F}_{jk} \ dx \right| \\ &\le C\lambda^{-1} (\|\lambda \partial_t \dot{\rho}\|^2 + \|\lambda \nabla \partial_t \dot{\rho}\|^2 + \|\partial_t \dot{u}\|^2). \end{aligned}$$

Thus, we have

$$\frac{1}{2} \frac{d}{dt} \widetilde{E}(\partial_t U(t)) + \mu \|\nabla \partial_t \dot{u}\|^2 \le C\varepsilon E(\partial_t U(t)) + \frac{\mu}{2} \|\nabla \partial_t \dot{u}\|^2, \tag{3.76}$$

provided λ is sufficiently large.

By summing up (3.69)–(3.76), we arrive at

$$\frac{d}{dt} [\widetilde{E}(\partial_t U(t)) + \widetilde{E}(\partial_t \nabla^{s-1} U(t))] + \mu \|\nabla \partial_t \dot{u}\|_{s-1}^2 \le C\varepsilon E_{s-1}(\partial_t U(t)). \tag{3.77}$$

Noting (2.4), (2.5), and (2.11), we have

$$\begin{aligned} E_{s-1}(\partial_t U(0)) &\le C\|\lambda(u_0+\widetilde{u}_0^\lambda)\nabla\widetilde{\rho}_0^\lambda\|_{s-1}^2 + \|\lambda(\widetilde{\rho}_0^\lambda+1)\nabla\cdot\widetilde{u}_0^\lambda\|_{s-1}^2 + \|\lambda^2\nabla\widetilde{\rho}_0^\lambda\|_{s-1}^2 \\ &\quad + \|\nabla(\nabla\cdot\widetilde{u}_0^\lambda) + \nabla\cdot[(1+\widetilde{\rho}_0^\lambda)(F_0+\widetilde{F}_0^\lambda)(F_0+\widetilde{F}_0^\lambda)^T]\|_{s-1}^2 \\ &\quad + \|(\Delta(u_0+\widetilde{u}_0^\lambda))\|_{s-1}^2 + \|(u_0+\widetilde{u}_0^\lambda)\cdot\nabla(u_0+\widetilde{u}_0^\lambda)\|_{s-1}^2 \\ &\quad + \|(u_0+\widetilde{u}_0^\lambda)\cdot\nabla(F_0+\widetilde{F}_0^\lambda) - \nabla(u_0+\widetilde{u}_0^\lambda)(F_0+\widetilde{F}_0^\lambda)\|_{s-1} \\ &\le C(\delta_0^2 + \|u_0\|_{s+1}^2 + \|F_0-\overline{F}\|_s^2). \end{aligned} \tag{3.78}$$

Combining (3.77) and (3.78), one can use Gronwall's inequality to get

$$\widetilde{E}(\partial_t U(t)) + \widetilde{E}(\partial_t\nabla^{s-1}U(t)) + \mu\int_0^t \|\nabla\partial_t\dot{u}\|_{s-1}^2\,dt \le C\exp C\varepsilon t \tag{3.79}$$

for $0 \le t \le T_\lambda$.

This completes the proof of Theorem 2.2.

4. Proof of Theorem 2.3. In this section, we will tie everything together and complete the proof Theorem 2.3. For every $\overline{T} > 0$, we have from section 3 that

$$\begin{aligned} &E_s(U^\lambda(t)) \le C\varepsilon^2, \qquad t\in[0,T^\lambda), \\ &E_{s-1}(\partial_t U^\lambda(t)) \le C\exp Ct, \qquad 0\le t\le \overline{T}, \end{aligned} \tag{4.1}$$

provided $\lambda > C\overline{T}\exp C\overline{T}$ is sufficiently large.

We have, as $\lambda \longrightarrow \infty$, that $\rho^\lambda \longrightarrow 1$ in $L^\infty(0,\overline{T};H^s)\cap Lip([0,\overline{T}],H^{s-1})$. Moreover, a standard compactness argument based on the Lions–Aubin lemma (see [23], for example) implies that any subsequence of $(u^\lambda,\ F^\lambda)$ has a subsequence with a limit $(u,\ F)$ with $(u, F-\overline{F}) \in L^\infty(0,\overline{T};H^s)\cap C([0,\overline{T}],H^{s-\epsilon})$ and $(u_t,F_t)\in L^\infty(0,\overline{T};H^{s-1})$, where ϵ is a small positive constant. Now let $\phi(t,x)$ and $\varphi(t,x)$ be two smooth test functions with compact supports in $[0,\overline{T}]$ and $\nabla\cdot\phi = 0$. Then

$$\begin{aligned} &\int_0^{\overline{T}}\int_{\mathsf{T}^2}\phi\left(\partial_t\dot{u}^\lambda + \dot{u}^\lambda\cdot\nabla\dot{u}^\lambda - \frac{\mu}{\rho^\lambda}\Delta\dot{u}^\lambda - \frac{\mu}{\rho^\lambda}\nabla\nabla\cdot\dot{u}^\lambda - \frac{1}{\rho^\lambda}\nabla\cdot\rho^\lambda\dot{F}^\lambda(\dot{F}^\lambda)^T \right.\\ &\qquad\qquad \left. - \frac{1}{\rho^\lambda}\nabla\cdot[\rho^\lambda\dot{F}^\lambda\overline{F}^T + \rho^\lambda\overline{F}(\dot{F}^\lambda)^T]\right)\,dxdt \\ &= \int_0^{\overline{T}}\int_{\mathsf{T}^2} -\phi\lambda^2\nabla\int_1^{\rho^\lambda}\frac{p'(\xi)}{\xi}\,d\xi\,dxdt \\ &= 0 \end{aligned} \tag{4.2}$$

and

$$\int_0^{\overline{T}}\int_{\mathsf{T}^2}\varphi(\partial_t\dot{F}^\lambda + \dot{u}^\lambda\cdot\nabla\dot{F}^\lambda - \nabla\dot{u}^\lambda\dot{F}^\lambda - \nabla\dot{u}^\lambda\overline{F})\,dxdt = 0.$$

On the other hand, from the first equation of system (1.1), we have

$$\nabla\cdot\dot{u}^\lambda = \partial_t\dot{\rho}^\lambda + \dot{u}^\lambda\cdot\nabla\dot{\rho}^\lambda + \dot{\rho}^\lambda\nabla\cdot\dot{u}^\lambda.$$

Let $\lambda \longrightarrow \infty$. We obtain that (u,F) satisfies (2.9) and (2.10) in time interval $t\in[0,\overline{T}]$. By the uniqueness of the classical solution of system (2.9), (2.10), it follows that the convergence is in fact valid for the sequences u^λ and F^λ themselves. Since $\overline{T}$ is arbitrary, (u,F) in fact is the unique global classical solution of the incompressible Oldroyd system (1.2) with the initial data (2.10).

Acknowledgment. The first author would like to thank Professor Chun Liu for many helpful discussions and suggestions.

REFERENCES

[1] S. Alinhac, *Blowup for Nonlinear Hyperbolic Equations*, Birkhäuser, Boston, 1995.

[2] J.-Y. Chemin and N. Masmoudi, *About lifespan of regular solutions of equations related to viscoelastic fluids*, SIAM J. Math. Anal., 33 (2001), pp. 84–112.

[3] D. Ebin, *Global solutions of the equations of elastodynamics of incompressible neo-Hookean materials*, Proc. Natl. Acad. Sci. USA, 90 (1993), pp. 3802–3805.

[4] D. Ebin, *Global solutions of the equations of elastodynamics of incompressible materials*, Electron. Res. Announc. Amer. Math. Soc., 2 (1996), pp. 50–59.

[5] J. Hron, J. Málek, J. Nečas, and K. R. Rajagopal, *Numerical simulations and global existence of solutions of two-dimensional flows of fluids with pressure- and shear-dependent viscosities*, Math. Comput. Simulation, 61 (2003), pp. 297–315.

[6] D. D. Joseph, *Instability of the rest state of fluids of arbitrary grade greater than one*, Arch. Ration. Mech. Anal., 75 (1980/81), pp. 251–256.

[7] S. Kawashima and Y. Shibata, *Global existence and exponential stability of small solutions to nonlinear viscoelasticity*, Comm. Math. Phys., 148 (1992), pp. 189–208.

[8] S. Klainerman, *Uniform decay estimates and the Lorentz invariance of the classical wave equation*, Comm. Pure Appl. Math., 38 (1985), pp. 321–332.

[9] S. Klainerman, *The null condition and global existence to nonlinear wave equations*, in Nonlinear Systems of Partial Differential Equations in Applied Mathematics, Part 1, (Santa Fe, NM, 1984), Lectures in Appl. Math. 23, AMS, Providence, RI, 1986, pp. 293–326.

[10] S. Klainerman and A. Majda, *Singular limits of quasilinear hyperbolic system with large parameters and the incompressible limit of compressible fluids*, Comm. Pure Appl. Math., 34 (1981), pp. 481–524.

[11] S. Klainerman and T. C. Sideris, *On almost global existence for nonrelativistic wave equations in* 3*D*, Comm. Pure Appl. Math., 49 (1996), pp. 307–321.

[12] O. A. Ladyzhenskaya and G. A. Seregin, *On the regularity of solutions of two-dimensional equations of the dynamics of fluids with nonlinear viscosity*, Zap. Nauchn. Sem. S.-Peterburg Otdek Mat. Inst. Steklon (POMI), 259 (1999), pp. 145–166.

[13] Z. Lei, *Global existence of classical solutions for some Oldroyd-B model via the incompressible limit*, Chinese Ann. Math. Ser. B, to appear.

[14] F. H. Lin and C. Liu, *Existence of solutions for Erichsen-Leslie system*, Arch. Ration. Mech. Anal., 154 (2000), pp. 135–156.

[15] F. H. Lin, C. Liu, and P. Zhang, *On hydrohynamics of viscoelastic fluids*, Comm. Pure Appl. Math., to appear.

[16] P. L. Lions and N. Masmoudi, *Global solutions for some Oldroyd models of non-Newtonian flows*, Chinese Ann. Math. Ser. B, 21 (2000), pp. 131–146.

[17] C. Liu and N. J. Walkington, *An Eulerian description of fluids containing visco-hyperelastic particles*, Arch. Ration. Mech. Anal., 159 (2001), pp. 229–252.

[18] J. Málek, J. Nečas, and K. R. Rajagopal, *Global analysis of solutions of the flows of fluids with pressure-dependent viscosities*, Arch. Ration. Mech. Anal., 165 (2002), pp. 243–269.

[19] W. R. Schowalter, *Mechanics of Non-Newtonian Fluids*, Pergamon Press, Oxford, 1978.

[20] T. C. Sideris, *Nonresonance and global existence of prestressed nonlinear elastic waves*, Ann. of Math. (2), 151 (2000), pp. 849–874.

[21] T. C. Sideris and B. Thomases, *Global existence for* 3*D incompressible isotropic elastodynamics via the incompressible limit*, Comm. Pure Appl. Math., 57 (2004), pp. 1–39.

[22] M. Slemrod, *Constitutive relations for Rivlin-Erichsen fluids based on generalized rational approximation*, Arch. Ration. Mech. Anal., 146 (1999), pp. 73–93.

[23] R. Teman, *Navier-Stokes Equations*, North-Holland, Amsterdam, 1977.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 815–840

OSCILLATING SOLUTIONS OF INCOMPRESSIBLE MAGNETOHYDRODYNAMICS AND DYNAMO EFFECT*

DAVID GÉRARD-VARET†

Abstract. We are interested in the stability properties of some solutions of viscous incompressible magnetohydrodynamics equations. These solutions are highly oscillating, with frequency involving a small parameter ε. They arise in the study of small-scale dynamo mechanisms. We prove both nonlinear stability and instability results, depending on the time scale under consideration.

Key words. magnetohydrodynamics, dynamo theory, oscillations

AMS subject classifications. 35Q30, 35Q35

DOI. 10.1137/S0036141004444603

1. Introduction. This paper deals with some oscillatory solutions of the equations of magnetohydrodynamics (MHD). It is motivated by the study of small-scale dynamo mechanisms. Before we state precisely our main results, let us first specify the general framework.

The incompressible MHD equations read in a dimensionless form:

$$\begin{cases} \partial_t u + u \cdot \nabla u + \nabla p - \dfrac{1}{\mathrm{Re}} \Delta u = \operatorname{curl} b \times b + f, \\ \partial_t b - \operatorname{curl}(u \times b) - \dfrac{1}{\mathrm{Rm}} \Delta b = 0, \\ \operatorname{div}\, u = \operatorname{div}\, b = 0. \end{cases} \tag{1.1}$$

They describe the evolution of an incompressible and electrically conducting fluid. They are derived from the incompressible Navier–Stokes equations, the Maxwell equations, and Ohm's law in a conducting medium (see [15]). Functions

$$u = u(t, \mathrm{x}) \in \mathbb{R}^3, \quad b = b(t, \mathrm{x}) \in \mathbb{R}^3, \quad f = f(t, \mathrm{x}) \in \mathbb{R}^3$$

model, respectively, the fluid velocity, the magnetic field, and an additional forcing term; for instance, due to convection. The space and time variables are $t \in \mathbb{R}^+$, $\mathrm{x} = (x, y, z) \in \mathbb{R}^3$. We denote

$$\Delta = \partial_x^2 + \partial_y^2 + \partial_z^2, \quad \nabla = (\partial_x, \partial_y, \partial_z)^t,$$

and for any $v = (v_1(\mathrm{x}), v_2(\mathrm{x}), v_3(\mathrm{x}))^t \in \mathbb{R}^3$,

$$\operatorname{div}\, v = \partial_x v_1 + \partial_y v_2 + \partial_z v_3, \quad \operatorname{curl} v = (\partial_y v_3 - \partial_z v_2, \partial_z v_1 - \partial_x v_3, \partial_x v_2 - \partial_y v_1)^t.$$

Constants Re and Rm are called hydrodynamic and magnetic Reynolds numbers. To lighten the notation, we will assume in what follows that Re = Rm = 1. Note that the divergence-free condition on b is preserved by (1.1b). As soon as it is satisfied initially, it is satisfied for all positive times.

*Received by the editors June 21, 2004; accepted for publication (in revised form) January 15, 2005; published electronically December 2, 2005.
http://www.siam.org/journals/sima/37-3/44460.html

†C.N.R.S./D.M.A, U.M.R. 8553, E.N.S., 45 Rue d'Ulm, 75230 Paris Cedex 05, France (David.Gerard-Varet@ens.fr).

Roughly speaking, *dynamo theory* deals with the stability of solutions

$$(u, b) = (u(t, \mathrm{x}), 0)$$

of system (1.1). More precisely, it studies the generation of magnetic field from the fluid flow u. The basic idea is that the "self-excited" term $\mathrm{curl}\,(u \times b)$ may amplify the magnetic field in the manner of an instability. As long as the fluid motion is strong enough, this transfer from kinetic to magnetic energy may thus prevent the decay of the magnetic field, despite the dissipation term $-\Delta b$.

It is widely accepted that dynamo action takes place in the Earth, in the sun, and in many other planets and stars. Therefore, the understanding of dynamo mechanisms is a major physical issue. It has been the subject of an expansive literature: we refer to the recent review papers by Gilbert [8] and Fearn [5] for a good introduction and appropriate lists of references. Note that most of these references are limited to *kinematic dynamos*: the Laplace force is neglected, and only the induction equation (1.1b) is considered, at imposed velocity u.

Among the mechanisms that have been identified, one of the most famous is the so-called *alpha effect*. It is based on a *scale separation*: the velocity and magnetic fields are assumed to vary on (turbulent) time and length scales τ and l, much smaller than the typical macro scales T and L. Introducing the ratios $\lambda = \tau/T \ll 1$ and $\beta = l/L \ll 1$, one can write this with little formalism:

$$\begin{aligned} u &\approx u_* \left(t, \mathrm{x}, \lambda^{-1} t,\, \beta^{-1}\mathrm{x}\right) \;+\; \bar{u}\,(t, \mathrm{x})\,, \\ b &\approx b_* \left(t, \mathrm{x}, \lambda^{-1} t,\, \beta^{-1}\mathrm{x}\right) \;+\; \bar{b}\,(t, \mathrm{x})\,, \end{aligned} \tag{1.2}$$

where u_* (resp., b_*) is the fluctuating part of the field, and $\bar{u}$ (resp., $\bar{b}$) is its mean part. The basic idea is that the "average" of the fluctuating term $\mathrm{curl}\,(u_* \times b_*)$ can have a destabilizing effect on the mean field $\bar{b}$, generating a dynamo.

This idea was first introduced by Parker [13] in 1955, and in a geophysical context by Braginsky [2]. It has been generalized by Steenbeck, Krause, and Rädler [16]. Let us also mention the important works [14] and [3] on periodic dynamos. Note that the alpha effect has since been confirmed experimentally [17].

The present paper is a small step towards the mathematical study of this mechanism. Namely, we will investigate the stability properties of solutions $(u, 0)$ of (1.1) given by

$$(u, 0) = (\varepsilon^{-1} u^\varepsilon, 0), \quad u^\varepsilon(t, \mathrm{x}) = U(\varepsilon^{-4} t, \varepsilon^{-2}\mathrm{x}), \tag{1.3}$$

where $U = U(\tau, \theta)$ satisfies

$$U \in H^\infty \left(\mathbb{T} \times \mathbb{T}^3\right)^3, \quad \int_{\mathbb{T}\times\mathbb{T}^3} U = 0, \quad \mathrm{div}_\theta\, U = 0. \tag{1.4}$$

Remark 1. The set $\mathcal{P}$ of profiles U satisfying (1.4) is a Fréchet space, where the topology is induced by the family of norms

$$\|U\|_m^2 = \sum_{(\omega,\xi)\in\mathbb{Z}^4} \left(|\omega|^2 + |\xi|^2\right)^m |\hat{U}(\omega, \xi)|^2, \quad m \geq 0,$$

where $\hat{U}$ is the Fourier transform with respect to (τ, θ). We denote by $d_{\mathcal{P}}$ a metric defining this topology.

Remark 2. We assume that $\varepsilon \ll 1$ and that $\int U = 0$, which means we consider fast oscillations with zero mean flow. This is reminiscent of the (somehow crude) modeling of turbulence that we have in mind.

Remark 3. The amplitude and time and length scales in (1.3) are classical (see [8]). In short, one can say using the notation of (1.2) that they correspond to the case

$$|u_*| \gg |\bar{u}|, \quad |b_*| \ll |\bar{b}|.$$

Indeed, the oscillatory part $u_* = \varepsilon^{-1}u^\varepsilon$ of the velocity field is $O(\varepsilon^{-1})$, bigger than the $O(1)$ potential mean part $\bar{u}$. On the contrary, our choices $O(\varepsilon^4)$ and $O(\varepsilon^2)$ for time and length scales of u^ε ensure small amplitude for the oscillatory part of b. This can be seen formally from (1.1b), as the oscillation $b_* = b_*(t, \mathrm{x}, \tau = \varepsilon^{-4}t, \theta = \varepsilon^{-2}\mathrm{x})$ satisfies

$$\varepsilon^{-4}\left(\partial_\tau b_* - \Delta_\theta b_*\right) = o(1).$$

We emphasize that other choices for solutions u^ε are possible. In particular, it would be interesting to study oscillating fields with larger wavelength, so as to emphasize the role of the hyperbolic part of (1.1).

We will show that solutions u^ε given by (1.3) *are stable on times $t = O(1)$ and "generically" unstable on times $t \sim |\ln(\varepsilon)|$.* Substituting $u = \varepsilon^{-1}u^\varepsilon + v$ into (1.1), we will rather work with the system

$$\begin{cases} \partial_t v + \varepsilon^{-1}u^\varepsilon \cdot \nabla v + \varepsilon^{-1} v \cdot \nabla u^\varepsilon + v \cdot \nabla v + \nabla p - \Delta v = \operatorname{curl} b \times b, \\ \partial_t b - \varepsilon^{-1}\operatorname{curl}(u^\varepsilon \times b) - \operatorname{curl}(v \times b) - \Delta b = 0, \\ \operatorname{div} v = \operatorname{div} b = 0. \end{cases} \tag{1.5}$$

Note that for all $\varepsilon > 0$ and all divergence-free fields $v_0^\varepsilon, b_0^\varepsilon \in L^2\left(\mathbb{R}^3\right)^3$, system (1.5) has global weak solutions

$$v, b \in L^\infty_{loc}\left(\mathbb{R}^+;\, L^2\left(\mathbb{R}^3\right)\right)^3 \cap L^2_{loc}(\mathbb{R}^+;\, \dot{H}^1(\mathbb{R}^3))^3,$$

with initial data $v_0^\varepsilon, b_0^\varepsilon$. Indeed, u^ε and its derivatives are bounded functions, so that a classical Leray-type existence theorem for MHD equations extends easily to (1.5).

We will first prove the following stability result.

THEOREM 1.1 (nonlinear stability result). *Let $U \in \mathcal{P}$, and let $\{u^\varepsilon\}_{\varepsilon>0}$ satisfy* (1.3). *Let $m \in \mathbb{N}$, and let v_0, b_0 be in $H^\infty\left(\mathbb{R}^3\right)^3$ divergence-free.*

For all $T \geq 0$, there exist $\delta > 0$, $\varepsilon_0 > 0$ such that if

$$m \geq 1 \quad \text{or} \quad \|(v_0, b_0)\|_{H^{1/2}} \leq \delta,$$

the Cauchy problem (1.5) *with initial data $\varepsilon^m v_0, \varepsilon^m b_0$ has a unique solution*

$$v^\varepsilon, b^\varepsilon \in C^0\left([0,T];\, H^\infty\left(\mathbb{R}^3\right)\right)^3$$

for all $\varepsilon < \varepsilon_0$. Moreover, it satisfies the following for a positive constant C, large s, and small enough ε:

$$\sup_{0\leq t\leq T} \|(v^\varepsilon, b^\varepsilon)(t,\cdot)\|_{L^2} \;\leq\; C\,\varepsilon^m\,\|(v_0,b_0)\|_{H^s},$$
$$\sup_{0\leq t\leq T} \|(v^\varepsilon, b^\varepsilon)(t,\cdot)\|_{L^\infty} \;\leq\; C\,\varepsilon^m\,\|(v_0,b_0)\|_{H^s}.$$

We will then prove the following instability result.

THEOREM 1.2 (nonlinear instability result). *There exists a dense and open subset Ω of $\mathcal{P}$ such that*

- *for all* $U \in \Omega$, *and* $\{u^\varepsilon\}_{\varepsilon>0}$ *satisfying* (1.3), *and*
- *for all* $m \in \mathbb{N}$,

one can find $\delta > 0$, *times* $t(\varepsilon) = O(|\ln(\varepsilon)|)$, *and families of solutions* $\{(v^\varepsilon, b^\varepsilon)^t\}_{\varepsilon>0}$ *of* (1.5) *with*

$$v^\varepsilon, b^\varepsilon \in C^0\left(\mathbb{R}^+; H^\infty\left(\mathbb{R}^3\right)\right)^3,$$

$$\| \partial_{\mathrm{x}}^\alpha (v^\varepsilon, b^\varepsilon)|_{t=0} \|_{L^2} \le C_\alpha\, \varepsilon^{m-2|\alpha|+1} \quad \forall \alpha \in \mathbb{N}^3,$$

and

$$\|b^\varepsilon|_{t=t(\varepsilon)}\|_{L^2} \ge \delta.$$

Remark 4. Note that the lower bound in Theorem 1.2 applies to b^ε, which is exactly the mathematical expression of an alpha effect: small-scale velocity u^ε generates destabilization of $b = 0$.

Remark 5. Theorem (1.2) extends and justifies linear computations carried in [8]. It is also reminiscent of the classical linear computations of Roberts [14] on periodic dynamos. In [14] Roberts studied the equation

$$\partial_t b + \mathrm{curl}\,(u^\varepsilon \times b) - \nu\varepsilon\Delta b = 0, \quad u^\varepsilon = U\left(\varepsilon^{-1}t, \varepsilon^{-1}\mathrm{x}\right),$$

for U periodic with zero mean. He showed an instability result for ν large enough. However, his analysis, which relied heavily on perturbation theory, does not adapt to our nonlinear framework. Henceforth, we use a drastically different approach, based on energy estimates.

The paper is structured as follows. In section 2, we introduce an auxiliary singular system. This system involves additional variables, which take into account the dependence of $(v^\varepsilon, b^\varepsilon)$ on $\varepsilon^{-4}t$, $\varepsilon^{-2}\mathrm{x}$. We construct approximate solutions of this system (section 2.1) and perform a priori estimates (section 2.2). The proof of Theorem 1.1 follows (section 2.3). In section 3, we focus on the instability mechanism. We show that the approximate solutions of section 2 have generically exponential growth (section 3.1) and give precise estimates on this growth (section 3.2). We end with the proof of Theorem 1.2 (section 3.3).

2. Singular system. From the structure of the small-scale flow u^ε, we expect solutions $v^\varepsilon, b^\varepsilon$ of (1.5) to exhibit rapid oscillations, involving $\varepsilon^{-4}t$ and $\varepsilon^{-2}\mathrm{x}$. In particular, we expect the derivatives of $v^\varepsilon, b^\varepsilon$ to behave badly. Hence, we do not expect good H^s energy estimates on system (1.5). To override this difficulty, we will follow ideas of nonlinear geometric optics (cf. [10]): we will work directly in the class of solutions of the type

$$(v^\varepsilon, b^\varepsilon)^t(t,\mathrm{x}) = V^\varepsilon\left(t, \mathrm{x}, \varepsilon^{-4}t, \varepsilon^{-2}\mathrm{x}\right), \tag{2.1}$$

where $V^\varepsilon = V^\varepsilon(t,\mathrm{x},\tau,\theta)$ is periodic in τ and θ. We will get Sobolev bounds on V^ε, which will allow us to control v^ε, b^ε in L^2 and L^∞ (cf. Theorem 1.2).

We introduce the following singular system, of unknown $V = (w, \beta)^t$:

$$\begin{cases} \partial_t V + \varepsilon^{-4}\partial_\tau V + \varepsilon^{-1}(B_{\mathrm{x}} + \varepsilon^{-2}B_\theta)(\tilde{U}, V) + \dfrac{1}{2}(B_{\mathrm{x}} + \varepsilon^{-2}B_\theta)(V,V) \\[2mm] \qquad - \left(\nabla_{\mathrm{x}} + \varepsilon^{-2}\nabla_\theta\right)^2 V = \left(\left(\nabla_{\mathrm{x}} + \varepsilon^{-2}\nabla_\theta\right) p,\, 0\right)^t, \\[2mm] \mathrm{Div}_{\mathrm{x}} V + \varepsilon^{-2}\mathrm{Div}_\theta V = 0, \end{cases} \tag{2.2}$$

with $\tilde{U} := (U,0)^t$, $U = U(\tau,\theta) \in \mathcal{P}$, where for all $V = (w,\beta)$, $\tilde{V} = (\tilde{w},\tilde{\beta})$,

$$B_{\mathrm{x}}(V,\tilde{V}) := \begin{pmatrix} \mathrm{div}_{\mathrm{x}}\left(w \otimes \tilde{w} - \beta \otimes \tilde{\beta}\right) + \mathrm{div}_{\mathrm{x}}\left(\tilde{w} \otimes w - \tilde{\beta} \otimes \beta\right) \\ -\mathrm{curl}_{\mathrm{x}}\left(w \times \tilde{\beta}\right) - \mathrm{curl}_{\mathrm{x}}\left(\tilde{w} \times \beta\right) \end{pmatrix},$$

$$B_{\theta}(V,\tilde{V}) := \begin{pmatrix} \mathrm{div}_{\theta}\left(w \otimes \tilde{w} - \beta \otimes \tilde{\beta}\right) + \mathrm{div}_{\theta}\left(\tilde{w} \otimes w - \tilde{\beta} \otimes \beta\right) \\ -\mathrm{curl}_{\theta}\left(w \times \tilde{\beta}\right) - \mathrm{curl}_{\theta}\left(\tilde{w} \times \beta\right) \end{pmatrix},$$

and where

$$\mathrm{Div}_{\mathrm{x}} = \begin{pmatrix} \mathrm{div}_{\mathrm{x}} & \\ & \mathrm{div}_{\mathrm{x}} \end{pmatrix}, \quad \mathrm{Div}_{\theta} = \begin{pmatrix} \mathrm{div}_{\theta} & \\ & \mathrm{div}_{\theta} \end{pmatrix}.$$

Note that the quadratic terms in (1.5a) satisfy

$$u^{\varepsilon} \cdot \nabla v + v \cdot \nabla u^{\varepsilon} = \mathrm{div}\ (u^{\varepsilon} \otimes v) + \mathrm{div}\ (v \otimes u^{\varepsilon}),$$

$$\mathrm{curl}\, b \times b = b \cdot \nabla b - \frac{1}{2}\nabla|b|^2 = \mathrm{div}\ (b \otimes b) - \frac{1}{2}\nabla|b|^2,$$

using the fact that v, b are divergence-free. We thus see that any regular solution V^{ε} of (2.2) provides a solution $v^{\varepsilon}, b^{\varepsilon}$ of (1.5) through identity (2.1).

2.1. Approximate solutions. *Up to the end of the section, we fix time $T > 0$, and $m \in \mathbb{N}$.* For any

$$f = f(t,\mathrm{x},\tau,\theta) = \sum_{\omega,k} f_{\omega,k}(t,\mathrm{x})\, e^{i(\omega\tau + k\cdot\theta)},$$

we denote $\overline{f} = f_{0,0}$, $f_* = f - \overline{f}$.

We first construct approximate solutions of (2.2) of the following type:

$$\begin{aligned} V^{\varepsilon}(t,\mathrm{x},\tau,\theta) &\approx \varepsilon^m \sum \varepsilon^i V^i(t,\mathrm{x},\tau,\theta), \\ p^{\varepsilon}(t,\mathrm{x},\tau,\theta) &\approx \varepsilon^{m-1} \sum \varepsilon^i p^i(t,\mathrm{x},\tau,\theta), \end{aligned} \tag{2.3}$$

where for all $i \geq 0$,

$$\left(V^i, p^i\right)^t = (w^i, \beta^i, p^i)^t \in C^{\infty}\left([0,T];\ H^{\infty}\left(\mathbb{R}^3 \times \mathbb{T} \times \mathbb{T}^3\right)\right)^7, \quad \int \overline{p}^i(t,\cdot) = 0.$$

We plug approximation (2.3) into the system (2.2). We identify terms of order ε^{m+i-4} in (2.2a) and of order ε^{m+i-2} in (2.2b). This yields, for all $i \geq 0$,

$$\begin{cases} (\partial_\tau - \Delta_\theta)V^i = F^i, \\ \mathrm{Div}_\theta V^i = -\mathrm{Div}_{\mathrm{x}} V^{i-2}, \end{cases} \tag{S_i}$$

where $V^j := 0$, $p^j := 0$ for $j < 0$, and

$$\begin{aligned} F^i = &-(\partial_t - \Delta_{\mathrm{x}})V^{i-4} - B_{\mathrm{x}}\left(\tilde{U}, V^{i-3}\right) + \left(\mathrm{div}_{\mathrm{x}}\nabla_\theta + \mathrm{div}_\theta\nabla_{\mathrm{x}}\right)V^{i-2} \\ &- \sum_{j+J=i-m-4} B_{\mathrm{x}}\left(V^j, V^J\right) - \sum_{j+J=i-m-2} B_\theta\left(V^j, V^J\right) - B_\theta\left(\tilde{U}, V^{i-1}\right) \\ &+ \left(\nabla_\theta p^{i-1}, 0\right)^t + \left(\nabla_{\mathrm{x}} p^{i-3}, 0\right)^t. \end{aligned} \tag{2.4}$$

Note that (S_0) is equivalent to $V_*^0 \equiv 0$. For $i \geq 0$, we take the oscillatory part of (S_{i+1}) and the average of (S_{i+4}). We get

$$(T_i) \qquad \begin{cases} (\partial_\tau - \Delta_\theta) V_*^{i+1} = \left(\nabla_\theta p_*^i, 0\right)^t - B_\theta\left(\tilde{U}, \overline{V}^i\right) + G_*^i, \\ \mathrm{Div}_\theta V_*^{i+1} = -\mathrm{Div}_{\mathrm{x}} V_*^{i-1}, \\ (\partial_t - \Delta_{\mathrm{x}})\overline{V}^i = \left(\nabla_{\mathrm{x}} \overline{p}^{i+1}, 0\right)^t - \overline{B_{\mathrm{x}}\left(\tilde{U}, V_*^{i+1}\right)} - \displaystyle\sum_{j+J=i-m} \overline{B_{\mathrm{x}}\left(V^j, V^J\right)}, \\ \mathrm{Div}_{\mathrm{x}} \overline{V}^i = 0, \end{cases}$$

where $G_*^i := F_*^{i+1} - (\nabla_\theta p_*^i, 0)^t + B_\theta(\tilde{U}, \overline{V}^i)$ depends only on $V^0, \ldots, V^{i-1}$, V_*^i, and $\nabla_{\mathrm{x}} p_*^{i-1}$.

Thus, introducing

$$X^i := \left(\overline{V}^i, V_*^{i+1}, p_*^i, \overline{p}^{i+1}\right),$$

(T_i) can be seen as a system of unknown X^i, with data depending on $X^0, X^1, \ldots, X^{i-1}$. We will show inductively on $i \geq 0$ the solvability of (T_i).

Case $i = 0$. Recall that $V_*^0 \equiv 0$. As $\overline{p}^0$ does not appear in systems (T_i), we can also assume $\overline{p}^0 \equiv 0$. System (T_0) depends on the values of m.

- $m = 0$. The system (T_0) reads

$$\begin{cases} (\partial_\tau - \Delta_\theta) V_*^1 = \left(\nabla_\theta p_*^0, 0\right)^t - B_\theta\left(\tilde{U}, \overline{V}^0\right), \\ \mathrm{Div}_\theta V_*^1 = 0, \\ (\partial_t - \Delta_{\mathrm{x}})\overline{V}^0 = \left(\nabla_{\mathrm{x}} \overline{p}^1, 0\right)^t - \overline{B_{\mathrm{x}}\left(\tilde{U}, V_*^1\right)} - B_{\mathrm{x}}\left(\overline{V}^0, \overline{V}^0\right), \\ \mathrm{Div}_{\mathrm{x}} \overline{V}^0 = 0. \end{cases}$$

Applying Div_θ to $(T_0\mathrm{a})$ and using $(T_0\mathrm{b})$ leads to

$$(\Delta_\theta p_*^0, 0)^t = \mathrm{Div}_\theta B_\theta\big(\tilde{U}, \overline{V}^0\big).$$

As the second component of the right-hand side is $\mathrm{div}_\theta\, \mathrm{curl}_\theta\big(U \times \overline{\beta}^0\big) \equiv 0$, such an equation has a unique solution:

$$(p_*^0, 0)^t = \Delta_\theta^{-1} \mathrm{Div}_\theta B_\theta\big(\tilde{U}, \overline{V}^0\big). \tag{2.5}$$

Here, Δ_θ^{-1} denotes the inverse of Δ_θ in the set of L^2 periodic functions of (τ, θ) with zero average in θ. Thus, we are left with

$$\begin{cases} (\partial_\tau - \Delta_\theta) V_*^1 = L_\theta \overline{V}^0, \\ (\partial_t - \Delta_{\mathrm{x}})\overline{V}^0 = \left(\nabla_{\mathrm{x}} \overline{p}^1, 0\right)^t - \overline{B_{\mathrm{x}}\left(\tilde{U}, V_*^1\right)} - B_{\mathrm{x}}\left(\overline{V}^0, \overline{V}^0\right), \\ \mathrm{Div}_{\mathrm{x}} \overline{V}^0 = 0, \end{cases}$$

where

$$L_\theta \overline{V}^0 := \nabla_\theta \Delta_\theta^{-1} \mathrm{Div}_\theta B_\theta \left(\tilde{U}, \overline{V}^0\right) - B_\theta \left(\tilde{U}, \overline{V}^0\right).$$

Note that

$$V_*^1 = (\partial_\tau - \Delta_\theta)^{-1} L_\theta \overline{V}^0, \tag{2.6}$$

where $(\partial_\tau - \Delta_\theta)^{-1}$ denotes the inverse of $\partial_\tau - \Delta_\theta$ in the set of L^2 periodic functions of (τ, θ) with zero average in θ. We end up with

$$\begin{cases} (\partial_t - \Delta_{\mathrm{x}})\overline{V}^0 = \left(\nabla_{\mathrm{x}} \overline{p}^1, 0\right)^t - \overline{B_{\mathrm{x}}\left(\tilde{U}, (\partial_\tau - \Delta_\theta)^{-1} L_\theta \overline{V}^0\right)} - B_{\mathrm{x}}\left(\overline{V}^0, \overline{V}^0\right), \\ \mathrm{Div}_{\mathrm{x}} \overline{V}^0 = 0. \end{cases} \tag{2.7}$$

This last system is of the type (1.1), up to the additional linear term

$$\mathcal{A}\overline{V}^0 := -\overline{B_{\mathrm{x}}\left(\tilde{U}, (\partial_\tau - \Delta_\theta)^{-1} L_\theta \overline{V}^0\right)}.$$

Classical existence results for smooth solutions of Navier–Stokes-type equations extend without difficulty (see [7]). In particular, there exists $\delta = \delta(T)$ such that for all

$$\overline{V}_0^0 \in H^\infty\left(\mathbb{R}^3\right)^6, \quad \mathrm{Div}_{\mathrm{x}} \overline{V}_0^0 = 0, \quad \|\overline{V}_0^0\|_{H^{1/2}} \le \delta,$$

system (2.7) has a unique solution

$$\left(\overline{V}^0, \overline{p}^1\right) \in C^\infty\left([0,T];\ H^\infty\left(\mathbb{R}^3\right)\right)^7, \quad \int \overline{p}^1(t,\cdot) = 0,$$

with $\overline{V}^0|_{t=0} = \overline{V}_0^0$. Together with (2.5) and (2.6), this provides a unique solution

$$X^0 = \left(\overline{V}^0, V_*^1, p_*^0, \overline{p}^1\right)$$

of system (T_0).

- $m \ge 1$. The situation is even simpler, as the quadratic term disappears. We are left with

$$\begin{cases} (\partial_t - \Delta_{\mathrm{x}})\overline{V}^0 = \left(\nabla_{\mathrm{x}} \overline{p}^1, 0\right)^t + \mathcal{A}\overline{V}^0, \\ \mathrm{Div}_{\mathrm{x}} \overline{V}^0 = 0, \end{cases} \tag{2.8}$$

which has regular solutions

$$\left(\overline{V}^0, p^1\right)^t \in C^0\left(\mathbb{R}^+;\ H^\infty\left(\mathbb{R}^3\right)\right)^7, \quad \int \overline{p}^1(t,\cdot) = 0$$

for any initial data $\overline{V}_0^0$ in $H^\infty\left(\mathbb{R}^3\right)^6$, $\mathrm{Div}_{\mathrm{x}} \overline{V}^0 = 0$.

Case $i \ge 1$. The solvability of (T_i) is proved inductively. Let $i \ge 1$, and let $X^0, \dots, X^{i-1}$ solve $(T_0), \dots, (T_{i-1})$ on the time interval $[0,T]$. Applying Div_θ to $(T_i\mathrm{a})$ and using $(T_i\mathrm{b})$ leads to

$$(\Delta_\theta p_*^i, 0)^t = \mathrm{Div}_\theta\left(B_\theta\left(\tilde{U}, \overline{V}^i\right) - G_*^i\right) - (\partial_\tau - \Delta_\theta)\, \mathrm{Div}_{\mathrm{x}} V_*^{i-1}. \tag{2.9}$$

To solve (T_i), we need to verify that the right-hand side of (2.9) has zero second component. This compatibility condition ensures that the magnetic component of our approximation remains divergence-free. It is reminiscent to the fact that $\operatorname{div} b = 0$ is preserved by (1.5b).

If we denote $G^i_* = \left(g^i_*, h^i_*\right)^t$, then (2.9) has a unique solution if and only if

$$\mathrm{div}_\theta h^i_* = -(\partial_\tau - \Delta_\theta)\mathrm{div}_\mathrm{x} \beta^{i-1}_*,$$

i.e.,

$$\mathrm{div}_\theta h^i_* = -\mathrm{div}_\mathrm{x} h^{i-2}_*. \tag{2.10}$$

The expression of h^i_* yields

$$\begin{aligned}
\mathrm{div}_\theta h^i_* &= \mathrm{div}_\theta \bigg(-(\partial_t - \Delta_\mathrm{x})\beta^{i-3} + \left(\mathrm{div}_\mathrm{x}\nabla_\theta + \mathrm{div}_\theta\nabla_\mathrm{x}\right)\beta^{i-1} \\
&\qquad + \mathrm{curl}_\mathrm{x}\left(U \times \beta^{i-2}\right) + \mathrm{curl}_\mathrm{x} \sum_{j+J=i-m-3} w^j \times \beta^J \bigg) \\
&= -(\partial_t - \Delta_\mathrm{x})\mathrm{div}_\theta \beta^{i-3} + \left(\mathrm{div}_\mathrm{x}\nabla_\theta + \mathrm{div}_\theta\nabla_\mathrm{x}\right)\mathrm{div}_\theta \beta^{i-1} \\
&\qquad - \mathrm{div}_\mathrm{x}\mathrm{curl}_\theta \left(U \times \beta^{i-2}\right) - \mathrm{div}_\mathrm{x}\mathrm{curl}_\theta \sum_{j+J=i-m-3} w^j \times \beta^J.
\end{aligned}$$

Using that $\mathrm{div}_\theta \beta^i = -\mathrm{div}_\mathrm{x}\beta^{i-2}$ for all i and $\mathrm{div}_\mathrm{x}\mathrm{curl}_\mathrm{x} \equiv 0$, we get

$$\begin{aligned}
\mathrm{div}_\theta h^i_* = -\mathrm{div}_\mathrm{x} \bigg(&-(\partial_t - \Delta_\mathrm{x})\beta^{i-5} + \left(\mathrm{div}_\mathrm{x}\nabla_\theta + \mathrm{div}_\theta\nabla_\mathrm{x}\right)\beta^{i-3} \\
&+ \mathrm{curl}_\theta\left(U \times \beta^{i-2}\right) + \mathrm{curl}_\mathrm{x}\left(U \times \beta^{i-4}\right) \\
&+ \mathrm{curl}_\theta \sum_{j+J=i-m-3} w^j \times \beta^J + \mathrm{curl}_\mathrm{x} \sum_{j+J=i-m-5} w^j \times \beta^J \bigg),
\end{aligned}$$

which is exactly (2.10). Hence,

$$\left(p^i_*, 0\right)^t = \Delta_\theta^{-1}\mathrm{Div}_\theta B_\theta\left(\tilde{U}, \overline{V}^i\right) + H^i_*, \tag{2.11}$$

where H^i_* depends only on $X^0, \ldots, X^{i-1}$. Solving $(T_i\mathrm{a})$ yields in turn

$$V^{i+1}_* = (\partial_\tau - \Delta)^{-1} L_\theta \overline{V}^i + I^i_*,$$

where I^i_* depends only on $X^0, \ldots, X^{i-1}$. We are left with equations of the following type:

$$\begin{cases}
(\partial_t - \Delta_\mathrm{x})\overline{V}^i = \left(\nabla_\mathrm{x}\overline{p}^{i+1}, 0\right)^t + \mathcal{A}\,\overline{V}^i + \overline{J}^i \\
\qquad\qquad - \delta_{0m}\left(B_\mathrm{x}\left(\overline{V}^0, \overline{V}^i\right) + B_\mathrm{x}\left(\overline{V}^i, \overline{V}^0\right)\right), \\
\mathrm{Div}_\mathrm{x}\overline{V}^i = 0,
\end{cases} \tag{2.12}$$

where δ_{0m} is the Kronecker symbol, and where

$$\overline{J}^i \in C^\infty\left([0,T];\ H^\infty\left(\mathbb{R}^3\right)\right)^6$$

depends on $X^0, \dots, X^{i-1}$. This system is linear and of parabolic type. It follows easily that for all initial data $\overline{V}^i|_{t=0} = \overline{V}^i_0$ in $H^\infty\left(\mathbb{R}^3\right)^6$, $\mathrm{Div_x}\overline{V}^i_0 = 0$, such a system has a unique solution

$$\left(\overline{V}^i, \overline{p}^i\right) \in C^\infty\left([0,T];\, H^\infty\left(\mathbb{R}^3\right)\right)^7, \quad \int \overline{p}^i(t,\cdot) = 0.$$

Back to system (T_i), this ends the induction.

2.2. A priori estimates. We now establish some stability estimates on systems of type (2.2) that will be used in sections 2.3 and 3.3. More precisely, let $\{T^\varepsilon\}_{\varepsilon>0}$ be a family of times, and let $\left\{V^\varepsilon_{app}\right\}_{\varepsilon>0}$, $\{F^\varepsilon\}_{\varepsilon>0}$ be families of functions satisfying,

$$\forall\varepsilon, \quad V^\varepsilon_{app}, F^\varepsilon_{app} \in C\left([0,T^\varepsilon];\, H^\infty\left(\mathbb{T}\times\mathbb{T}^3\times\mathbb{R}^3\right)\right)^6,$$

and such that

$$\sup_{\varepsilon}\sup_{0\le t\le T^\varepsilon}\left\|\partial_t^\alpha\partial_{\mathrm{x},\tau,\theta}^\beta V^\varepsilon_{app}(t,\cdot)\right\|_{L^\infty} \le C_{\alpha,\beta} \quad \forall\alpha,\beta. \tag{2.13}$$

We define

$$U^\varepsilon := \tilde{U} + \varepsilon V^\varepsilon_{app}, \quad \tilde{U} = (U,0)^t, \quad U \in \mathcal{P},$$

and we consider the following equations:

$$\begin{cases} \partial_t V + \varepsilon^{-4}\partial_\tau V + \varepsilon^{-1}(B_\mathrm{x} + \varepsilon^{-2}B_\theta)(U^\varepsilon, V) + \dfrac{1}{2}(B_\mathrm{x} + \varepsilon^{-2}B_\theta)(V,V) \\ \qquad -\left(\nabla_\mathrm{x} + \varepsilon^{-2}\nabla_\theta\right)^2 V = \left(\left(\nabla_\mathrm{x} + \varepsilon^{-2}\nabla_\theta\right)p,\, 0\right)^t + F^\varepsilon, \\ \mathrm{Div_x}V + \varepsilon^{-2}\mathrm{Div}_\theta V = 0. \end{cases} \tag{2.14}$$

We distinguish between the low-frequency part V_l and the high-frequency part V_h of V. We introduce $\chi = \chi(\zeta,\xi) \in C^\infty\left(\mathbb{R}^3\times\mathbb{T}^3\right)$ such that

$$\begin{aligned} \chi(\zeta,\xi) &= 1 \quad \text{for } |\zeta+\xi| \le \delta, \\ \chi(\zeta,\xi) &= 0 \quad \text{for } |\zeta+\xi| \ge 2\delta, \end{aligned}$$

where δ is a fixed number satisfying $0 < \delta < 1/4$. Then we set

$$V_l = \chi(\varepsilon^2 D_\mathrm{x}, D_\theta)V, \quad p_l = \chi(\varepsilon^2 D_\mathrm{x}, D_\theta)p, \quad V_h = V - V_l, \quad p_h = p - p_l,$$

where for any f, $\chi(\varepsilon^2 D_\mathrm{x}, D_\theta)f$ is the Fourier multiplier defined as

$$\mathcal{F}\left(\chi(\varepsilon^2 D_\mathrm{x}, D_\theta)f\right)(\zeta,\xi) = \chi(\varepsilon^2\zeta,\xi)\mathcal{F}(f)(\zeta,\xi),$$

$\mathcal{F}$ being the Fourier transform with respect to x and θ. Finally, we define for all $s \in N$ and for all $t \in [0,T^\varepsilon]$

$$\begin{aligned} \Psi_s(V;t) &:= \|V_l(t)\|^2_{H^s} + \varepsilon^2\|V_h(t)\|^2_{H^s} + \int_0^t \|\left(\nabla_\mathrm{x} + \varepsilon^{-2}\nabla_\theta\right)V_l(u)\|^2_{H^s}\,du \\ &\qquad + \varepsilon^{-2}\int_0^t \|V_h(u)\|^2_{H^s}\,du + \varepsilon^{-2}\int_0^t \|(\varepsilon^2\nabla_\mathrm{x} + \nabla_\theta)V_h(u)\|^2_{H^s}\,du, \\ \alpha_s(V;t) &= \sup_{0\le u\le t}\Psi_s(V;u). \end{aligned}$$

We show the following.

Proposition 2.1. *Let $V \in C^0\left([0,T^\varepsilon]; H^\infty\right)^6$, satisfying* (2.14). *Then the following inequality holds, for all $s \geq 5$, for ε small enough:*

$$\alpha_s(V;t) \leq \mathcal{C}_s\left(\alpha_s(V;0) + \varepsilon^6 \int_0^t \|F_h^\varepsilon(u)\|_{H^s}^2\, du + \int_0^t \|F_l^\varepsilon(u)\|_{H^s}^2\, du \right. \\ \left. + (1+\alpha_s(V;t)) \int_0^t \alpha_s(V;u)\, du \, + \, \alpha_s(V;t)^2\right). \tag{2.15}$$

Proof. We start with an L^2 estimate on (2.14).

L^2 estimates.

- *Estimates on V_h.* We apply $\left(1-\chi\left(\varepsilon^2 D_{\mathrm{x}}, D_\theta\right)\right)$ to (2.14). Multiplication by V_h and integration yield

$$\|V_h(t)\|_{L^2}^2 + \varepsilon^{-4}\int_0^t \|(\varepsilon^2\nabla_{\mathrm{x}} + \nabla_\theta)V_h\|_{L^2}^2 \leq \|V_h(0)\|_{L^2}^2 \\ + \frac{\|U^\varepsilon\|_\infty}{\varepsilon^3}\int_0^t \|V\|_{L^2}\, \|(\varepsilon^2\nabla_{\mathrm{x}} + \nabla_\theta)V_h\|_{L^2} \\ + \int_0^t \|V\otimes V\|_{L^2}\, \|(\nabla_{\mathrm{x}} + \varepsilon^{-2}\nabla_\theta)V_h\|_{L^2} + \int_0^t \|F_h^\varepsilon\|_{L^2}\, \|V_h\|_{L^2}.$$

Recall that notation $V\otimes V$ refers to the matrix $(V_j\, V_k)_{j,k}$.

By the Plancherel formula,

$$\|(\varepsilon^2\nabla_{\mathrm{x}} + \nabla_\theta)V_h\|_{L^2}^2 = \frac{1}{(2\pi)^6}\int |\varepsilon^2\zeta + \xi|^2 |\mathcal{F}(V_h)(\zeta,\xi)|^2 d\zeta d\xi \\ \geq \frac{\delta^2}{(2\pi)^6}\int |\mathcal{F}(V_h)(\zeta,\xi)|^2 d\zeta d\xi = \delta^2 \|V^h\|_{L^2}^2.$$

Hence, for ε small enough,

$$\|V_h(t)\|_{L^2}^2 + \varepsilon^{-4}\int_0^t \|V_h\|_{L^2}^2 + \varepsilon^{-4}\int_0^t \|(\varepsilon^2\nabla_{\mathrm{x}} + \nabla_\theta)V_h\|_{L^2}^2 \leq C\biggl(\|V_h(0)\|_{L^2}^2 \\ + \varepsilon^{-2}\int_0^t \|V_l\|_{L^2}^2 + \int_0^t \|V\otimes V\|_{L^2}^2 + \varepsilon^4\int_0^t \|F_h^\varepsilon\|_{L^2}^2\biggr). \tag{2.16}$$

- *Estimates on V_l.* The low-frequency part V_l satisfies

$$\partial_t V_l - (\nabla_{\mathrm{x}} + \varepsilon^{-2}\nabla_\theta))^2 V_l = -\left((\nabla_{\mathrm{x}} + \varepsilon^{-2}\nabla_\theta)p_l, 0\right)^t \\ - \frac{1}{2}\chi(\varepsilon^2 D_{\mathrm{x}}, D_\theta)\left(B_{\mathrm{x}} + \varepsilon^{-2}B_\theta\right)(V,V) \, - \, \varepsilon^{-1}\chi(\varepsilon^2 D_{\mathrm{x}}, D_\theta)\, \left(B_{\mathrm{x}} + \varepsilon^{-2}B_\theta\right)(U^\varepsilon, V).$$

The key observation is that

$$\chi(\varepsilon^2 D_{\mathrm{x}}, D_\theta)\left(B_{\mathrm{x}} + \varepsilon^{-2}B_\theta\right)(U^\varepsilon, V) = \chi(\varepsilon^2 D_{\mathrm{x}}, D_\theta)\left(B_{\mathrm{x}} + \varepsilon^{-2}B_\theta\right)(\tilde{U}, V_h) \\ + \varepsilon\, \chi(\varepsilon^2 D_{\mathrm{x}}, D_\theta)\left(B_{\mathrm{x}} + \varepsilon^{-2}B_\theta\right)(V_{app}^\varepsilon, V).$$

Indeed,

$$\begin{aligned}\text{supp } \mathcal{F}\left(\left(B_{\mathrm{x}}+\varepsilon^{-2}B_\theta\right)\left(\tilde{U},V_l\right)\right) &\subset \text{ supp } \mathcal{F}(U) \ + \ \text{ supp } \mathcal{F}(V_l)\\ &\subset \bigcup_{\xi'\in\mathbb{Z}^3-\{0\}} \{(0,\xi')\} \ + \ \left\{(\zeta,\xi),\ |\varepsilon^2\zeta+\xi|\le 2\delta\right\}\\ &\subset \left\{(\zeta,\xi), \quad |\varepsilon^2\zeta+\xi|\ge 1-2\delta\right\}\end{aligned}$$

As $\delta < 1/4$, we get

$$\text{supp } \mathcal{F}(U) \ + \ \text{ supp } \mathcal{F}(V_l) \ \subset \ \left\{(\zeta,\xi), \quad |\varepsilon^2\zeta+\xi| > 2\delta\right\}$$

so that

$$\chi(\varepsilon^2 D_{\mathrm{x}}, D_\theta)\left(B_{\mathrm{x}}+\varepsilon^{-2}B_\theta\right)(\tilde{U},V_l)=0.$$

We end with the following energy estimates:

$$\begin{aligned}\|V_l(t)\|_{L^2}^2 + \int_0^t \|(\nabla_{\mathrm{x}}+\varepsilon^{-2}\nabla_\theta)V_l\|_{L^2}^2 \le C\Big(&\|V_l(0)\|_{L^2}^2\\ &+\varepsilon^{-2}\int_0^t \|V_h\|_{L^2}^2 + \int_0^t \|V_l\|_{L^2}^2 + \int_0^t \|V\otimes V\|_{L^2}^2 + \int_0^t \|F_l^\varepsilon\|_{L^2}^2\Big).\end{aligned}$$

Combining with (2.16), we get

$$\Psi_0(V;t) \le C\Big(\Psi_0(V;0)+\int_0^t \|V_l\|_{L^2}^2+\int_0^t \|V\otimes V\|_{L^2}^2+\varepsilon^6\int_0^t \|F_h^\varepsilon\|_{L^2}^2+\int_0^t \|F_l^\varepsilon\|_{L^2}^2\Big).$$

H^s *estimates.* We do not detail the derivation of the H^s estimates, $s\ge 1$. They follow from differentiating system (2.14) and applying the same argument as above. We get, for all s and for small enough ε,

$$\begin{aligned}\Psi_s(V;t) \ \le \ C_s\Big(&\Psi_s(V;0)+\int_0^t \|V_l\|_{H^s}^2 + \int_0^t \|V\otimes V\|_{H^s}^2\\ &+\varepsilon^6\int_0^t \|F_h^\varepsilon\|_{H^s}^2 + \int_0^t \|F_l^\varepsilon\|_{H^s}^2\Big).\end{aligned}\tag{2.17}$$

It remains to handle the quadratic terms. As $s \ge 5 > 7/2$, $H^s\left(\mathbb{R}^3\times\mathbb{T}\times\mathbb{T}^3\right)$ is a Banach algebra. Hence, we obtain the bounds

$$\begin{aligned}\int_0^t \|V_l\otimes V_h\|_{H^s}^2 &\le \int_0^t \|V_l\|_{H^s}^2\|V_h\|_{H^s}^2\\ &\le \left(\sup_{0\le u\le t}\|V_l(u)\|_{H^s}^2\right)\left(\int_0^t \|V_h\|_{H^s}^2\right)\end{aligned}$$

and, in the same way,

$$\int_0^t \|V_l\otimes V_l\|_{H^s}^2 \le \left(\sup_{0\le u\le t}\|V_l(u)\|_{H^s}^2\right)\left(\int_0^t \|V_l\|_{H^s}^2\right),$$

$$\int_0^t \|V_h \otimes V_h\|_{H^s}^2 \leq \left(\sup_{0\leq u\leq t} \varepsilon^2 \|V_h(u)\|_{H^s}^2\right) \left(\varepsilon^{-2}\int_0^t \|V_h\|_{L^2}^2\right).$$

If we inject these bounds in (2.17), we obtain

$$\Psi_s(V;t) \leq C_s \left(\Psi_s(V;0) + \varepsilon^6 \int_0^t \|F_h^\varepsilon(u)\|_{H^s}^2 du + \int_0^t \|F_l^\varepsilon(u)\|_{H^s}^2 du \right.$$
$$\left. + (1+\alpha_s(V;t)) \int_0^t \Psi_s(V;u)du + \alpha_s(V;t)\Psi_s(V;t)\right). \tag{2.18}$$

Bound (2.15) follows. This ends the proof of the proposition.

2.3. Proof of Theorem 1.1. We now turn to the proof of Theorem 1.1. Let $\overline{V}_0 = (v_0, b_0)$ in $H^\infty\left(\mathbb{R}^3\right)^6$, such that $\mathrm{Div}_{\mathrm{x}}\overline{V}_0 = 0$. From computations of section 2.1, we infer the existence of $\delta > 0$, such that if

$$m \geq 1 \quad \text{or} \quad \|\overline{V}_0\|_{H^{1/2}} \leq \delta, \tag{2.19}$$

there exist approximate solutions of (2.2), indexed by $n \in \mathbb{N}$,

$$V_{app}^{\varepsilon,n} = \varepsilon^m \left(\overline{V}^0(t,\mathrm{x}) + \sum_{i=1}^n \varepsilon^i V^i\left(t,\mathrm{x},\tau,\theta\right) + \varepsilon^{n+1} V_*^{n+1}\right), \tag{2.20}$$

$$p_{app}^{\varepsilon,n} = \varepsilon^m \left(p_*^0(t,\mathrm{x},\tau,\theta) + \sum_{i=1}^n \varepsilon^i p^i\left(t,\mathrm{x},\tau,\theta\right) + \varepsilon^{n+1} \overline{p}^{n+1}\right), \tag{2.21}$$

with $\overline{V}^0|_{t=0} = \overline{V}_0$, and $\overline{V}^i|_{t=0} = 0$ for $i \geq 1$. Profiles V^i and p^i are found using the recursion of the previous section. They satisfy

$$\begin{cases} \partial_t V_{app}^{\varepsilon,n} + \varepsilon^{-4}\partial_\tau V_{app}^{\varepsilon,n} + \varepsilon^{-1}\left(B_{\mathrm{x}} + \varepsilon^{-2}B_\theta\right)(\tilde{U}, V_{app}^{\varepsilon,n}) \\ \qquad + \dfrac{1}{2}\left(B_{\mathrm{x}} + \varepsilon^{-2}B_\theta\right)(V_{app}^{\varepsilon,n}, V_{app}^{\varepsilon,n}) - \left(\nabla_{\mathrm{x}} + \varepsilon^{-2}\nabla_\theta\right)^2 V_{app}^{\varepsilon,n} \\ \qquad = -\left(\left(\nabla_{\mathrm{x}} + \varepsilon^{-2}\nabla_\theta\right) p_{app}^{\varepsilon,n},\, 0\right)^t + R_{app}^{\varepsilon,n}, \\ \left(\mathrm{Div}_{\mathrm{x}} \,+\, \varepsilon^{-2}\mathrm{Div}_\theta\right) V_{app}^{\varepsilon,n} = r_{app}^{\varepsilon,n}, \end{cases} \tag{2.22}$$

where for all $s, \alpha \in \mathbb{N}$, the remainders $R_{app}^{\varepsilon,n}$ and $r_{app}^{\varepsilon,n}$ satisfy the following estimates:

$$\sup_{0\leq t\leq T^\varepsilon} \|\partial_t^\alpha R_{app}^{\varepsilon,n}(t)\|_{H^s} \leq C_{\alpha,s}\,\varepsilon^{m+n-2}, \qquad \sup_{0\leq t\leq T^\varepsilon} \|\partial_t^\alpha r_{app}^{\varepsilon,n}(t)\|_{H^s} \leq C_{\alpha,s}\,\varepsilon^{m+n}. \tag{2.23}$$

We want the second equation of (2.22) to become homogeneous. This is useful in the energy estimates for getting rid of the gradient pressure term. We need to build a function W_{app}^ε, "sufficiently small," such that

$$\left(\mathrm{Div}_{\mathrm{x}} \,+\, \varepsilon^{-2}\mathrm{Div}_\theta\right) W_{app}^\varepsilon = -r_{app}^{\varepsilon,n}.$$

A natural attempt to do so would be to look for W_{app}^ε in the form

$$W_{app}^\varepsilon = \begin{pmatrix} \nabla_{\mathrm{x}} + \varepsilon^{-2}\nabla_\theta & \\ & \nabla_{\mathrm{x}} + \varepsilon^{-2}\nabla_\theta \end{pmatrix} \Psi^\varepsilon, \tag{2.24}$$

with

$$-\left(\nabla_{\mathrm{x}} + \varepsilon^{-2}\nabla_\theta\right)^2 \Psi^\varepsilon = r^{\varepsilon,n}_{app}. \tag{2.25}$$

However, it is not obvious that (2.25) has a solution. Indeed, the operator $-(\nabla_{\mathrm{x}} + \varepsilon^{-2}\nabla_\theta)^2$ is not elliptic: its symbol $(\zeta + \varepsilon^{-2}\xi)^2$ cancels for all $\zeta = -\varepsilon^{-2}\xi,\ \xi \in \mathbb{Z}$.

To get rid of this difficulty, we must again distinguish between average and oscillations, and between low frequencies and high frequencies. We first notice that

$$r^{\varepsilon,n}_{app} = \varepsilon^{m+n}\mathrm{Div}_{\mathrm{x}}\left(V^n_* + \varepsilon V^{n+1}_*\right) = r^{\varepsilon,n}_{app,*}.$$

Then we infer from the regularity of $V^{\varepsilon,n}_{app}$ that

$$\begin{aligned}
&\|\partial_t^\alpha\partial_\tau^\beta V^{\varepsilon,n}_{app,*,l}(t,\cdot,\tau,\cdot)\|^2_{H^s}\\
&\le \frac{1}{(2\pi)^6}\int_{\{|\varepsilon^2\zeta+\xi|\,\le\,2\delta\}}(1+|\zeta|^2+|\xi|^2)^s\,\left|\mathcal{F}(\partial_t^\alpha\partial_\tau^\beta V^{\varepsilon,n}_{app,*}(t,\cdot,\tau,\cdot))\right|^2\,d\zeta d\xi\\
&\le \frac{1}{(2\pi)^6}\int_{\{|\zeta|\,\ge\,(1-2\delta)\,\varepsilon^{-2}\}}(1+|\zeta|^2+|\xi|^2)^s\,\left|\mathcal{F}(\partial_t^\alpha\partial_\tau^\beta V^{\varepsilon,n}_{app,*}(t,\cdot,\tau,\cdot))\right|^2\,d\zeta d\xi\\
&\le C_k\,\varepsilon^{4k}\,\|\partial_t^\alpha\partial_\tau^\beta V^{\varepsilon,n}_{app}(t,\cdot,\tau,\cdot)\|_{H^{s+k}} = O(\varepsilon^{4k}) \qquad \forall k>0.
\end{aligned} \tag{2.26}$$

Thus, redefining

$$V^{\varepsilon,n}_{app} := \overline{V}^{\varepsilon,n}_{app} + V^{\varepsilon,n}_{app,*,h},$$

we see that $V^{\varepsilon,n}_{app}$ still satisfies a system of type (2.22), with estimates (2.23), and the new remainder $r^{\varepsilon,n}_{app}$ is such that

$$r^{\varepsilon,n}_{app,l} = 0.$$

In particular, the function W^ε_{app} given by (2.24), (2.25) is well-defined and is bounded by

$$\|\partial_t^\alpha W^\varepsilon_{app}\|_{H^{s-1}} \le C_{\alpha,s}\,\varepsilon^2\,\|\partial_t^\alpha r^\varepsilon_{app}\|_{H^s} \quad \forall \alpha, s.$$

Finally, we set $V^\varepsilon_{app} = V^{\varepsilon,n}_{app} + W^\varepsilon_{app}$. This solves

$$(2.27)\qquad \begin{cases} \partial_t V^\varepsilon_{app} + \varepsilon^{-4}\partial_\tau V^\varepsilon_{app} + \varepsilon^{-1}\left(B_{\mathrm{x}} + \varepsilon^{-2}B_\theta\right)(\tilde{U}, V^\varepsilon_{app}) \\ \qquad\qquad + \dfrac{1}{2}\left(B_{\mathrm{x}} + \varepsilon^{-2}B_\theta\right)(V^\varepsilon_{app}, V^\varepsilon_{app}) - \left(\nabla_{\mathrm{x}} + \varepsilon^{-2}\nabla_\theta\right)^2 V^\varepsilon_{app} \\ \qquad\qquad = -\left(\left(\nabla_{\mathrm{x}} + \varepsilon^{-2}\nabla_\theta\right)p^\varepsilon_{app},\, 0\right)^t + R^\varepsilon_{app}, \\ \mathrm{Div}_{\mathrm{x}} V \;+\; \varepsilon^{-2}\mathrm{Div}_\theta V = 0, \end{cases}$$

where

$$(2.28)\qquad \sup_{0\le t\le T^\varepsilon} \|R^\varepsilon_{app}(t)\|_{H^s} \le C_s\,\varepsilon^{m+n-2} \quad \forall s.$$

We can now use the results of section 2.2. We fix $n \ge 4$, $s \ge 5$ and assume (2.19). For ε small enough, we will show the existence of a solution of (2.2), $V^\varepsilon \in C^0\left([0,T];\, H^\infty\left(\mathbb{R}^3\times\mathbb{T}\times\mathbb{T}^3\right)\right)^6$, $V^\varepsilon|_{t=0} = \varepsilon^m\overline{V}_0$.

The local existence theory of smooth solutions for (2.2) is classical (see, for instance, [11]). For all $\varepsilon > 0$, there exists a unique maximal solution

$$V^\varepsilon \in C^0\left([0,T^\varepsilon_*);\, H^\infty\left(\mathbb{R}^3\times\mathbb{T}\times\mathbb{T}^3\right)\right)^6, \quad V^\varepsilon|_{t=0} = \varepsilon^m\overline{V}_0.$$

Moreover, the lifespan T^ε_* satisfies one of the following conditions:

- $T^\varepsilon_* \ge T$,
- $T^\varepsilon_* < T$ and $\liminf_{t\to T^\varepsilon_*} \|V^\varepsilon(t)\|_{H^s} \to +\infty$.

It is enough to show that the second possibility does not occur for small enough ε. Let $T^\varepsilon < \min(T^\varepsilon_*, T)$, and define $W^\varepsilon := V^\varepsilon - V^\varepsilon_{app}$ on $[0,T^\varepsilon]$. Then W^ε is a solution of (2.14), where V^ε_{app} defined above satisfies (2.13), and $F^\varepsilon = R^\varepsilon_{app}$ satisfies (2.28).

We apply Proposition 2.1. This yields, for ε small enough,

$$(2.29)\qquad \alpha_s(W^\varepsilon;t) \le C\Bigg(\varepsilon^{2m+4} + \alpha_s(W^\varepsilon;0) + (1+\alpha_s(W^\varepsilon;t))\int_0^t \alpha_s(W^\varepsilon;u)du \\ + \alpha_s(W^\varepsilon;t)^2\Bigg).$$

Note that $W^\varepsilon|_{t=0} = \varepsilon^m\overline{V}_0 - V^\varepsilon_{app}|_{t=0}$ satisfies

$$\overline{W}^\varepsilon|_{t=0} = 0, \quad \|W^\varepsilon_*|_{t=0}\|_{H^s} = O(\varepsilon^{m+1}).$$

As in (2.26), we deduce that

$$(2.30)\qquad \|W^\varepsilon_l|_{t=0}\|_{H^s} = O(\varepsilon^k) \quad \forall k$$

and that

$$\alpha_s(W^\varepsilon;0) = O(\varepsilon^{2m+4}).$$

We introduce

$$T(\varepsilon) := \sup\left\{t\in[0,T^\varepsilon), \quad \alpha_s(W^\varepsilon;t) < \varepsilon^{2m+3}\right\}.$$

For $\varepsilon > 0$ small enough, $T(\varepsilon)$ is well-defined and positive. Moreover, using (2.29), for all $t < T(\varepsilon)$,

$$(1-\varepsilon^{2m+3})\,\alpha_s(W^\varepsilon;t) \le C\left(\varepsilon^{2m+4} \;+\; (1+\varepsilon^{2m+3})\int_0^t \alpha_s(W^\varepsilon;u)du\right).$$

The Gronwall's lemma implies, for all $t < T(\varepsilon)$, $\alpha_s(W^\varepsilon;t) \le C'\,\varepsilon^{2m+4}$. This last inequality shows that $T(\varepsilon) = T^\varepsilon$, and that

$$\sup_{0\le t\le T^\varepsilon} \alpha_s(W^\varepsilon;t) \;\le\; C''\varepsilon^{2m+4}.$$

In particular, we get

$$\sup_{0\le t\le T^\varepsilon} \|W^\varepsilon(t,\cdot)\|_{H^s} \le C\varepsilon^{m+1}.$$

Back to $V^\varepsilon = V^\varepsilon_{app} + W^\varepsilon$, we obtain, for all s,

$$\begin{aligned}\sup_{0\le t\le T^\varepsilon}\|V^\varepsilon(t,\cdot)\|_{H^s} &\le \sup_{0\le t\le T^\varepsilon}\|V^\varepsilon_{app}(t,\cdot)\|_{H^s} + \sup_{0\le t\le T^\varepsilon}\|W^\varepsilon(t,\cdot)\|_{H^s}\\ &\le\; C\varepsilon^m\|\overline{V}_0\|_{H^s}.\end{aligned} \tag{2.31}$$

This yields $T^\varepsilon_* \ge T$ and shows the existence on $[0,T]$ of a smooth solution V^ε with initial data $\overline{V}_0$.

As explained at the beginning of section 2, this provides a smooth solution

$$(v^\varepsilon, b^\varepsilon)^t(t,\mathrm{x}) = V^\varepsilon\left(t,\mathrm{x},\varepsilon^{-4}t,\varepsilon^{-2}\mathrm{x}\right)$$

of (1.5), with initial data $(\varepsilon^m v_0, \varepsilon^m b_0)^t$. Uniqueness of the solution $(v^\varepsilon, b^\varepsilon)$ is a direct consequence of its regularity. Finally, the estimates of the theorem follow from (2.31). This ends the proof.

3. Instability mechanism. We now begin the description of the instability mechanism leading to Theorem 1.2. As we will see, this mechanism is connected to the behavior of the WKB solutions (2.3).

3.1. Spectral analysis. To understand the instability process requires the study of system (2.8), satisfied by $V^0 = \overline{V}^0(t,\mathrm{x})$ when $m \ge 1$. This system reads

$$\begin{cases}\partial_t\overline{w}^0 = \mathcal{A}_1\overline{w}^0 + \nabla_{\mathrm{x}}\overline{p}^1 + \Delta_{\mathrm{x}}\overline{w}^0,\\ \partial_t\overline{\beta}^0 = \mathcal{A}_2\overline{\beta}^0 + \Delta_{\mathrm{x}}\overline{\beta}^0,\\ \mathrm{div}_{\mathrm{x}}\overline{w}^0 = \mathrm{div}_{\mathrm{x}}\overline{\beta}^0 = 0,\end{cases} \tag{3.1}$$

where the operators

$$\mathcal{A}_1\overline{w} = \mathrm{div}_{\mathrm{x}}(A_1\overline{w}), \quad \mathcal{A}_2\overline{\beta} = \mathrm{curl}_{\mathrm{x}}(A_2\overline{\beta})$$

involve the linear operators $A_1 \in \mathcal{L}\left(\mathbb{R}^3, \mathcal{M}_3\left(\mathbb{R}\right)\right)$, $A_2 \in \mathcal{L}\left(\mathbb{R}^3,\mathbb{R}^3\right)$. They are defined in the following way: for all $b \in \mathbb{R}^3$,

$$\begin{aligned}A_1 b &= -\int_{\tau,\theta}\left((\partial_\tau - \Delta_\theta)^{-1}\left(\nabla_\theta\Delta_\theta^{-1}\mathrm{div}^2_\theta(U\otimes b) \;-\; \mathrm{div}_\theta\left(U\otimes b\right)\right)\right)\otimes U,\\ A_2 b &= \int_{\tau,\theta} U\times\left((\partial_\tau - \Delta_\theta)^{-1}\mathrm{curl}_\theta(U\times b)\right).\end{aligned}$$

As

$$\mathrm{div}_\theta(U\otimes b) = b\cdot\nabla_\theta U = \mathrm{curl}_\theta(U\times b),$$

we deduce $\mathrm{div}^2_\theta(U\otimes b)=0$ and

$$A_1 b = \int_{\tau,\theta} \otimes\Big((\partial_\tau-\Delta_\theta)^{-1}\mathrm{curl}_\theta\,(U\times b)\Big)\otimes U.$$

Then

$$\begin{aligned}\mathcal{A}_1 a &= -\mathrm{div}_{\mathrm{x}}\int_{\tau,\theta}\Big((\partial_\tau-\Delta_\theta)^{-1}\mathrm{curl}_\theta\,(U\times b)\Big)\otimes U\\ &= \int_{\tau,\theta} U\cdot\nabla_{\mathrm{x}}(\partial_\tau-\Delta_\theta)^{-1}\mathrm{curl}_\theta(U\times b) = -\mathcal{A}_2 b.\end{aligned}$$

If we relabel, for all $b\in\mathbb{R}^3$,

$$Ab = \int_{\tau,\theta} U\times\Big((\partial_\tau-\Delta_\theta)^{-1}\mathrm{curl}_\theta(U\times b)\Big),$$

the system on $\overline{V}^0$ eventually reads

$$\begin{cases}\partial_t\overline{w}^0 = -\mathrm{curl}_{\mathrm{x}}A\,\overline{w}^0+\nabla_{\mathrm{x}}\overline{p}^1+\Delta_{\mathrm{x}}\overline{w}^0,\\ \partial_t\overline{\beta}^0 = \mathrm{curl}_{\mathrm{x}}A\,\overline{\beta}^0+\Delta_{\mathrm{x}}\overline{\beta}^0,\\ \mathrm{div}_{\mathrm{x}}\overline{w}^0 = \mathrm{div}_{\mathrm{x}}\overline{\beta}^0=0.\end{cases}\tag{3.2}$$

Remark 6. In the physicists' community, the matrix A is often denoted α, which motivates the expression *alpha effect*.

We focus on solutions of (3.2) with initial data $(0,\overline{\beta}_0)^t$, $\mathrm{div}_{\mathrm{x}}\,\overline{\beta}_0=0$. Thus, $\overline{w}_0=0$ for all times, and the divergence-free condition $\mathrm{div}_{\mathrm{x}}\overline{\beta}^0=0$ is fulfilled for all times. If we set $b^0:=\overline{\beta}^0$, we are left with the study of

$$\partial_t b^0 - \mathrm{curl}_{\mathrm{x}}\left(Ab^0\right) - \Delta_{\mathrm{x}} b^0 = 0.\tag{3.3}$$

We first state some properties of the matrix A.

Lemma 3.1. *For all $U\in\mathcal{P}$, the matrix $A=A(U)$ is real-symmetric. Moreover, the set*

$$\Omega = \{U\in\mathcal{P},\quad A(U)\ \textit{has simple nonzero eigenvalues}\}$$

is dense and open in $\mathcal{P}$.

Proof. Using the Fourier coefficients of U, we compute, for all $b\in\mathbb{R}^3$,

$$A\,b = \sum_{\substack{(\omega,k)\in\mathbb{Z}^4\\(\omega,k)\neq 0}} \frac{\hat{U}(\omega,k)\times\hat{U}(-\omega,-k)\,(-ib\cdot k)}{i\omega+|k|^2}.\tag{3.4}$$

Note that the change of indices $(\omega, k) \to (\omega' = -\omega, k' = -k)$ yields

$$\begin{aligned} A\,b &= \sum_{(\omega',k')} \frac{\hat{U}(-\omega',-k') \times \hat{U}(\omega',k')\,(ib\cdot k')}{-i\omega' + |k'|^2} \\ &= \sum_{(\omega',k')} \frac{\hat{U}(\omega',k') \times \hat{U}(-\omega',-k')\,(-ib\cdot k')}{-i\omega' + |k'|^2}, \end{aligned}$$

so that

$$A\,b = \int_{\mathbb{T}^4} U \times (-\partial_\tau - \Delta_\theta)^{-1} \mathrm{curl}_\theta\,(U \times b)\,. \tag{3.5}$$

As U is real-valued, $\hat{U}(-\omega,-k) = \hat{U}^*(\omega,k)$, where the asterisk denotes complex conjugation. We deduce, for all $b \in \mathbb{R}^3$,

$$\begin{aligned} A^*\,b &= \sum_{(\omega,k)} \frac{\hat{U}^*(\omega,k) \times \hat{U}^*(-\omega,-k)\,(ib\cdot k)}{-i\omega + |k|^2} \\ &= \sum_{(\omega,k)} \frac{\hat{U}(-\omega,-k) \times \hat{U}(\omega,k)\,(ib\cdot k)}{-i\omega + |k|^2} \\ &= \sum_{(\omega',k')} \frac{\hat{U}(\omega',k') \times \hat{U}(-\omega',-k')\,(-ib\cdot k')}{i\omega' + |k'|^2} = \alpha\,b, \end{aligned}$$

so that α is real. Then we compute

$$\begin{aligned} A\,b\cdot\tilde{b} &= \int_{\mathbb{T}^4} \Big(U \times (\partial_\tau - \Delta_\theta)^{-1} \mathrm{curl}_\theta\,(U\times b) \Big) \cdot \tilde{b} \\ &= -\int_{\mathbb{T}^4} (U \times \tilde{b}) \cdot \Big((\partial_\tau - \Delta_\theta)^{-1} \mathrm{curl}_\theta\,(U\times b) \Big) \\ &= -\int_{\mathbb{T}^4} (U \times \tilde{b}) \cdot \Big(\mathrm{curl}_\theta\,(\partial_\tau - \Delta_\theta)^{-1}\,(U\times b) \Big) \\ &= -\int_{\mathbb{T}^4} \Big((-\partial_\tau - \Delta_\theta)^{-1} \mathrm{curl}_\theta \Big(U\times\tilde{b}\Big) \Big) \cdot \Big(U\times b \Big) \\ &= \int_{\mathbb{T}^4} \Big(U \times (-\partial_\tau - \Delta_\theta)^{-1} \mathrm{curl}_\theta \Big(U\times\tilde{b}\Big) \Big) \cdot b = \alpha\,\tilde{b}\cdot b \end{aligned}$$

using (3.5). Thus, A is symmetric.

From (3.4), we deduce that

$$\mathcal{P} \mapsto \mathcal{M}_3\,(\mathbb{R})\,, \quad U \mapsto A(U)$$

is continuous. This clearly implies that Ω is open in $\mathcal{P}$. Let $\varepsilon > 0$, and $U \in \mathcal{P} - \Omega$. Let

$$U^n(\tau,\theta) = \sum_{|\omega|+|k|\le n} \hat{U}(\omega,k)\,e^{i(\omega\tau + k\cdot\theta)}.$$

There exists N, such that $d_{\mathcal{P}}\left(U, U^N\right) < \varepsilon/2$. If $U^N \in \Omega$, we are done. Otherwise, we consider

$$\tilde{U} = U^N \; + \; \sum_{i=1}^{3} \delta_i \, V^i, \quad \delta_i > 0 \quad \forall i,$$

where

$$\begin{aligned} V^1(\theta_1,\theta_2,\theta_3) = \Big(& \cos((N+1)\theta_2), \sin((N+1)\theta_1), \\ & \cos((N+1)\theta_1) + \sin((N+1)\theta_2) \Big)^t, \\ V^2(\theta_1,\theta_2,\theta_3) = \Big(& \sin((N+2)\theta_3), \sin((N+2)\theta_1) + \cos((N+2)\theta_3), \\ & \cos((N+2)\theta_1) \Big)^t, \\ V^3(\theta_1,\theta_2,\theta_3) = \Big(& \sin((N+3)\theta_3) + \cos((N+3)\theta_2), \cos((N+3)\theta_3), \\ & \sin((N+3)\theta_2) \Big)^t. \end{aligned}$$

Note that the V^i's are special cases of the famous ABC flows [1] of the type

$$\begin{aligned} V(\theta_1,\theta_2,\theta_3) = \Big(& C\sin(M\theta_3) + B\cos(M\theta_2), A\sin(M\theta_1) + C\cos(M\theta_3), \\ & B\sin(M\theta_2) + A\cos(M\theta_1) \Big)^t. \end{aligned}$$

They satisfy the Beltrami property that $\operatorname{curl} u = k\, u$, $k > 0$. A simple calculation shows that

$$A(\tilde{U}) = A(U^N) - \delta_1 \begin{pmatrix} 1 & & \\ & 1 & \\ & & 0 \end{pmatrix} - \delta_2 \begin{pmatrix} 1 & & \\ & 0 & \\ & & 1 \end{pmatrix} - \delta_3 \begin{pmatrix} 0 & & \\ & 1 & \\ & & 1 \end{pmatrix}.$$

Thus, for appropriate choices of δ_1, δ_2, δ_3, one has

$$A(\tilde{U}) \in \Omega, \quad d_{\mathcal{P}}\left(U^n, \tilde{U}\right) < \varepsilon/2,$$

so that $d_{\mathcal{P}}(U, \tilde{U}) < \varepsilon$. Thus, Ω is dense, which ends the proof of the lemma.

We can now perform a spectral analysis of (3.3).

Proposition 3.2. *Let Ω be the subset of $\mathcal{P}$ defined in Lemma* 3.1. *For all U in Ω, there exist $\zeta^0 \in \mathbb{R}^3$, $\delta > 0$, and two smooth functions*

$$\Lambda_+ : B(\zeta^0, \delta) \mapsto \mathbb{R}_*^+, \quad \hat{b} : B(\zeta^0, \delta) \mapsto \left(\mathbb{C}_*\right)^3,$$

such that for all $\zeta \in B(\zeta^0, \delta)$,

$$b^\zeta(t, \mathrm{x}) \; = \; \hat{b}(\zeta) \, \exp\left(\Lambda_+(\zeta)\, t\right) \, \exp\left(i\zeta \cdot \mathrm{x}\right)$$

is a divergence-free solution of (3.3). *Moreover, one can assume that Λ has a non-degenerate maximum over $B(\zeta^0, \delta)$ at ζ^0.*

Proof. Let U in Ω. We apply the Fourier transform to (3.3): we get, for all $\zeta \in \mathbb{R}^3$,

$$\partial_t \mathcal{F}\left(b^0\right)(t,\zeta) = A^\zeta A\,\mathcal{F}\left(b^0\right)(t,\zeta) - |\zeta|^2 \mathcal{F}\left(b^0\right)(t,\zeta),$$

where,

$$\forall \zeta \in \mathbb{R}^3, \quad A^\zeta = \begin{pmatrix} 0 & -i\zeta_3 & i\zeta_2 \\ i\zeta_3 & 0 & -i\zeta_1 \\ -i\zeta_2 & i\zeta_1 & 0 \end{pmatrix}$$

is the matrix corresponding to cross-product by $i\zeta$. As A is real-symmetric, there exists an orthogonal matrix P with $P^t A\, P = \operatorname{diag}(\alpha_1, \alpha_2, \alpha_3)$, $\alpha_i \in \mathbb{R}$. Introducing $\xi = P^t \zeta$, $\tilde{b}(t,\xi) = P^t \mathcal{F}(b^0)(t,\zeta)$, the previous equation reads

$$\partial_t\, \tilde{b}(t,\xi) = A^\xi D\, \tilde{b}(t,\xi) - |\xi|^2\, \tilde{b}(t,\xi). \tag{3.6}$$

A rapid calculation shows that the eigenvalues λ of $A^\xi D - |\xi|^2 I_d$ satisfy $\lambda = -|\xi|^2$ or

$$\left(\lambda + |\xi|^2\right)^2 \;=\; \xi_1^2 \alpha_2\alpha_3 \;+\; \xi_2^2\alpha_1\alpha_3 \;+\; \xi_3^2 \alpha_1\alpha_2. \tag{3.7}$$

As $U \in \Omega$, the α_i's are distinct and nonzero. Consequently, the products $\alpha_i\alpha_j$, $i \neq j$, are also distinct and nonzero. Let

$$f(\xi) := \xi_1^2\alpha_2\alpha_3 \;+\; \xi_2^2\alpha_1\alpha_3 \;+\; \xi_3^2\alpha_1\alpha_2, \quad \mathcal{U} := \{\xi,\ f(\xi) > 0\}.$$

There are two possibilities:

- All the α_i's have the same sign. Then all the $\alpha_i\alpha_j$, $i \neq j$, are positive, and $\mathcal{U}$ is $\mathbb{R}^3_*$.
- The α_i's have different signs. Then among the products $\alpha_i\alpha_j$, $i \neq j$, two are negative and one is positive—for instance, $\alpha_2\alpha_3$. In this case, $\mathcal{U}$ is the cone $\{|\xi_1|^2 > -\frac{\alpha_1\alpha_3}{\alpha_2\alpha_3}|\xi_2|^2 - \frac{\alpha_1\alpha_2}{\alpha_2\alpha_3}|\xi_3|^2\}$.

On $\mathcal{U}$, one can define

$$\lambda_\pm(\xi) = \pm f(\xi)^{1/2} \;-\; |\xi|^2.$$

Note that $\lambda_+(\xi)$ takes positive values for some ξ. Indeed, if $\bar{\xi}$ is such that $f(\bar{\xi}) > 0$, then

$$\lambda_+\left(\delta\bar{\xi}\right) = \delta f\left(\bar{\xi}\right)^{1/2} \;-\; \delta^2\left|\bar{\xi}\right|^2$$

is positive for $\delta > 0$ small enough. Moreover, using $\lambda_+(\xi) \xrightarrow{|\xi|\to+\infty} -\infty$, we deduce that λ_+ has a global positive maximum in $\mathcal{U}$, say at $\xi = \xi^0$. As ξ^0 is a critical point, we obtain

$$\begin{aligned} 0 = \lambda_+'(\xi^0) &= \frac{1}{2} f(\xi^0)^{-1/2}\nabla f(\xi^0) \;-\; 2\xi^0 \\ &= f(\xi^0)^{-1/2}\begin{pmatrix} \alpha_2\alpha_3 & & \\ & \alpha_1\alpha_3 & \\ & & \alpha_1\alpha_2 \end{pmatrix}\xi^0 - 2\xi^0. \end{aligned}$$

Up to reindex the eigenvalues, this implies that

$$2 f(\xi^0)^{1/2} = \alpha_2\alpha_3, \quad \xi^0 = (\xi_1^0, 0, 0)^t. \tag{3.8}$$

Then we compute

$$\lambda_+''(\xi^0) = -\frac{1}{2} f(\xi^0)^{-3/2} \nabla f(\xi^0) \otimes \nabla f(\xi^0) + \frac{1}{2} f(\xi^0)^{-1/2} f''(\xi^0) - 2I_d,$$

and using (3.8) leads to

$$\lambda_+''(\xi^0) = \begin{pmatrix} -\frac{1}{2f(\xi^0)^{3/2}}(\xi_1^0)^2 & & \\ & \frac{\alpha_1\alpha_3}{f(\xi^0)^{1/2}} - 2 & \\ & & \frac{\alpha_1\alpha_2}{f(\xi^0)^{1/2}} - 2 \end{pmatrix}.$$

In particular, $\det(\lambda_+''(\xi^0)) \neq 0$, which means that the maximum at ξ^0 is nondegenerate.

Let $\delta > 0$ such that $B(\xi^0, \delta) \subset \mathcal{U}$. For all $\xi \in B(\xi^0, \delta)$, $\lambda_+(\xi)$ has multiplicity 1 as an eigenvalue of $A^\xi D$. Therefore, classical smooth dependence results on eigenvalues and eigenvectors yield the existence of a smooth function $\tilde{b} : B(\xi^0, \delta) \mapsto (\mathbb{C}_*)^3$ such that

$$A^\xi D\, \tilde{b}(\xi) = \lambda_+(\xi)\, \tilde{b}(\xi).$$

Back to original variables, we define

$$\zeta^0 = P\xi^0, \quad \Lambda_\pm(\zeta) = \lambda_\pm(P^t\zeta), \quad \hat{b}(\zeta) = P\tilde{b}(P^t\zeta),$$

so that

$$\left(A^\xi A - |\zeta|^2\right) \hat{b}(\zeta) = \Lambda_+(\zeta)\, \hat{b}(\zeta).$$

This shows that, for all $\zeta \in B(\zeta^0)$,

$$b^\zeta(t, \mathrm{x}) \;=\; \hat{b}(\zeta)\, e^{i\Lambda(\zeta)t}\, e^{i\zeta\cdot \mathrm{x}}$$

solves (3.3). Note also that b^ζ is divergence-free, as $\zeta \cdot \hat{b}(\zeta) = 0$. This ends the proof of the proposition.

3.2. Construction of unstable wavepackets. Thanks to Proposition 3.2, we are now able to build approximate solutions having exponential growth. More precisely, we show the following.

PROPOSITION 3.3. *Let $U \in \Omega$. There exists $\lambda^0 > 0$, and for every integer $n \in \mathbb{N}$ families $\{X^i = (\overline{V}^i, V_*^{i+1}, p_*^i, \overline{p}^{i+1})\}_{0\le i\le n}$ such that:*

(i) *For all i, X^i satisfies (T_i).*

(ii) *As $t \to +\infty$, $V^0 = \overline{V}^0 = (0, \overline{\beta}^0)$ has the asymptotic behavior*

$$\|\overline{\beta}^0(t,\cdot)\|_{L^2} \;\sim\; \frac{C}{\sqrt{t}} e^{\lambda^0 t}. \tag{3.9}$$

(iii) *For all $i = km + l$, with $l \in \{0, \dots, m-1\}$, for all $\alpha, s \in \mathbb{N}$, and for all t,*

$$\|\partial_t^\alpha X^i(t,\cdot)\|_{H^s} \le \frac{C_{\alpha,i,s}}{\sqrt{1+t}^{k+1}}\, t^l\, e^{(k+1)\lambda^0 t}, \quad C_{\alpha,i,s} > 0. \tag{3.10}$$

Proof. We treat separately the cases $i = 0$, $1 \le i \le m-1$, and $i \ge m$ (for which (T_i) includes quadratic terms).

Construction of X^0. We first use Proposition 3.2 to build X^0. Let

$$\Lambda_\pm : B(\zeta^0,\delta) \mapsto \mathbb{R}_*^+, \quad \hat{b} : B(\zeta^0,\delta) \mapsto (\mathbb{C}_*)^3$$

as in Proposition 3.3. As $\Lambda^0 := \Lambda_+(\zeta^0)$ is a nondegenerate maximum, one can assume, up to take a smaller δ, that

$$\Lambda_+(\zeta) = \Lambda_+(\zeta^0) + \nabla_\zeta \Lambda_+(\zeta^0)\cdot(\zeta-\zeta^0) - |\zeta-\zeta^0|^2 \frac{\alpha(\zeta-\zeta^0)}{2}, \tag{3.11}$$

where $0 < \underline{\alpha} < \alpha(\cdot) < \overline{\alpha}$ in $B(0,\delta)$. We extend functions $\Lambda_\pm$ and $\hat{b}$ to

$$B_\delta := B(\zeta^0,\delta) \cup B(-\zeta^0,\delta)$$

by the following: for all $\zeta \in B(-\zeta^0,\delta)$,

$$\Lambda_\pm(\zeta) := \Lambda_\pm(-\zeta), \quad \hat{b}(\zeta) := \hat{b}(-\zeta)^*.$$

With this continuation,

$$b^\zeta(t,\mathrm{x}) = \hat{b}(\zeta)\, e^{\Lambda_+(\zeta)t}\, e^{i\zeta\cdot\mathrm{x}} \tag{3.12}$$

is a divergence-free solution of (3.3) for all ζ in B_δ. Now let ϕ be a smooth real-valued function supported in B_δ, such that $\phi(\zeta^0) = 1$, $\phi(-\zeta) = \phi(\zeta)^*$. We set

$$\overline{\beta}^0(t,\mathrm{x}) = \int_{B_\delta} \phi(\zeta)\,\hat{b}(\zeta)\, e^{\Lambda_+(\zeta)t}\, e^{i\zeta\cdot\mathrm{x}}\, d\zeta. \tag{3.13}$$

Then we define $\overline{V}^0 := (0,\overline{\beta}^0)^t$, V_*^1 by (2.6), and $X^0 = (\overline{V}^0, V_*^1, 0, 0)$. From the properties of b^ζ, we deduce easily that $\overline{V}^0$ satisfies (3.2), and consequently that X^0 solves (T_0). Finally, points (ii) and (iii) of the proposition are derived from standard computations involving (3.11). For the sake of brevity, we do not detail these computations and refer to [4] for complete treatment in a very similar framework.

Construction of X^i, $1 \le i \le m-1$. Let X^0 be defined as above, and for $1 \le i \le m-1$, let X^i be defined inductively by the following: *X^i is the solution of (T_i) with $\overline{V}^i|_{t=0} = 0$.* As seen in section 3.1, such a definition makes sense. We show by induction on $i \ge 0$ the following property:

(P_i): *Function X^i has an expression of the type*

$$\begin{aligned} X^i(t,\mathrm{x},\tau,\theta) &\\ = \int_{B_\delta} &\left(P^{i,+}_{\zeta,\tau,\theta}(t)\, e^{\Lambda_+(\zeta)t} + P^{i,-}_{\zeta,\tau,\theta}(t)\, e^{\Lambda_-(\zeta)t} + P^{i,-}_{\zeta,\tau,\theta}(t)\, e^{-|\zeta|^2 t}\right) e^{i\zeta\cdot\mathrm{x}} d\zeta, \end{aligned} \tag{3.14}$$

where $P^{i,\pm}_{\zeta,\tau,\theta}$, $P^{i,0}_{\zeta,\tau,\theta}$ are polynomials in t, of degree $\le i$, with coefficients smooth and compactly supported in $B_\delta \times \mathbb{T} \times \mathbb{T}^3$.

- *Case* $i = 0$. (P_0) is true by definition of X^0.
- *Case* $i \ge 1$. Let $i \ge 1$, and assume $(P_0), \dots, (P_{i-1})$. Recall that

$$\begin{aligned} \left(p_*^i, 0\right)^t &= \Delta_\theta^{-1} \mathrm{Div}_\theta B_\theta\left(\tilde{U}, \overline{V}^i\right) + H_*^i, \\ V_*^{i+1} &= (\partial_\tau - \Delta)^{-1} L_\theta \overline{V}^i + I_*^i, \end{aligned}$$

where H_*^i, I_*^i are i-linear in $(X^0,\dots,X^{i-1})$. Thus, X^i is of type (3.14) as soon as $\overline{V}^i$ and $\overline{p}^{i+1}$ are.

These functions satisfy

$$\begin{cases}\partial_t \overline{V}^i = \begin{pmatrix}-\mathrm{curl}_{\mathrm{x}} A & \\ & \mathrm{curl}_{\mathrm{x}} A\end{pmatrix} V^i + \Delta_{\mathrm{x}} V^i + \left(\nabla_{\mathrm{x}} \overline{p^{i+1}}, 0\right)^t + \overline{J}^i, \\ \mathrm{Div}_{\mathrm{x}} V^i = 0. \end{cases} \tag{3.15}$$

Taking the curl of the first line of (3.15) yields

$$\overline{p}^{i+1} = -(\Delta_{\mathrm{x}})^{-1}\, \overline{J}^i_1,$$

which shows that

$$\overline{p}^{i+1}(t,\mathrm{x}) = \int_{B_\delta} \left(Q^{i,+}_\zeta(t)\, e^{\Lambda_+(\zeta)t} + Q^{i,-}_\zeta(t)\, e^{\Lambda_-(\zeta)t} + Q^{i,0}_\zeta(t)\, e^{-|\zeta|^2 t}\right) e^{i\zeta\cdot x} d\zeta,$$

where $Q^{i,\pm}_\zeta$, $Q^{i,0}_\zeta$ are polynomials in t, of degree $\le i-1$, with coefficients smooth and compactly supported in B_δ.

Replacing $\overline{p}^{i+1}$ by its expression leads to

$$\begin{cases}\partial_t \overline{V}^i = \begin{pmatrix}-\mathrm{curl}_{\mathrm{x}} A & \\ & \mathrm{curl}_{\mathrm{x}} A\end{pmatrix} \overline{V}^i + \Delta_{\mathrm{x}} \overline{V}^i + \overline{K}^i, \\ \mathrm{Div}_{\mathrm{x}} \overline{V}^i = 0, \end{cases}$$

where $\overline{K}^i$ is i-linear on $(X^0,\dots,X^{i-1})$. We define

$$\mathcal{V}^i(t,\zeta) := \mathcal{F}\left(\overline{V}^i\right)(t,\zeta),$$

which satisfies

$$\partial_t \mathcal{V}^i(t,\zeta) = \begin{pmatrix}-A^\zeta A & \\ & A^\zeta A\end{pmatrix} - |\xi|^2 \mathcal{V}^i(t,\zeta) + \mathcal{K}^i(t,\zeta),$$

$$\mathcal{K}^i(t,\zeta) = R^{i,+}_\zeta(t)\, e^{\Lambda_+(\zeta)t} + R^{i,-}_\zeta(t)\, e^{\Lambda_-(\zeta)t} + R^{i,0}_\zeta(t)\, e^{-|\zeta|^2 t},$$

where $R^{i,\pm}_\zeta$, $R^{i,0}_\zeta$ are polynomials in t, of degree $\le i-1$, smooth and compactly supported in B_δ.

At fixed ζ, such an equation is an ordinary differential system of the type

$$\frac{d}{dt}\mathcal{V} + M\mathcal{V} = S^+(t)e^{\Lambda_+ t} + S^-(t)e^{\Lambda_- t} + S^0(t)e^{\Lambda^0 t},$$

where Λ_+, Λ_-, Λ_0, which stand for $\Lambda_+(\zeta)$, $\Lambda_-(\zeta)$, $-|\zeta|^2$, are simple eigenvalues of the matrix M, which stands for $\left(\begin{smallmatrix}-A^\zeta A & \\ & A^\zeta A\end{smallmatrix}\right)$. It is well known that the solution

$$\mathcal{V}(t) = \int_0^t e^{M(t-s)} \left(S^+(s)e^{\Lambda^+ s} + S^-(s)e^{\Lambda^- s} + S^0(s)e^{\Lambda^0 s}\right) ds$$

with $\mathcal{V}(0) = 0$ is of the type

$$\mathcal{V}(t) = T^+(t)e^{\Lambda^+ t} + T^-(t)e^{\Lambda^- t} + T^0(t)e^{\Lambda^0 t},$$

where $T^\pm$ (resp., T^0) is a polynomial such that $\deg T^\pm \leq \deg S^\pm + 1$ (resp., $\deg T^0 \leq \deg S^0 + 1$).

Back to the original system, (P_i) follows, which ends the induction. Point (iii) is again a classical consequence of expression (3.14) and (3.11).

Construction of X^i, $i \geq m$. As $i \geq m$, quadratic terms enter system (T_i). More precisely, one checks that

- H^i_* and I^i_* are made of terms that are i-linear in $(X^0, \dots, X^{i-1})$, and of quadratic terms that involve the pairs $\{V^j, V^J\}$ with $j + J = i - m - 1$, or $i - m - 3$;
- $\overline{J}^i$ and $\overline{K}^i$ are made of terms that are i-linear in $(X^0, \dots, X^{i-1})$, and of quadratic terms involving the pairs $\{V^j, V^J\}$ with $j + J = i - m$, $i - m - 1$ or $i - m - 3$.

Note that

$$\Lambda_+(j\zeta^0) \,<\, j\,\Lambda_+(\zeta^0) \quad \forall 2 \leq j \leq n,$$

so that up to take δ smaller, one can assume that for all $2 \leq j \leq n$ and for all $\zeta_1, \dots, \zeta_j$ in B_δ,

$$\Lambda_+(\zeta_1 + \cdots + \zeta_j) \,<\, \Lambda_+(\zeta_1) \,+ \cdots +\, \Lambda_+(\zeta_j).$$

Under this assumption, one can show inductively that for general $i = km + l$, $k \geq 0$, $1 \leq l \leq m - 1$, the solution X^i of (T_i) with $\overline{V}^i|_{t=0} = 0$ has an expression of the type

$$X^i(t, \mathrm{x}, \tau, \theta) = \sum_{j=1}^{k+1} \int_{B_\delta^{j+1}} Y^{i,j}(\zeta_1, \dots, \zeta_j, \tau, \theta, t)\, e^{i(\zeta_1 + \cdots + \zeta_j)\cdot \mathrm{x}}\, d\zeta_1 \dots d\zeta_j$$
$$+ \int_{B_\delta^{k+1}} P^i(\zeta_1, \dots, \zeta_{k+1}, \tau, \theta, t)\, e^{(\Lambda_+(\zeta_1) + \cdots + \Lambda_+(\zeta_{k+1}))\, t}\, e^{i(\zeta_1 + \cdots + \zeta_{k+1})\cdot \mathrm{x}}\, d\zeta_1 \dots d\zeta_{k+1},$$

where

- P^i is a polynomial in t, of degree $\leq l$, with coefficients smooth and compactly supported.
- $Y^{i,j}$ is a finite sum of terms of the form

  $$Y^i_{\Lambda_1, \dots, \Lambda_j}(\zeta_1, \dots, \zeta_j, \tau, \theta, t) = Q^i_{\Lambda_1, \dots, \Lambda_j}(\zeta_1, \dots, \zeta_j, \tau, \theta, t)\, e^{(\Lambda_1(\zeta_1) + \cdots + \Lambda_j(\zeta_{k+1}))\, t},$$

  with $Q^i_{\Lambda_1, \dots, \Lambda_j}$ polynomial in t, and $\Lambda_1(\zeta_0) + \cdots + \Lambda_j(\zeta_0) < (k+1)\Lambda_+(\zeta^0)$.

We do not detail this induction, as it is very similar to the previous one and tedious. Once this expression for X^i is obtained, the estimate follows; see [4] for details.

3.3. Proof of Theorem 1.2. We now turn to the proof of the instability theorem, Theorem 1.2. We adapt ideas of [9], encountered in the stability study of Euler and Prandtl equations (see also [6, 12]).

Let k_0 in $\mathbb{N}^*$ to be chosen later, let $m \geq 1$, and let $n = k_0 m$. Let $U \in \Omega$, and take profiles $X^0, \dots, X^n$ as in Proposition 3.3. We have in particular

$$\varepsilon^m \|\overline{V}^0(t, \cdot)\|_{L^2} = \varepsilon^m \|\overline{\beta}^0(t, \cdot)\|_{L^2} \geq \frac{C_0\, \varepsilon^m\, e^{\lambda^0 t}}{(1+t)^{1/2}}.$$

We define

$$\mathcal{E}(t) := \frac{C_0\, \varepsilon^m\, e^{\lambda^0 t}}{(1+t)^{1/2}}$$

and $T^\varepsilon > 0$ such that $\mathcal{E}(T^\varepsilon) = 1$. Note that $T^\varepsilon = O(|\ln(\varepsilon)|)$. Moreover, thanks to point (iii) of Proposition (3.3), we get, for all $i = km + l$, $l \in \{0, \dots, m-1\}$, and for all α, s, t,

$$\begin{aligned}
\varepsilon^{m+i} \|\partial_t^\alpha X^i(t)\|_{H^s} &\le C\, \frac{\varepsilon^{m+i} e^{(k+1)\lambda_0 t}}{\sqrt{1+t}^{k+1}}\, t^l, \\
&\le C\, \left(\frac{\varepsilon^m e^{\lambda_0 t}}{\sqrt{1+t}}\right)^{k+1}\, (\varepsilon t)^l, \\
&\le C'\, \mathcal{E}(t)^{k+1}\, (\varepsilon t)^l.
\end{aligned} \tag{3.16}$$

As in section 2.3, we define $V_{app}^{\varepsilon,n}$ and $p_{app}^{\varepsilon,n}$ by (2.3), so that they satisfy equations of type (2.14). Thanks to (3.16), it is easy to check that

$$\|\overline{R}_{app}^{\varepsilon,n}(t)\|_{H^s} \le C_s\, \mathcal{E}(t)^{k_0 - 2},$$

$$\|R_{app,*}^{\varepsilon,n}(t)\|_{H^s} \le C_s\, \varepsilon^{-2}\, \mathcal{E}(t)^{k_0-2},$$

$$\|\partial_t^\alpha r_{app}^{\varepsilon,n}(t,\cdot)\|_{H^s} \le C'_{s,\alpha}\, \left(\varepsilon^m\, \mathcal{E}(t)^{k_0+1} \;+\; \varepsilon\, \mathcal{E}(t)^{k_0+1}\right).$$

The Fourier transform of V^n and V^{n+1} has compact support in ζ, so that $r_{app}^{\varepsilon,n} = \varepsilon^{m+n}(V_*^n + \varepsilon V_*^{n+1})$ is such that $r_{app,l}^{\varepsilon,n} = 0$. As in section 2.3, we can then define W_{app}^ε by (2.24), (2.25) and set $V_{app}^\varepsilon = V_{app}^{\varepsilon,n} + W_{app}^\varepsilon$. This solves (2.27), where the remainder R_{app}^ε satisfies, for all $t \in [0, T^\varepsilon]$,

$$\|\overline{R}_{app}^{\varepsilon}(t)\|_{H^s} \;+\; \varepsilon^2 \|R_{app,*}^\varepsilon(t)\|_{H^s} \;\le C_s\, \mathcal{E}(t)^{k_0-2}. \tag{3.17}$$

We deduce (as in (2.26))

$$\begin{aligned}
\|R_{app,l}^\varepsilon(t)\|_{H^s} &\le \|\overline{R}_{app,l}^{\varepsilon}(t)\|_{H^s} \;+\; \|R_{app,*,l}^\varepsilon(t)\|_{H^s} \\
&\le \|\overline{R}_{app}^{\varepsilon}(t)\|_{H^s} \;+\; C\,\varepsilon^2\, \|R_{app}^\varepsilon(t)\|_{H^{s+1/2}} \\
&\le C'_s\, \mathcal{E}(t)^{k_0-2}.
\end{aligned} \tag{3.18}$$

Conclusion. We fix $s \ge 5$, $\eta > 0$ to be chosen later. With the notation of Proposition 2.1, (2.15), we choose $k_0 > \max(4, 2 + \lambda^{-1}\mathcal{C}_s)$. Let V^ε be the solution of (2.2), with initial data $V^\varepsilon|_{t=0} = V_{app}^\varepsilon|_{t=0}$. Let $W^\varepsilon = V^\varepsilon - V_{app}^\varepsilon$, and let

$$T(\varepsilon) = \sup\left\{t \in [0, T^\varepsilon], \quad \forall u \in [0,t],\ \alpha_s(W^\varepsilon; u) \le 1/2\right\}.$$

We apply Proposition 2.1 to W^ε. This yields, for all $t \in [0, T(\varepsilon))$,

$$\alpha_s(W^\varepsilon; t) \le 2\mathcal{C}_s \int_0^t \alpha_s(W^\varepsilon; u)du + \frac{4}{3}\mathcal{C}_s \left(\varepsilon^6 \int_0^t \|R_{app,h}^\varepsilon(u)\|_{H^s}^2 du \right.$$
$$\left. + \int_0^t \|R_{app,l}^\varepsilon(u)\|_{H^s}^2 du\right).$$

We deduce from (3.18) that

$$\alpha_s(W^\varepsilon;t) \leq 2\mathcal{C}_s \int_0^t \alpha_s(W^\varepsilon;u)du \,+\, C_s\,\mathcal{E}(t)^{2(k_0-2)}.$$

With our choice for k_0, Gronwall's lemma implies

$$\alpha_s(W^\varepsilon;t) \leq\, C_s\,\mathcal{E}(t)^{2(k_0-2)}.$$

For $t^\varepsilon = T^\varepsilon - \sigma$, σ independent of ε, large enough,

$$C_s\,\mathcal{E}(t)^{2(k_0-2)} \,<\, e^{-2(k_0-2)\lambda^0\sigma} < \frac{1}{4}.$$

This shows that $T(\varepsilon) \geq t^\varepsilon$ and that

$$\|W_l^\varepsilon(t^\varepsilon)\|_{H^s}^2 + \varepsilon^2\|W^\varepsilon(t^\varepsilon)\|_{H^s}^2 < \exp(-2\lambda^0\sigma) \quad \forall s.$$

Recall also that, up to consider a larger σ,

$$\|V_{app}^\varepsilon - \overline{V}^0\|_{H^s}^2 \,\leq\, e^{-2\lambda^0\sigma}.$$

We can now conclude the proof of Theorem 1.2. We introduce

$$\begin{aligned}(v^\varepsilon, b^\varepsilon)^t(t,\mathrm{x}) &= V^\varepsilon\left(t,\mathrm{x},\varepsilon^{-4}t,\varepsilon^{-2}\mathrm{x}\right)\\ &= (w^\varepsilon,\beta^\varepsilon)\left(t,\mathrm{x},\varepsilon^{-4}t,\varepsilon^{-2}\mathrm{x}\right),\end{aligned}$$

which is solution of the original system (1.5). Rapid computations lead to

$$\begin{aligned}\|b^\varepsilon(t)\|_{L^2_\mathrm{x}}^2 &\geq \|\beta^\varepsilon(t)\|_{L^2_{\mathrm{x},\tau,\theta}}^2 \,-\, C\,\varepsilon^2\|\beta^\varepsilon(t)\|_{H^1_{\mathrm{x},\tau,\theta}}^2\\ &\geq \|\overline{\beta}_l^\varepsilon(t)\|_{L^2_\mathrm{x}}^2 \,-C\,\varepsilon^2\|\beta^\varepsilon(t)\|_{H^1_{\mathrm{x},\tau,\theta}}^2\\ &\geq \frac{\varepsilon^m}{2}\|\overline{\beta}_l^0(t)\|_{L^2_\mathrm{x}}^2 \,-\, C'\Big(\|\overline{V}_{app,l}^\varepsilon(t) - \varepsilon^m\overline{V}_{app,l}^0(t)\|_{L^2_\mathrm{x}}^2 + \|\overline{W}_l^\varepsilon(t)\|_{L^2_\mathrm{x}}^2\\ &\quad + \varepsilon^2\left(\varepsilon^m\|\overline{\beta}^0(t)\|_{H^1_\mathrm{x}}^2 + \|V_{app}^\varepsilon(t) - \varepsilon^m\overline{V}_{app}^0(t)\|_{H^1_{\mathrm{x},\tau,\theta}}^2 + \|W^\varepsilon(t)\|_{H^1_{\mathrm{x},\tau,\theta}}^2\right)\Big),\end{aligned}$$

where C and C' are positive constants independent of ε and η. As the Fourier transform of $\overline{\beta}^0$ has compact support, we deduce that

$$\overline{\beta}^0 = \overline{\beta}_l^0, \quad \|\overline{\beta}^0\|_{H^1_\mathrm{x}} \,\leq\, R\,\|\overline{\beta}^0\|_{L^2_\mathrm{x}}$$

for some $R > 0$.

Using previous bounds, this yields, for ε small enough,

$$\|b^\varepsilon(t)\|_{L^2}^2 \,\geq\, C_0\,e^{-\lambda^0\sigma} \,-\, C_1\,e^{-2\lambda^0\sigma} \,\geq\, \delta > 0,$$

up to consider a larger σ. Theorem 1.2 follows.

Acknowledgments. I wish to express my gratitude to B. Desjardins and E. Dormy, for careful reading of the manuscript and helpful comments.

REFERENCES

[1] V. Arnol'd, *Sur la topologie des écoulements stationnaires des fluides parfaits*, C. R. Acad. Sci. Paris, 261 (1965), pp. 17–20.

[2] S. Braginsky, *Theory of the hydromagnetic dynamo*, Soviet Phys. JETP, 20 (1964), pp. 1462–1471.

[3] S. Childress, *New solutions of the kinematic dynamo problem*, J. Math. Phys., 11 (1970), pp. 3063–3076.

[4] B. Desjardins and E. Grenier, *Linear instability implies nonlinear instability for various types of viscous boundary layers*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 20 (2003), pp. 87–106.

[5] D. Fearn, *Hydromagnetic flows in planetary cores*, Rep. Prog. Phys., 61 (1998), pp. 175–235.

[6] S. Friedlander, W. Strauss, and M. Vishik, *Nonlinear instability in an ideal fluid*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 187–209.

[7] H. Fujita and T. Kato, *On the Navier-Stokes initial value problem.* I, Arch. Rational Mech. Anal., 16 (1964), pp. 269–315.

[8] A. D. Gilbert, *Dynamo theory*, in Handbook of Mathematical Fluid Dynamics, Vol. II, North-Holland, Amsterdam, 2003, pp. 355–441.

[9] E. Grenier, *On the nonlinear instability of Euler and Prandtl equations*, Comm. Pure Appl. Math. 53 (2000), pp. 1067–1091.

[10] J.-L. Joly, G. Métivier, and J. Rauch, *Coherent and focusing multidimensional nonlinear geometric optics*, Ann. Sci. École Norm. Sup. (4), 28 (1995), pp. 51–113.

[11] G. Métivier, *Stability of Small Viscosity Noncharacteristic Boundary Layers*, Lecture notes, available at http://www.math.u-bordeaux.fr/~metivier/Preprints.html.

[12] G. Métivier, *Exemples d'instabilités pour des équations d'ondes non-linéaires (d'après Gilles Lebeau)*, Astérisque, 294 (2004), pp. 63–75.

[13] E. Parker, *Hydromagnetic dynamo models*, Astrophys. J., 122 (1955), pp. 293–314.

[14] G. Roberts, *Spatially periodic dynamos*, Philos. Trans. Roy. Soc. London Ser. A, 266 (1970), pp. 535–558.

[15] P. Roberts, *An Introduction to Magnetohydrodynamics*, Longmans Green, London, 1967.

[16] M. Steenbeck, F. Krause, and K. Rädler, *A calculation of the mean electromotive force in an electrically conducting fluid in turbulent motion, under the influence of Coriolis forces*, Z. Naturforsch., 21 (1966), pp. 369–376.

[17] R. Stieglitz and U. Muller, *Experimental demonstration of a homogeneous two-scale dynamo*, Phys. Fluids, 13 (2001), pp. 561–564.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 841–866

OPTIMAL RATE OF CONVERGENCE OF THE BENCE–MERRIMAN–OSHER ALGORITHM FOR MOTION BY MEAN CURVATURE*

KATSUYUKI ISHII†

Abstract. Bence, Merriman, and Osher proposed an algorithm for computing the motion a hypersurface by mean curvature in terms of solutions of the usual heat equation, continually reinitialized after short time steps. In this paper, applying some techniques of asymptotic analysis for the Allen–Cahn equation, we give a rate of convergence of their algorithm for the motion of a smooth and compact hypersurface by mean curvature. We also consider the special case of a circle evolving by curvature and show that our rate is optimal.

Key words. motion by mean curvature, numerical algorithm, rate of convergence, optimality

AMS subject classifications. 35K05, 35K55, 65M15

DOI. 10.1137/04061862X

1. Introduction.

In 1992, Bence, Merriman, and Osher proposed in [2] an algorithm for computing the motion of a hypersurface by its mean curvature. It is described as follows.

Given a closed set $C_0 \subset \mathbb{R}^N$, we solve the initial-value problem for the heat equation

$$\begin{cases} u_t - \Delta u = 0 & \text{in } (0,+\infty)\times\mathbb{R}^N, \\ u(0,x) = \begin{cases} 1, & x \in C_0, \\ -1, & x \in \mathbb{R}^N \setminus C_0. \end{cases} \end{cases} \tag{1.1}$$

Fix a time step $h > 0$ and set

$$C_1 = \{x \in \mathbb{R}^N \mid u(h,x) \geq 0\}.$$

Next we solve (1.1) with C_0 replacing C_1 and define a new set C_2 with the solution u replaced by that of (1.1) with the new initial data. Repeating this procedure, we have a sequence $\{C_k\}_{k=0,1,\ldots}$ of closed sets in $\mathbb{R}^N$. Then we define

$$C_t^h = C_k \quad \text{if } kh \leq t < (k+1)h,\ k = 0,1,\ldots,$$

for $t \geq 0$. Letting $h \searrow 0$, we obtain

$$\partial C_t^h \longrightarrow \Gamma_t,\ \Gamma_0 = \partial C_0,$$

and Γ_t moves by its $((N-1)$-times) mean curvature.

The convergence of the Bence–Merriman–Osher (BMO) algorithm was proved by Mascarenhas [19], Evans [5], Barles and Georgelin [1], and Goto, Ishii, and Ogawa [10]. The generalizations of this algorithm were considered by Ishii [14], Ishii, Pires,

*Received by the editors November 10, 2004; accepted for publication (in revised form) March 14, 2005; published electronically December 2, 2005. This work was partially supported by JSPS grant-in-aid for scientific research grants 14540117 and 15340051.
http://www.siam.org/journals/sima/37-3/61862.html

†Faculty of Maritime Sciences, Kobe University, Higashinada, Kobe 658-0022, Japan (ishii@maritime.kobe-u.ac.jp).

and Souganidis [16], Ishii and Ishii [15], Vivier [23], and Leoni [18]. However, to the author's knowledge, there are a few results on the rate of convergence of the BMO algorithm. In [22] Ruuth gave an error estimate for the case of the planar graph on $[0,h]$. Ishii and Nakamura [17] proved that the Hausdorff distance between the motion by mean curvature Γ_t and the approximate interface $\Gamma_t^h := \partial C_t^h$ is an order of $h^{1/2}$ as $h \searrow 0$. This estimate is valid before the onset of singularities, but not optimal.

The purpose of this paper is to show the optimal rate of convergence of the BMO algorithm, valid before the onset of singularities, for the Hausdorff distance between Γ_t and Γ_t^h. In fact, assuming $\{\Gamma_t\}_{0\le t<T_0}$ is the motion of a smooth and compact hypersurface by mean curvature, we prove that, for any $T < T_0$,

$$\sup_{t\in[0,T]} d_H(\Gamma_t, \Gamma_t^h) \le Lh,$$

where L is a constant depending on T, but independent of small $h > 0$, and d_H denotes the Hausdorff distance. This estimate is optimal and improves that of [17].

Both of the order in h and the optimality are the consequence of the maximum principle and the explicit constructions of sub- and supersolutions of (1.1), which are inspired by the asymptotic analysis of solutions of the Allen–Cahn equation (see, e.g., Fife [7] and de Mottoni and Schatzman [4]). As for the relation between the BMO algorithm and the Allen–Cahn equation, from the viewpoint of the splitting methods in numerical analysis, Vivier [23] first pointed out that we may think the Allen–Cahn equation is an approximation of the BMO algorithm. Leoni [18] and Goto, Ishii, and Ogawa [10] gave the proofs of the convergence of the BMO algorithm and a generalized scheme by applying some techniques of the asymptotic analysis for the Allen–Cahn equation. The arguments in this paper also rely on them.

This paper is organized as follows. In section 2 we discuss the formal asymptotic expansion of the radially symmetric solution of (1.1). To derive formally the equation of the motion of the interface, we consider the asymptotic behavior as $t \searrow 0$ of the zero of the solution of (1.1). In section 3 we construct sub- and supersolutions of (1.1) by using some functions provided by the formal asymptotic expansion. We treat the nonradial case in subsection 3.2 and the radial case in subsection 3.3. Section 4 is devoted to the rate of convergence of the BMO algorithm to the motion of a smooth and compact hypersurface by mean curvature. The arguments in sections 3–4 are very similar to those in [18], although Leoni considered in her paper a different situation from ours. In section 5, we return to the special case of a circle evolving by curvature. By the radial symmetry, we have only to consider the asymptotic behavior of the radius R_h of the approximate circle as $h \searrow 0$. In subsection 5.1 we obtain the short-time asymptotics of R_h. In subsection 5.2 we formally derive a corrector for R_h. Based on these results, in subsection 5.3 we obtain the behavior of R_h and show the optimality of our estimate obtained in section 4. The considerations of section 5 are motivated by Nochetto, Paolini, and Verdi [20]. In [20] they obtained the optimal error estimate of the approximate interface given by the solution of a variational inequality to the smooth motion by mean curvature. The appendix is devoted to the proof of a lemma given in section 2.

In the following of this paper, we denote by K various constants depending only on known ones, and the notation $g = O(f)$ means that $|g| \le K'f$ for some constant $K' > 0$ independent of small $t > 0$.

2. Formal asymptotic expansion to the radial case. In this section we briefly discuss the formal asymptotic expansion to the simplest situation, the radially symmetric solution of (1.1) as $t \searrow 0$. Even though this presentation is only formal, it shows us several crucial aspects for the constructions of sub- and supersolutions of (1.1) in the next section. For each $x_0 \in \mathbb{R}^N$, put $B(x_0, R) = \{x \in \mathbb{R}^N \mid |x - x_0| < R\}$.

If $u = u(t, r)$ $(r = |x|)$ and $C_0 = \overline{B(0, R)}$, then the problem (1.1) turns to

$$\text{(2.1)} \qquad \begin{cases} \mathcal{L}u := u_t - u_{rr} - \dfrac{N-1}{r} u_r = 0 & \text{in } (0, +\infty) \times (0, +\infty), \\ u_r(t, 0) = 0, \quad t > 0, \\ u(0, r) = \begin{cases} 1, & r \in [0, R], \\ -1, & r \in (R, +\infty). \end{cases} \end{cases}$$

Set $\widetilde{\Gamma}_0 = \partial B(0, R)$ and $\widetilde{\Gamma}_t = \{x \in \mathbb{R}^N \mid u(t, |x|) = 0\}$ for $t > 0$. Then we can easily verify that $\widetilde{\Gamma}_t$ is a sphere in $\mathbb{R}^N$.

As to the behavior of the solution of (2.1) away from $\widetilde{\Gamma}_t$, we have the following.

LEMMA 2.1. *For any $\delta \in (0, R/2)$, there exist $M_0 > 0$ and $t_0 \in (0, 1)$ such that, for all $t \in (0, t_0)$,*

$$\begin{aligned} &1 - M_0 e^{-\delta^2/32t} \le u(t, r) \le 1 \quad \textit{for all } 0 \le r \le R - \delta, \\ &-1 \le u(t, r) \le -1 + M_0 e^{-\delta^2/32t} \quad \textit{for all } r \ge R + \delta. \end{aligned}$$

See Goto, Ishii, and Ogawa [10, Proposition 6.1] for the proof. From this lemma, it is sufficient for us to consider the asymptotics of the solution of (2.1) near $\widetilde{\Gamma}_t$. Fix $\delta \in (0, R/2)$ and take $t_0 > 0$ so small that Lemma 2.1 holds. Let $\widetilde{\phi}(t)$ be the radius of $\widetilde{\Gamma}_t$. We assume that the solution u of (2.1) is approximated by the following formal series in $(t, r) \in (0, t_0) \times (R - \delta, R + \delta)$:

$$\text{(2.2)} \quad u(t, r) = \sum_{j=0}^{+\infty} t^{j/2} U_j \left(t, \frac{\widetilde{d}(t, r)}{2\sqrt{t}} \right), \qquad \widetilde{\phi}(t) = \sum_{j=0}^{+\infty} t^j \phi_j(t), \qquad \widetilde{\phi}(0) = \phi_0(0) = R.$$

Here U_j and ϕ_j are assumed to be bounded in $(0, t_0) \times (R - \delta, R + \delta)$ for each $j \in \mathbb{N} \cup \{0\}$ and $\widetilde{d}(t, r)$ is the signed distance function to $\widetilde{\Gamma}_t$ given by

$$\widetilde{d}(t, r) = \widetilde{\phi}(t) - r.$$

Before choosing U_j and ϕ_j $(j \in \mathbb{N} \cup \{0\})$, we give a lemma on the effect of the diffusion on $\partial B(0, R)$. Select $\widetilde{r} \ge 0$ so that $u(t, R - \widetilde{r}) = 0$. Then $\widetilde{r} = \widetilde{\phi}(0) - \widetilde{\phi}(t)$ and it is the normal distance between $\partial B(0, R)$ and $\widetilde{\Gamma}_t$.

LEMMA 2.2. *We have*

$$\widetilde{r} = \frac{(N-1)t}{R} + \frac{(N-1)(3N-1)t^2}{6R^3} + O(t^3) \quad \textit{as } t \searrow 0.$$

See the appendix for the proof. By this lemma, we formally have

$$\widetilde{\phi}' \Big|_{t=0} = - \widetilde{r}' |_{t=0} = - \frac{N-1}{R},$$

where $' = d/dt$. This suggests that we may set $\phi_0(t) = \phi(t)$, where $\phi(t) = \sqrt{R^2 - 2(N-1)t}$ and it solves

$$\text{(2.3)} \qquad \phi'(t) = - \frac{N-1}{\phi(t)} \quad \text{for } t > 0, \ \phi(0) = R.$$

Using Lemma 2.2 and the definition of ϕ, we can estimate the distance between $\partial B(0,\phi(t))$ and $\widetilde{\Gamma}_t$ for small $t>0$.

PROPOSITION 2.3. *Let $\phi(t)=\sqrt{R^2-2(N-1)t}$ and let $\widetilde{r}$ be the normal distance between $\partial B(0,R)$ and $\widetilde{\Gamma}_t$. Then*

$$\widetilde{r}-(R-\phi(t))=\frac{(N-1)t^2}{3R^3}+O(t^3) \quad \textit{as } t\searrow 0.$$

Remark 2.1. Ruuth [22, Chapter 4] obtained a similar result to this proposition in the case of the graph in $\mathbb{R}^2$.

Proof of Proposition 2.3. It is easily seen by Taylor expansion to ϕ around $t=0$ that

$$\phi(t)=R-\frac{(N-1)t}{R}-\frac{(N-1)^2t^2}{2R^2}+O(t^3) \quad \text{as } t\searrow 0.$$

Hence, we have the result by using Lemma 2.2 and this expansion. ∎

We choose U_j and ϕ_j ($j\in\mathbb{N}\cup\{0\}$) of (2.2). First, we do some ϕ_j's. It is seen that, as $t\searrow 0$,

$$\begin{aligned}
&\widetilde{r}-(R-\phi(t))=-\sum_{j=1}^{+\infty}t^j\phi_j(t),\\
&\frac{1}{R}\approx\frac{1}{\widetilde{\phi}(t)}=\frac{1}{\phi(t)}\left\{1-\sum_{j=1}^{+\infty}t^j\frac{\phi_j(t)}{\phi(t)}+\left(\sum_{j=1}^{+\infty}t^j\frac{\phi_j(t)}{\phi(t)}\right)^2-\cdots\right\}.
\end{aligned}$$

It follows from these relations and Proposition 2.3 that

$$-\sum_{j=1}^{+\infty}t^j\phi_j(t)=\frac{t^2(N-1)}{3(\phi(t))^3}\left\{1-\sum_{j=1}^{+\infty}t^j\frac{\phi_j(t)}{\phi(t)}+\left(\sum_{j=1}^{+\infty}t^j\frac{\phi_j(t)}{\phi(t)}\right)^2-\cdots\right\}^3+O(t^3)$$

for sufficiently small $t>0$. Comparing the coefficients of t^j ($j=1,2$) on both sides, we have

$$t\text{−term}:\phi_1(t)=0,\ t^2\text{−term}:\phi_2(t)=-\frac{N-1}{3(\phi(t))^3}.$$

We omit the choices of ϕ_j's ($j\geq 3$).

Second, we select some U_j's. We set $\widetilde{\phi}(t)=\phi(t)+t^2\phi_2(t)$ for simplicity. Put $\rho=\widetilde{d}/2\sqrt{t}$. Since $r=\phi-(2\sqrt{t}\rho-t^2\phi_2)$, we get

$$\frac{1}{r}=\frac{1}{\phi-(2\sqrt{t}\rho-t^2\phi_2)}=\frac{1}{\phi}\sum_{j=0}^{+\infty}\left\{\frac{1}{\phi}\left(2\sqrt{t}\rho-t^2\phi_2\right)\right\}^j \quad \text{for small } t>0. \tag{2.4}$$

Besides we easily see that

$$d_t=\phi'+2t\phi_2+t^2\phi_2',\ d_r=-1,\ d_{rr}=0.$$

Thus we use (2.2), (2.4) and these identities to compute that

$$
\begin{aligned}
\mathcal{L}u = &-\frac{1}{4t}\left(U_{0,\rho\rho}+2\rho U_{0,\rho}\right)+\frac{1}{\sqrt{t}}\left\{\frac{U_{0,\rho}}{2}\left(\phi'+\frac{N-1}{\phi}\right)-\left(\frac{1}{4}U_{1,\rho\rho}+\frac{\rho}{2}U_{1,\rho}-\frac{1}{2}U_1\right)\right\}\\
&+\left\{\frac{(N-1)\rho U_{0,\rho}}{\phi^2}+\frac{U_{1,\rho}}{2}\left(\phi'+\frac{N-1}{\phi}\right)-\left(\frac{1}{4}U_{2,\rho\rho}+\frac{\rho}{2}U_{2,\rho}-U_2\right)\right\}\\
&+\sqrt{t}\left\{U_{0,\rho}\phi_2+U_{1,t}+(N-1)\left(\frac{2\rho^2}{\phi^3}U_{0,\rho}+\frac{\rho U_{1,\rho}}{\phi^2}\right)+\frac{U_{2,\rho}}{2}\left(\phi'+\frac{N-1}{\phi}\right)\right.\\
&\qquad\left.-\left(\frac{1}{4}U_{3,\rho\rho}+\frac{\rho}{2}U_{3,\rho}-\frac{3}{2}U_3\right)\right\}+\cdots\\
= &\ 0,
\end{aligned}
$$

where $U_{i,\rho}=\partial U_i/\partial\rho$ and$U_{i,\rho\rho}=\partial^2 U_i/\partial\rho^2$.

We compare the coefficients of $t^{j/2}$ $(j=-2,-1,0,1,2,\ldots)$. In the case of the t^{-1}-term, we can derive

$$
U_{0,\rho\rho}+2\rho U_{0,\rho}=0 \quad \text{on } \mathbb{R}^1. \tag{2.5}
$$

Taking Lemma 2.1 into account, we impose the following condition on U_0:

$$
U_0(t,\rho)\longrightarrow\begin{cases}1 & \text{as } \rho\to+\infty,\\ -1 & \text{as } \rho\to-\infty\end{cases}\qquad \text{for any small } t>0. \tag{2.6}
$$

Then we have

$$
U_0=U_0(\rho)=\frac{2}{\sqrt{\pi}}\int_0^\rho e^{-s^2}ds. \tag{2.7}
$$

As for the $t^{-1/2}$-term, by (2.3) we obtain

$$
\frac{1}{4}U_{1,\rho\rho}+\frac{\rho}{2}U_{1,\rho}-\frac{1}{2}U_1=0 \quad \text{on } \mathbb{R}^1. \tag{2.8}
$$

Since the rate of convergence (2.6) is faster than the exponential one, combining Lemma 2.1 with this fact, we have the following condition on U_1:

$$
U_1(t,\rho)\longrightarrow 0 \quad \text{as } \rho\to\pm\infty \quad \text{for any small } t>0. \tag{2.9}
$$

Therefore we have $U_1\equiv 0$ because the uniqueness of solutions of (2.8) under (2.9) holds in the class of bounded functions.

In the case of the $t^{j/2}$-term $(j=0,1)$, from (2.3) and the fact that $U_1\equiv 0$ we get

$$
\frac{1}{4}U_{2,\rho\rho}+\frac{\rho}{2}U_{2,\rho}-U_2=\frac{(N-1)\rho U_{0,\rho}}{\phi^2} \quad \text{on } \mathbb{R}^1, \tag{2.10}
$$

$$
\frac{1}{4}U_{3,\rho\rho}+\frac{\rho}{2}U_{3,\rho}-\frac{3}{2}U_3=(N-1)\left(2\rho^2-\frac{1}{3}\right)\frac{U_{0,\rho}}{\phi^3} \quad \text{on } \mathbb{R}^1. \tag{2.11}
$$

By the same reason as above, the following condition is imposed on U_j's $(j=2,3)$:

$$
U_j(t,\rho)\longrightarrow 0 \quad \text{as } \rho\to\pm\infty \quad \text{for any small } t>0. \tag{2.12}
$$

Solving (2.10) and (2.11) under (2.12), we obtain

$$
U_2(t,\rho)=-\frac{(N-1)\rho e^{-\rho^2}}{\sqrt{\pi}(\phi(t))^2}, \qquad U_3(t,\rho)=-\frac{4(N-1)\rho^2 e^{-\rho^2}}{3\sqrt{\pi}(\phi(t))^3}. \tag{2.13}
$$

We omit selecting U_j $(j\geq 4)$.

Remark 2.2. In the above discussion, we have not applied the Fredholm alternative to derive the equation of the motion of the interface, which is used to do so in the case of the Allen–Cahn equation (see, e.g., Fife [7], de Mottoni and Schatzman [4], and Nochetto, Paolini, and Verdi [20]), because such equations as (2.8) under (2.9) have only the trivial solution. For this reason, we have used other methods such as Lemma 2.2 and Proposition 2.3 to determine ϕ_0, ϕ_1, and ϕ_2.

3. Subsolutions and supersolutions. We construct sub- and supersolutions of (1.1) in $(kh,(k+1)h)\times\mathbb{R}^N$ for $h>0$ and $k\in\mathbb{N}\cup\{0\}$. These functions will be used in sections 4 and 5 to derive the optimal rate of convergence of the BMO algorithm.

In this and the next section we assume that $\{\Gamma_t\}_{0\le t<T_0}$ is a motion of a smooth and compact hypersurface by mean curvature. The precise assumption on $\{\Gamma_t\}_{0\le t<T_0}$ is given in subsection 3.1. In addition, the existence, uniqueness, and behavior of $\{\Gamma_t\}_{0\le t<T_0}$ are mentioned in Remark 4.1 of section 4.

3.1. Signed distance function. For each $t\in[0,T_0)$, the signed distance function $d=d(t,x)$ to Γ_t is defined by

$$d\,(t,x)=\begin{cases}\operatorname{dist}\,(x,\Gamma_t) & \text{for } x\in D_t^+,\\ 0 & \text{for } x\in\Gamma_t,\\ -\operatorname{dist}\,(x,\Gamma_t) & \text{for } x\in D_t^-,\end{cases}\tag{3.1}$$

where D_t^+ denotes the bounded domain enclosed by Γ_t and $D_t^-=\mathbb{R}^N\backslash(D_t^+\cup\Gamma_t)$. Then d satisfies

$$d_t=\Delta d\quad\text{on } \Gamma_t,\ t>0.\tag{3.2}$$

For any $T\in(0,T_0)$ and $\delta>0$, let $\mathcal{N}_\delta$ be the tubular neighborhood of $\{(t,x)\in[0,T]\times\mathbb{R}^N\mid x\in\Gamma_t\}$:

$$\mathcal{N}_\delta:=\{(t,x)\in[0,T]\times\mathbb{R}^N\mid |d(t,x)|\le\delta\}.$$

We assume that $\{\Gamma_t\}_{0\le t<T_0}$ is so smooth that, for any $T<T_0$, there exists a $\delta>0$ satisfying

$$d_t,d_{x_i},d_{x_ix_j},d_{x_ix_jt},d_{x_ix_jx_k},d_{x_ix_jx_kx_l}\in L^\infty(\mathcal{N}_{5\delta})\quad\text{for } i,j,k,l=1,\dots,N.\tag{3.3}$$

It follows from this condition that for any $(t,x)\in\mathcal{N}_{5\delta}$, there is a unique $y(t,x)\in\Gamma_t$ satisfying

$$|d(t,x)|=|x-y(t,x)|.\tag{3.4}$$

Let $\widetilde{\kappa}=\widetilde{\kappa}(t,x)$ be the sum of the square of all principal curvatures at $x\in\Gamma_t$. Define

$$\kappa^s=\kappa^s(t,x):=\widetilde{\kappa}(t,y(t,x))\quad\text{for } (t,x)\in\mathcal{N}_{5\delta}.$$

Then (3.3) yields that

$$\kappa^s,\kappa^s_t,\kappa^s_{x_i},\kappa^s_{x_ix_j}\in L^\infty(\mathcal{N}_{5\delta})\quad\text{for } i,j=1,2,\dots,N.\tag{3.5}$$

Moreover, we observe by this property that there exists a $\kappa_1>0$ such that

$$|d_t-\Delta d-\kappa^s d|\le\kappa_1 d^2\quad\text{on } \mathcal{N}_{5\delta}.\tag{3.6}$$

For the details to (3.2)–(3.6), see, e.g., Chen [3], Gilbarg and Trudinger [9], and Paolini and Verdi [21].

3.2. Nonradial case. In this subsection we construct a sub- and a supersolution of (1.1) in $(kh,(k+1)h)\times\mathbb{R}^N$ for each $h>0$ and $k\in\mathbb{N}\cup\{0\}$.

We modify slightly the signed distance function d in (3.1). For any $(t,x)\in\mathcal{N}_{5\delta}$, $k\in\mathbb{N}\cup\{0\}$, and $\alpha_k\geq 0$, set

$$\underline{d}_k(t,x)=d(t,x)-\alpha_k h^2,\qquad \overline{d}_k(t,x)=d(t,x)+\alpha_k h^2.$$

We introduce smooth functions η and ζ satisfying

$$\eta(r)=\begin{cases} r, & s\leq\delta,\\ 2\delta, & s\geq 3\delta,\\ -2\delta, & s\leq -3\delta,\end{cases}\quad 0\leq\eta'\leq 1,\ |\eta''|\leq\frac{M_1}{\delta},\tag{3.7}$$

$$\zeta(r)=\begin{cases} 1, & |s|\leq\delta,\\ 0, & |s|\geq 3\delta,\end{cases}\quad 0\leq\zeta\leq 1,\ |\zeta'|\leq\frac{M_1}{\delta},\ |\zeta''|\leq\frac{M_1}{\delta^2},\tag{3.8}$$

where M_1 is a constant independent of δ. Motivated by the formal discussion in section 2, we define $\underline{u}$ and $\overline{u}$ by

$$\begin{aligned}\underline{u}\,(t,x)=U_0\left(\frac{\eta(\underline{d}_k(t,x))}{2\sqrt{t-kh}}\right)&+(t-kh)\zeta(\underline{d}_k(t,x))U_2\left(t,x,\frac{\underline{d}_k(t,x)}{2\sqrt{t-kh}}\right)\\ &-(t-kh)^{3/2}U_3-U_4\alpha_k h^2\sqrt{t-kh},\end{aligned}\tag{3.9}$$

$$\begin{aligned}\overline{u}\,(t,x)=U_0\left(\frac{\eta(\overline{d}_k(t,x))}{2\sqrt{t-kh}}\right)&+(t-kh)\zeta(\overline{d}_k(t,x))U_2\left(t,x,\frac{\overline{d}_k(t,x)}{2\sqrt{t-kh}}\right)\\ &+(t-kh)^{3/2}U_3+U_4\alpha_k h^2\sqrt{t-kh}\end{aligned}\tag{3.10}$$

in $(kh,(k+1)h)\times\mathbb{R}^N$ and $k\in\mathbb{N}\cup\{0\}$. Here $U_2=U_2(t,x,\rho)$ is given by

$$U_2(t,x,\rho)=-\frac{1}{\sqrt{\pi}}\kappa^s(t,x)\rho e^{-\rho^2},\tag{3.11}$$

and U_3 and U_4 are positive constants selected later. At $t=kh$, we set

$$\underline{u}(kh,x)=\begin{cases}1 & \text{if } \underline{d}_k(kh,x)\geq 0,\\ -1 & \text{if } \underline{d}_k(kh,x)<0,\end{cases}\qquad \overline{u}(kh,x)=\begin{cases}1 & \text{if } \overline{d}_k(kh,x)\geq 0,\\ -1 & \text{if } \overline{d}_k(kh,x)<0.\end{cases}\tag{3.12}$$

We note that U_2 satisfies

$$\frac{1}{4}U_{2,\rho\rho}+\frac{\rho}{2}U_{2,\rho}-U_2=\kappa^s\rho U_{0,\rho}\quad\text{for } (t,x)\in\mathcal{N}_{5\delta},\ \rho\in\mathbb{R}^1.\tag{3.13}$$

PROPOSITION 3.1. *Let d satisfy* (3.3) *for some $\delta>0$. Then there exist $h_1>0$, $U_3>0$, and $U_4>0$ such that for each $h\in(0,h_1)$, $k\in\mathbb{N}\cup\{0\}$, and $\alpha_k\geq 0$, $\underline{u}$ and $\overline{u}$ are, respectively, a subsolution and a supersolution of* (1.1) *in $(kh,(k+1)h)\times\mathbb{R}^N$.*

Proof. We set $k=0$ and prove the subsolution case.

For the notational simplicity, put $\underline{\rho}=\underline{d}_0/2\sqrt{t}$, $\underline{z}(t,x)=\eta(\underline{d}_0(t,x))$, and $\rho_{\underline{z}}=\underline{z}/2\sqrt{t}$. We denote by $\mathbb{L}$

$$\mathbb{L}u=u_t-\Delta u\quad\text{for } u=u(t,x).$$

It is seen by calculations that

$$\begin{aligned}\mathbb{L}\underline{u}=&-\frac{1}{4t}(U_{0,\rho\rho}|D\underline{z}|^2+2\rho_{\underline{z}}U_{0,\rho})+\frac{U_{0,\rho}}{2\sqrt{t}}(\underline{z}_t-\Delta\underline{z})-\zeta\left\{\frac{1}{4}U_{2,\rho\rho}|D\underline{d}_0|^2+\frac{\rho}{2}U_{2,\rho}-U_2\right\}\\ &+\sqrt{t}\left\{\zeta U_{2,\rho}(\underline{d}_{0,t}-\Delta\underline{d}_0)+\zeta'U_{2,\rho}|D\underline{d}_0|^2+\zeta\langle DU_{2,\rho},D\underline{d}_0\rangle\right\}\\ &+t\{\zeta(U_{2,t}-\Delta U_2)+\zeta'U_2(\underline{d}_{0,t}-\Delta\underline{d}_0)\}-\frac{3}{2}\sqrt{t}U_3-\frac{U_4\alpha_0h^2}{2\sqrt{t}},\end{aligned}$$

where $Df=(f_{x_1},\ldots,f_{x_N})$. We divide our consideration into two cases.

Case 1. $|\underline{d}_0(t,x)| \le \delta$.

In this case, $\eta = \zeta = 1$. Thus $\underline{z} = \underline{d}_0$, $\rho_{\underline{z}} = \underline{\rho}$ and

$$(3.14)\ |D\underline{z}| = |D\underline{d}_0| = |Dd| = 1, \qquad \underline{z}_t - \Delta\underline{z} = \underline{d}_{0,t} - \Delta\underline{d}_0 = d_t - \Delta d \quad \text{on } \mathcal{N}_{5\delta}.$$

Moreover, we observe that

$$(3.15)\quad \sup_{\substack{\rho\in\mathbb{R}^1\\ l=0,2}} |\rho^l U_{0,\rho}| + \sup_{(t,x)\in\mathcal{N}_{5\delta},\rho\in\mathbb{R}^1} (|U_2| + |U_{2,\rho}| + |U_{2,t}| + |\Delta U_2| + |DU_{2,\rho}|) \le K.$$

Here we have used (3.3) and (3.5) to obtain the boundedness for the second term on the left-hand side of this inequality.

It follows from (2.5), (3.3), (3.14), and (3.15) that

$$\mathbb{L}\underline{u} \le \frac{U_{0,\rho}}{2\sqrt{t}}(d_t - \Delta d) - \left\{\frac{1}{4}U_{2,\rho\rho} + \frac{\rho}{2}U_{2,\rho} - U_2\right\} + \sqrt{t}\left\{K(1+\sqrt{t}) - \frac{3}{2}U_3\right\} - \frac{U_4\alpha_0 h^2}{2\sqrt{t}}.$$

We see by the positivity of $U_{0,\rho}$, (3.6) and (3.13) that

$$\frac{U_{0,\rho}}{2\sqrt{t}}(d_t - \Delta d) - \left\{\frac{1}{4}U_{2,\rho\rho} + \frac{\rho}{2}U_{2,\rho} - U_2\right\} \le 2\sqrt{t}\kappa_1\underline{\rho}^2 U_{0,\rho} + \frac{U_{0,\rho}}{2\sqrt{t}}(\kappa^s + 2\kappa_1 d)\alpha_0 h^2.$$

Using (3.5), (3.15) and this inequality, we get

$$\mathbb{L}\underline{u} \le \sqrt{t}\left(K(1+\sqrt{t}) - \frac{3}{2}U_3\right) + \frac{h^2}{2\sqrt{t}}(U_{4,1}(U_{4,2} + 2\kappa_1\delta) - U_4)\alpha_0,$$

where $U_{4,1} = \|U_{0,\rho}\|_{L^\infty(\mathbb{R})}$ and $U_{4,2} = \|\kappa^s\|_{L^\infty(\mathcal{N}_{5\delta})}$. Therefore we can take

$$U_3 \ge \frac{4}{3}K, \qquad U_4 \ge U_{4,1}(U_{4,2} + 2\kappa_1\delta)$$

to obtain

$$\mathbb{L}\underline{u} \le 0 \quad \text{in } \{(t,x) \in (0,h)\times\mathbb{R}^N \mid |\underline{d}_0(t,x)| \le \delta\}.$$

Case 2. $|\underline{d}_0(t,x)| \ge \delta$.

In this case, $|\underline{z}| \ge \delta$. Then we see by (3.3), (3.5), and (3.7) that

$$(3.16)\qquad \begin{aligned} &|\rho_{\underline{z}}^l U_{0,\rho}(\rho_{\underline{z}})| + |U_2(t,x,\underline{\rho})| + |U_{2,\rho}(t,x,\underline{\rho})| + |U_{2,t}(t,x,\underline{\rho})| \\ &\quad + |\Delta U_2(t,x,\underline{\rho})| + |DU_{2,\rho}(t,x,\underline{\rho})| \le Ke^{-\delta^2/8t} \end{aligned}$$

for small $t > 0$ and $l = 0,1,2,3$. By (3.3) and (3.7) we get

$$|w_t - \Delta w| \le K \quad (w = z, \underline{d}_0).$$

Using (2.5), (3.8), (3.13), (3.16) and this estimate, we obtain

$$\mathbb{L}\underline{u} \le Ke^{-\delta^2/8t} - \frac{3\sqrt{t}}{2}U_3.$$

We choose $h_1 > 0$ so small that $e^{-\delta^2/8t} \le \sqrt{t}$ for all $t \in (0,h_1)$. Hence, letting $U_3 \ge 2K/3$ and $U_4 \ge 0$, we have

$$\mathbb{L}\underline{u} \le 0 \quad \text{in } \{(t,x) \in (0,h)\times\mathbb{R}^N \mid |\underline{d}_0(t,x)| \ge \delta\}$$

for all $h \in (0,h_1)$.

Consequently, taking a large $U_3 > 0$ and setting $U_4 = U_{4,1}(U_{4,2} + 2\kappa_1\delta)$, we obtain

$$\mathbb{L}\underline{u} \leq 0 \quad \text{in } (0,h) \times \mathbb{R}^N$$

for all $h \in (0, h_1)$. The supersolution case can be shown by a method similar to the above.

For the case $k \geq 1$, we can show the assertion of this proposition by using the same h_1, U_3, and U_4 as in the case $k = 0$. $\square$

3.3. Radial case. This subsection is devoted to the construction of a sub- and a supersolution of (1.1) which are radially symmetric. We recall that if $u = u(t,r)$ $(r = |x|)$ and $C_0 = B(0,R)$, then the problem (1.1) turns to (2.1). We assume $N = 2$ to simplify our arguments.

For any $R > 0$, put

$$\begin{aligned} \phi(t) &= \sqrt{R^2 - 2t},\ \phi_1(t) = \frac{1}{3(\phi(t))^3} \quad \text{on } [0,h], \\ \widetilde{d}(t,r) &= \phi(t) - t^2\phi_1(t) - r \quad \text{on } [0,h] \times \mathbb{R}^1. \end{aligned} \tag{3.17}$$

Let η and ζ be the same functions as (3.7) and (3.8), respectively. Set $z(t,r) = \eta(\widetilde{d}(t,r))$ and define $\underline{u}$ and $\overline{u}$ by

$$\underline{u}(t,r) = U_0\left(\frac{z(t,r)}{2\sqrt{t}}\right) + \zeta(\widetilde{d}(t,r))\left\{tU_2\left(t, \frac{\widetilde{d}(t,r)}{2\sqrt{t}}\right) + t^{3/2}U_3\left(t, \frac{\widetilde{d}(t,r)}{2\sqrt{t}}\right)\right\} - t^2U_4, \tag{3.18}$$

$$\overline{u}(t,r) = U_0\left(\frac{z(t,r)}{2\sqrt{t}}\right) + \zeta(\widetilde{d}(t,r))\left\{tU_2\left(t, \frac{\widetilde{d}(t,r)}{2\sqrt{t}}\right) + t^{3/2}U_3\left(t, \frac{\widetilde{d}(t,r)}{2\sqrt{t}}\right)\right\} + t^2U_4 \tag{3.19}$$

for $t > 0$, $r \in \mathbb{R}^1$. Here U_2, U_3 are given by (2.13) and U_4 is a constant selected later. At $t = 0$, we put

$$\underline{u}(0,r) = \overline{u}(0,r) = \begin{cases} 1, & \text{if } \widetilde{d}(0,r) \geq 0, \\ -1, & \text{if } \widetilde{d}(0,r) < 0. \end{cases} \tag{3.20}$$

PROPOSITION 3.2. *Fix $\delta \in (0, R/5 \wedge 1)$. Then there exist $h_2 > 0$ and $U_4 > 0$ such that for each $h \in (0, h_2)$, $\underline{u}$ and $\overline{u}$ are, respectively, a subsolution and a supersolution of* (2.1) *in $(0,h) \times \mathbb{R}^1$. In addition, they satisfy the boundary condition of* (2.1).

Proof. We assume $k = 0$ and prove the subsolution case. Set $\rho_z = z/2\sqrt{t}$ and $\widetilde{\rho} = \widetilde{d}/2\sqrt{t}$ for the notational simplicity.

It is observed by calculations that

$$\begin{aligned} \mathcal{L}\underline{u} = &-\frac{1}{4t}\left(U_{0,\rho\rho}z_r^2 + 2\rho_z U_{0,\rho}\right) + \frac{U_{0,\rho}}{2\sqrt{t}}\left(z_t - z_{rr} - \frac{1}{r}z_r\right) - \zeta\left(\frac{1}{4}U_{2,\rho\rho}\widetilde{d}_r^2 + \frac{\widetilde{\rho}}{2}U_{2,\rho} - U_2\right) \\ &+ \sqrt{t}\zeta\left\{U_{2,\rho}\left(\widetilde{d}_t - \widetilde{d}_{rr} - \frac{1}{r}\widetilde{d}_r\right) - \left(\frac{1}{4}U_{3,\rho\rho}\widetilde{d}_r^2 + \frac{\widetilde{\rho}}{2}U_{3,\rho} - \frac{3}{2}U_3\right)\right\} \\ &+ t\left[\zeta\left\{U_{2,t} + \sqrt{t}U_{3,t} + \frac{U_{3,\rho}}{2}\left(\widetilde{d}_t - \widetilde{d}_{rr} - \frac{1}{r}\widetilde{d}_r\right)\right\} - 2U_4\right] \\ &- 2\zeta'\widetilde{d}_r\left\{tU_{2,\rho}\frac{\widetilde{d}_r}{2\sqrt{t}} + t^{3/2}U_{3,\rho}\frac{\widetilde{d}_r}{2\sqrt{t}}\right\} \\ &+ t\left\{\zeta'\left(\widetilde{d}_t - \widetilde{d}_{rr} - \frac{1}{r}\widetilde{d}_r\right) - \zeta''\widetilde{d}_r^2\right\}(U_2 + \sqrt{t}U_3). \end{aligned}$$

We divide our consideration into two cases.

Case 1. $|\widetilde{d}(t,r)| \le \delta$.

In this case, $\eta = \zeta = 1$. Thus $z = \widetilde{d}$, $\rho_z = \widetilde{\rho}$ and

$$w_r = -1,\ w_t - w_{rr} - \frac{1}{r}w_r = \phi' - 2t\phi_1 - t^2\phi_1' + \frac{1}{r} \quad \text{for } w = z, \widetilde{d}. \tag{3.21}$$

Moreover, we see from (2.13) and (3.17) that

$$\sup_{\substack{\rho\in\mathbb{R}^1\\ l=0,1,2,3}} |\rho^l U_{0,\rho}| + \sup_{\substack{t\in[0,R^2/4],\rho\in\mathbb{R}^1\\ j=2,3,l=0,1,2,3}} \left(|\rho^l U_j| + |\rho^l U_{j,\rho}| + |U_{j,t}|\right) \le K. \tag{3.22}$$

By (2.5), (3.21), and this boundedness, we have

$$\begin{aligned}\mathcal{L}u \le{}& \frac{U_{0,\rho}}{2\sqrt{t}}\left(\phi' - 2t\phi_1 - t^2\phi_1' + \frac{1}{r}\right) - \left(\frac{1}{4}U_{2,\rho\rho} + \frac{\widetilde{\rho}}{2}U_{2,\rho} - U_2\right)\\ &+ t^{1/2}\left\{U_{2,\rho}\left(\phi' - 2t\phi_1 - t^2\phi_1' + \frac{1}{r}\right) - \left(\frac{1}{4}U_{3,\rho\rho} + \frac{\widetilde{\rho}}{2}U_{3,\rho} - \frac{3}{2}U_3\right)\right\}\\ &+ \frac{tU_{3,\rho}}{2}\left(\phi' - 2t\phi_1 - t^2\phi_1' + \frac{1}{r}\right) + t(K - 2U_4).\end{aligned}$$

We estimate $\phi' - 2t\phi_1 - t^2\phi_1' + 1/r$. We remark that $\widetilde{d} = 2\sqrt{t}\widetilde{\rho}$ and $r = \phi - (2\sqrt{t}\widetilde{\rho} + t^2\phi_1)$. It follows from (3.17) that there exists an $h_2 \in (0,\delta)$ such that $\phi(t) \ge 4\delta$ and $t^2\phi_1(t) \le \delta$ for $t \in (0,h_2)$. Thus we get $2\sqrt{t}|\widetilde{\rho}| + t^2\phi_1 \le 2\delta$ and

$$\frac{1}{r} \le \frac{1}{\phi}\left\{1 + \frac{1}{\phi}(2\sqrt{t}\widetilde{\rho} + t^2\phi_1) + \frac{1}{\phi^2}(2\sqrt{t}\widetilde{\rho} + t^2\phi_1)^2\right\} + \frac{1}{3\delta\phi^3}(2\sqrt{t}|\widetilde{\rho}| + t^2\phi_1)^3$$

for all $t \in (0,h_2)$. Since ϕ satisfies $\phi' + 1/\phi = 0$ and ϕ_1' is bounded, we observe that

$$\phi' - 2t\phi_1 - t^2\phi_1' + \frac{1}{r} \le \frac{2\sqrt{t}\widetilde{\rho}}{\phi^2} + \frac{4t\widetilde{\rho}^2}{\phi^3} - 2t\phi_1 + \frac{8t^{3/2}|\widetilde{\rho}|^3}{3\delta\phi^3} + Kt^2. \tag{3.23}$$

We observe from the positivity of $U_{0,\rho}$ on $\mathbb{R}^1$, (2.10), (2.11), (3.22), and this estimate that

$$\begin{aligned}&\frac{U_{0,\rho}}{2\sqrt{t}}\left(\phi' - 2t\phi_1 - t^2\phi_1' + \frac{1}{r}\right) - \left(\frac{1}{4}U_{2,\rho\rho} + \frac{\widetilde{\rho}}{2}U_{2,\rho} - U_2\right)\\ &\qquad -\sqrt{t}\left(\frac{1}{4}U_{3,\rho\rho} + \frac{\widetilde{\rho}}{2}U_{3,\rho} - \frac{3}{2}U_3\right) \le U_{0,\rho}(Kt\widetilde{\rho}^3 + Kt^{3/2}) \le tK(1+\sqrt{t}).\end{aligned}$$

In addition, (3.22) and (3.23) yield that

$$t^{1/2}U_{2,\rho}\left(\phi' - 2t\phi_1 - t^2\phi_1' + \frac{1}{r}\right) + \frac{tU_{3,\rho}}{2}\left(\phi' - 2t\phi_1 - t^2\phi_1' + \frac{1}{r}\right) \le Kt.$$

Therefore we obtain

$$\mathcal{L}\underline{u} \le t(K(1+\sqrt{t}) - 2U_4).$$

Consequently, taking U_4 sufficiently large, we obtain

$$\mathcal{L}\underline{u} \le 0 \quad \text{in } \{(t,r) \in (0,h) \times \mathbb{R}^1 \mid |\widetilde{d}(t,r)| \le \delta\}$$

for all $h \in (0,h_2)$.

Case 2. $|\widetilde{d}(t,r)| \geq \delta$.

In this case, $|z| \geq \delta$. Then we see by (2.13) that

$$|\rho_z^l U_{0,\rho}(\rho_z)| + |\widetilde{\rho}^l U_j(t,\widetilde{\rho})| + |\widetilde{\rho}^l U_{j,\rho}(t,\widetilde{\rho})| + |U_{j,t}(t,\widetilde{\rho})| \leq K e^{-\delta^2/8t} \tag{3.24}$$

for small $t > 0$, $j = 2, 3$, and $l = 0, 1, 2, 3$. Then it is easily observed by the fact that $\delta \leq r \leq R + 3\delta$ and by (3.7) and (3.17) that

$$\left| w_t - w_{rr} - \frac{1}{r} w_r \right| \leq K \quad \text{for } w = z, \widetilde{d}.$$

Thus we apply (3.7), (3.8), (3.24), and this inequality to obtain

$$\mathcal{L}\underline{u} \leq K e^{-\delta^2/8t} - 2tU_4.$$

Taking $h_2 > 0$ smaller if necessary, we have $e^{-\delta^2/8t} \leq t$ for all $t \in (0, h_2)$. Thus we choose $U_4 \geq K/2$ to have

$$\mathcal{L}\underline{u} \leq 0 \quad \text{in } \{(t,r) \in (0,h) \times \mathbb{R}^1 \mid |\widetilde{d}(t,r)| \geq \delta\}$$

for all $h \in (0, h_2)$.

Consequently, taking $U_4 > 0$ large and $h_2 > 0$ small, we obtain

$$\mathcal{L}\underline{u} \leq 0 \quad \text{in } (0,h) \times \mathbb{R}^1$$

for all $h \in (0, h_2)$. Since $\underline{u}(t,r) = U_0(\delta/\sqrt{t}) - t^2 U_4$ for $0 \leq r \ll 1$, it is easily verified that $\underline{u}_r(t,0) = 0$. The supersolution case can be shown in a similar way. ∎

4. Rate of convergence. In this section we consider the rate of convergence of the BMO algorithm to the motion of a smooth and compact hypersurface by mean curvature.

To state our theorem, we rewrite the BMO algorithm as follows. Let $\Gamma_0 \subset \mathbb{R}^N$ be a smooth and compact hypersurface and $C_0 \subset \mathbb{R}^N$ the compact set such that $\partial C_0 = \Gamma_0$. Fix a time step $h > 0$. Let $u^h = u^h(t,x)$ be the solution of

$$\begin{cases} u_t^h = \Delta u^h \quad \text{in } (kh, (k+1)h) \times \mathbb{R}^N, \\ u^h(kh, x) = \begin{cases} 1, & x \in C_k, \\ -1, & x \in \mathbb{R}^N \setminus C_k, \end{cases} \\ C_k = \begin{cases} \text{the above set } C_0 & \text{for } k = 0, \\ \left\{ x \in \mathbb{R}^N \;\middle|\; \lim\limits_{t \to kh-} u^h(t,x) \geq 0 \right\} & \text{for } k = 1, 2, \ldots. \end{cases} \end{cases} \tag{4.1}$$

Set

$$C_t^h = \begin{cases} \{x \in \mathbb{R}^N \mid u^h(t,x) \geq 0\} & \text{for } t \neq kh, \\ C_k & \text{for } t = kh, \end{cases}$$

$$\Gamma_t^h = \partial C_t^h (= \{x \in \mathbb{R}^N \mid u^h(t,x) = 0\}). \tag{4.2}$$

We note that C_{kh}^h coincides with C_k defined in the introduction and that Γ_t^h is a smooth and compact hypersurface for each $t \geq 0$, $h > 0$. Furthermore, there exists an $R_0 > 0$ such that $\Gamma_t^h \subset \overline{B(0, R_0)}$ for all $t \geq 0$ and $h > 0$ (see Barles and Georgelin [1, Lemma 5.1]). Using this formulation, we have the following theorem.

Theorem 4.1. *Let* $\{\Gamma_t\}_{0\le t<T_0}$ *be a smooth and compact motion by mean curvature satisfying* (3.3). *Let* Γ_t^h *be defined by* (4.2). *Then, for any* $T \in (0, T_0)$, *there exist* $h_0 > 0$ *and* $L > 0$ *such that*

$$\sup_{t\in[0,T]} d_H\left(\Gamma_t^h, \Gamma_t\right) \le Lh \tag{4.3}$$

for all $h \in (0, h_0)$. *Here* $d_H(A, B)$ *denotes the Hausdorff distance between* A, $B \subset \mathbb{R}^N$.

Remark 4.1. On the existence, uniqueness, and behavior of a motion by mean curvature $\{\Gamma_t\}_{0\le t<T_0}$, the following results are known. Assume that Γ_0 is the boundary of class $C^{k,\alpha}$ of a bounded domain ($k \ge 2, 0 < \alpha < 1$).

(i) For some $T_0 = T_0(\Gamma_0) > 0$, there uniquely exists a smooth and compact motion by mean curvature $\{\Gamma_t\}_{0\le t<T_0}$ starting from Γ_0. Moreover, the signed distance function d defined by (3.1) is of class $C^{(k+\alpha)/2,k+\alpha}(\mathcal{N}_{\delta_0})$ for some small $\delta_0 > 0$ (see Evans and Spruck [6]).

(ii) If $N = 2$ or Γ_0 is convex, then the motion $\{\Gamma_t\}_{0\le t<T_0}$ can be extended up to $T_0 = T_{max}$, where T_{max} is the extinction time for Γ_t (see Gage and Hamilton [8], Grayson [11], and Huisken [13]). In other cases the singularities may appear before Γ_t shrinks to a point (see, e.g., Grayson [12]).

Therefore (4.3) is valid before Γ_t shrinks to a point or develops the singularities.

Proof of Theorem 4.1. Set $k = 0$, $u = u^h$ and let $\underline{u}$ and $\overline{u}$ be defined by (3.9) and (3.10), respectively. Define

$$\underline{\Sigma}_t^h := \{x \in \mathbb{R}^N \mid \underline{u}(t,x) = 0\}, \qquad \underline{\Theta}_t^h := \{x \in \mathbb{R}^N \mid \underline{d}_0(t,x) = 0\},$$
$$\overline{\Sigma}_t^h := \{x \in \mathbb{R}^N \mid \overline{u}(t,x) = 0\}, \qquad \overline{\Theta}_t^h := \{x \in \mathbb{R}^N \mid \overline{d}_0(t,x) = 0\}.$$

Note that these sets are smooth and compact hypersurfaces.

Step 1. We prove that there exist $h_{0,1} > 0$, L_1 and $L_2 > 0$ such that

$$d_H(\underline{\Theta}_t^h, \underline{\Sigma}_t^h),\ d_H(\overline{\Theta}_t^h, \overline{\Sigma}_t^h) \le (L_1 h\alpha_0 + L_2)h^2 \tag{4.4}$$

for all $t \in (0, h)$ and $h \in (0, h_{0,1})$.

We easily see from (3.15) that there exists an $h_{0,1} = h_{0,1} > 0$ such that

$$|\underline{d}_0(t,x)| < 2\sqrt{t} \quad \text{on } \underline{\Sigma}_t^h \cup \underline{\Theta}_t^h \tag{4.5}$$

for all $t \in [0, h)$ and $h \in (0, h_{0,1})$. Moreover, taking $h_{0,1}$ smaller if necessary, we observe that for any $t \in (0, h_{0,1})$ and $x \in \mathbb{R}^N$ satisfying $|\underline{d}_0(t,x)| \le 2\sqrt{t}$,

$$\langle D\underline{u}(t,x), D\underline{d}_0(t,x)\rangle \ge \frac{U_{0,\rho}(1)}{2\sqrt{t}} - Kt \ge \frac{1}{10\sqrt{\pi t}}. \tag{4.6}$$

Let $\underline{x} \in \underline{\Theta}_t^h$ and take $\underline{y} \in \underline{\Sigma}_t^h$ so that $\underline{y} = \underline{x} + |\underline{x} - \underline{y}| D\underline{d}_0(t, \underline{x})$. Applying the mean value theorem, we obtain

$$\begin{aligned} 0 = \underline{u}(t,\underline{y}) &= \underline{u}(t,\underline{x}) + \langle D\underline{u}(t, \theta\underline{x} + (1-\theta)\underline{y}), \underline{y} - \underline{x}\rangle \quad (0 < \theta < 1) \\ &= -t^{3/2}U_3 - U_4\alpha_0 h^2\sqrt{t} + |\underline{x} - \underline{y}|\langle D\underline{u}(t, \theta\underline{y} + (1-\theta)\underline{x}), D\underline{d}_0(t,\underline{x})\rangle. \end{aligned}$$

It is easily seen that $D\underline{d}_0(t,\underline{x}) = Dd(t, \theta\underline{y} + (1-\theta)\underline{x})$. Hence we can use (4.6) to have $|\underline{x} - \underline{y}| \le 10\sqrt{\pi}t(U_4\alpha_0 h^2 + tU_3)$ and thus

$$\sup_{x\in\underline{\Theta}_t^h} \operatorname{dist}(x, \underline{\Sigma}_t^h) \le (L_1 h\alpha_0 + L_2)h^2,$$

where $L_1 = 10\sqrt{\pi}U_4$, $L_2 = 10\sqrt{\pi}U_3$ and U_3, U_4 are the same constants as in Proposition 3.1. Similarly, we can show that

$$\sup_{x\in\underline{\Sigma}_t^h} \operatorname{dist}(x, \underline{\Theta}_t^h),\ \sup_{x\in\overline{\Theta}_t^h} \operatorname{dist}(x, \overline{\Sigma}_t^h),\ \sup_{x\in\overline{\Sigma}_t^h} \operatorname{dist}(x, \overline{\Theta}_t^h) \le (L_1 h\alpha_0 + L_2)h^2.$$

Hence we obtain (4.4).

Step 2. We show that there exists an $h_0 > 0$ such that

$$d_H(\Gamma_t, \Gamma_t^h) \le \{(1 + L_1 h)\alpha_0 + L_2\}h^2 \quad \text{for all } t \in (0, h) \text{ and } h \in (0, h_0). \tag{4.7}$$

Let $h_1 > 0$ be given in Proposition 3.1. Set $h_0 = \min\{h_{0,1}, h_1\}$ and fix $h \in (0, h_0)$. Since it is easily verified by (3.12) that $\underline{u}(0, x) \le u(0, x) \le \overline{u}(0, x)$ on $\mathbb{R}^N$, we have $\underline{u}(t, x) \le u(t, x) \le \overline{u}(t, x)$ on $[0, h) \times \mathbb{R}^N$ by Proposition 3.1 and the comparison principle for the heat equation. This implies that

$$\Gamma_t^h \subset \{x \in \mathbb{R}^N \mid \underline{u}(t, x) \le 0 \le \overline{u}(t, x)\} \quad \text{for all } t \in [0, h), \tag{4.8}$$

that is, Γ_t^h lies between $\underline{\Sigma}_t^h$ and $\overline{\Sigma}_t^h$.

For any $x \in \Gamma_t$, we can find an $\underline{x} \in \underline{\Theta}_t^h$ such that $\operatorname{dist}(x, \underline{\Theta}_t^h) = \alpha_0 h^2 = |x - \underline{x}|$. From Step 1, we have

$$\operatorname{dist}(x, \underline{\Sigma}_t^h) \le |x - \underline{x}| + \operatorname{dist}(\underline{x}, \underline{\Sigma}_t^h) \le \{(1 + L_1 h)\alpha_0 + L_2\}h^2$$

for all $t \in [0, h)$. Since $x \in \Gamma_t$ is arbitrary, we get

$$\sup_{x\in\Gamma_t} \operatorname{dist}(x, \underline{\Sigma}_t^h) \le \{(1 + L_1 h)\alpha_0 + L_2\}h^2 \quad \text{for all } t \in [0, h). \tag{4.9}$$

Similarly, we can show that

$$\sup_{x\in\Gamma_t} \operatorname{dist}(x, \overline{\Sigma}_t^h) \le \{(1 + L_1 h)\alpha_0 + L_2\}h^2 \quad \text{for all } t \in [0, h), \tag{4.10}$$

with the same L_1, L_2 as above.

Hence, using (4.8)–(4.10), we obtain

$$\sup_{x\in\Gamma_t} \operatorname{dist}(x, \Gamma_t^h) \le \max\left\{\sup_{x\in\Gamma_t} \operatorname{dist}(x, \overline{\Sigma}_t^h), \sup_{x\in\Gamma_t} \operatorname{dist}(x, \underline{\Sigma}_t^h)\right\} \le \{(1 + L_1 h)\alpha_0 + L_2\}h^2$$

for all $t \in (0, h)$.

Since Γ_t also lies between $\underline{\Sigma}_t^h$ and $\overline{\Sigma}_t^h$, by the same argument as above, we get

$$\sup_{x\in\Gamma_t^h} \operatorname{dist}(x, \Gamma_t) \le \max\left\{\sup_{x\in\overline{\Sigma}_t^h} \operatorname{dist}(x, \Gamma_t), \sup_{x\in\underline{\Sigma}_t^h} \operatorname{dist}(x, \Gamma_t)\right\} \le \{(1 + L_1 h)\alpha_0 + L_2\}h^2$$

for all $t \in (0, h)$. Therefore we obtain (4.7).

Step 3. We consider the case $k = 1$. Put $\alpha_1 = (1+L_1 h)\alpha_0 + L_2$ and fix $h \in (0, h_0)$. Then we can see by Proposition 3.1 that $\underline{u}$ and $\overline{u}$ are, respectively, a subsolution and a supersolution of (1.1) in $(h, 2h) \times \mathbb{R}^N$. Since Γ_t^h moves continuously in t in the sense of the Hausdorff distance (cf. Goto, Ishii, and Ogawa [10, Corollary 3.1]), we observe by (4.7) that

$$\Gamma_h^h \subset \{x \in \mathbb{R}^N \mid \underline{d}_1(h, x) \le 0 \le \overline{d}_1(h, x)\}.$$

It is easily seen by (3.12) and this inclusion that $\underline{u}(h,x) \le u(h,x) \le \overline{u}(h,x)$ on $\mathbb{R}^N$, and hence we obtain $\underline{u}(t,x) \le u(t,x) \le \overline{u}(t,x)$ on $[h,2h) \times \mathbb{R}^N$ by the comparison principle for the heat equation. Therefore applying the argument in Step 2, we have

$$d_H(\Gamma_t, \Gamma_t^h) \le \{(1+L_1h)\alpha_1 + L_2\}h^2 \quad \text{for all } t \in [h,2h).$$

Step 4. We select $m \in \mathbb{N}$ satisfying $mh \le T < (m+1)h$ for each $h \in (0,h_0)$ and repeat the arguments in Steps 2–3 inductively. Set

$$\alpha_k = (1+L_1h)\alpha_{k-1} + L_2 \quad \text{for } k = 1,2,\dots,m.$$

Then it follows from Proposition 3.1 that $\underline{u}$ and $\overline{u}$ are, respectively, a subsolution and a supersolution of (1.1) in $(kh,(k+1)h) \times \mathbb{R}^N$. Since we can verify from (3.12) that $\underline{u}(kh,x) \le u(kh,x) \le \overline{u}(kh,x)$ on $\mathbb{R}^N$, we have $\underline{u}(t,x) \le u(t,x) \le \overline{u}(t,x)$ on $[kh,(k+1)h) \times \mathbb{R}^N$ by the comparison principle for the heat equation. Thus we obtain

$$d_H(\Gamma_t, \Gamma_t^h) \le \{(1+L_1h)\alpha_k + L_2\}h^2 \quad \text{for all } t \in [kh,(k+1)h) \text{ and } k = 0,1,2,\dots,m$$

by an argument similar to Step 2.

Step 5. We estimate the sequence $\{\alpha_k\}_{1\le k\le m}$. Since $\alpha_0 \ge 0$ is arbitrary, we can take $\alpha_0 = 0$. Then we observe from the definition of α_k that

$$\begin{aligned}\alpha_k &= (1+L_1h)\alpha_{k-1} + L_2 = (1+L_1h)^2\alpha_{k-2} + L_2\{1+(1+L_1h)\}\\ &= \cdots\\ &= L_2\sum_{l=1}^{k}(1+L_1h)^{l-1} \le L_2\frac{(1+L_1h)^m - 1}{L_1h}.\end{aligned}$$

By the choice of m, we get

$$\alpha_k \le \frac{L_2(e^{L_1T_0}-1)}{L_1h} \le \frac{L_2T_0e^{L_1T_0}}{h} \qquad \text{for } k = 0,1,\dots,m.$$

Thus we obtain

$$\sup_{t\in[0,T]} d_H(\Gamma_t,\Gamma_t^h) \le \left\{(1+L_1)L_2T_0e^{L_1T_0} + L_2\right\}h$$

for all $h \in (0,h_0)$. Therefore the proof is completed. □

5. Optimality. This section is devoted to the optimality for the estimate in Theorem 4.1. For this purpose, we consider a circle evolving by curvature.

Let $C_0 = \{x \in \mathbb{R}^2 \mid |x| \le 1\}$ and fix a time step $h > 0$. Let u^h be the radially symmetric solution of (4.1). Then we can easily verify that for any $t > 0$, Γ_t^h defined in (4.2) is a circle centered at the origin, and we denote by $R_h(t)$ the radius of Γ_t^h. Put $\phi(t) = \sqrt{1-2t}$ and let $\Gamma_t = \partial B(0,\phi(t))$. Take $\delta \in (0,1/5)$ and set

$$T_{max} = \frac{1}{2},\ T_\delta = \frac{1}{2} - \frac{25\delta^2}{2},\ m = \left[\frac{T_\delta}{h}\right], \tag{5.1}$$

where T_{max} is the extinction time for Γ_t and $[s]$ denotes the Gauss symbol for $s \in \mathbb{R}$. Applying Theorem 4.1, we see that for each $\delta \in (0,1/5)$, there exist $h_2 > 0$ and $M_1 > 0$ such that

$$\sup_{t\in[0,T_\delta]} |R_h(t) - \phi(t)| \le M_1h \quad \text{for all } h \in (0,h_2). \tag{5.2}$$

In the remainder of this section we consider a more precise behavior of R_h as $h \searrow 0$ than (5.2) and show that the estimate of Theorem 4.1 is optimal. Our main result of this section is stated as follows.

THEOREM 5.1. *For each $\delta \in (0, 1/5)$, there exist $h_0 > 0$ and $L > 0$ such that*

$$|R_h(t) - (\phi(t) - t^2\phi_1^0(t))| \le Lt^{5/2} \quad \text{for } t \in [t, h], \tag{5.3}$$

$$|R_h(t) - (\phi(t) - h\varphi(t))| \le Lh^{3/2} \quad \text{for } t \in [h, T_\delta] \tag{5.4}$$

for all $h \in (0, h_0)$. Here $\phi_1^0(t)$, $\varphi(t)$ are given by

$$\phi_1^0(t) = \frac{1}{3(\phi(t))^3}, \qquad \varphi(t) = -\frac{\log\phi(t)}{3\phi(t)}.$$

This theorem shows that Γ_t^h moves faster than Γ_t.

As a corollary of Theorem 5.1, we obtain an estimate of the distance between Γ_t and Γ_t^h.

COROLLARY 5.2. *For each $\delta \in (0, 1/5)$, there exist $h_1 \in (0, h_0)$, $\overline{L} > 0$, and $\underline{L} > 0$ such that*

$$\underline{L}t^2 \le d_H(\Gamma_t, \Gamma_t^h) \le \overline{L}t^2 \quad \text{for } t \in [0, h], \tag{5.5}$$

$$\underline{L}th \le d_H(\Gamma_t, \Gamma_t^h) \le \overline{L}th \quad \text{for } t \in [h, T_\delta] \tag{5.6}$$

for all $h \in (0, h_1)$.

This corollary shows that in the case where $\{\Gamma_t\}_{t\ge 0}$ is a motion of a smooth and compact hypersurface by mean curvature, the linear rate in h is optimal to the convergence of the BMO algorithm.

We prepare some functions which will be used in the following subsections. For $k = 0, 1, 2, \ldots, m$, we define

$$\phi^k(t) = \sqrt{(R_h(kh))^2 - 2t}, \; \phi_1^k(t) = \frac{1}{3(\phi^k(t))^3} \quad \text{for } t \in [0, h]. \tag{5.7}$$

Note that $\phi^0 = \phi$ and $(\phi^k)' = -1/\phi^k$ in $(0, h)$. It is easily seen by (5.2) and these facts that, for any $\delta > 0$, there exists an $h_0 > 0$ such that

$$\frac{1}{3(1+\delta)^3} \le \phi_1^k(t) \le \frac{1}{3\cdot(3\delta)^3}, \qquad \frac{1}{(1+\delta)^5} \le (\phi_1^k)'(t) \le \frac{1}{(3\delta)^5} \tag{5.8}$$

for all $t \in [0, h]$, $k = 0, 1, \ldots, m$, and $h \in (0, h_0)$.

5.1. Short-time asymptotics of . In this subsection, we prove the following theorem suggested by Proposition 2.3.

THEOREM 5.3. *There exist $h_3 > 0$ and $L_1 > 0$ such that*

$$|R_h(t + kh) - (\phi^k(t) - t^2\phi_1^k(t))| \le L_1 t^{5/2}$$

for all $t \in [0, h)$, $k = 0, 1, \ldots, m$, and $h \in (0, h_3)$.

Proof. Set $k = 0$ and $u = u^h$ for simplicity. Define $\widetilde{\phi}(t) = \phi(t) - t^2\phi_1^0(t)$ and $\widetilde{d}(t, r) = \widetilde{\phi}(t) - r$. Let $\underline{u}$ and $\overline{u}$ be defined by (3.18) and (3.19), respectively. Let h_2 be given in Proposition 3.2. Since $\underline{u}(0, r) = \overline{u}(0, r) = u(0, r)$ on $\mathbb{R}^1$ by (3.20), applying Proposition 3.2 and the comparison principle for the heat equation, we get

$$\underline{u}(t, r) \le u(t, r) \le \overline{u}(t, r) \quad \text{in } [0, h) \times \mathbb{R}^1 \tag{5.9}$$

for all $h \in (0, h_2)$. Let $\underline{\phi} = \underline{\phi}(t)$ and $\overline{\phi} = \overline{\phi}(t)$ be the zero of $\underline{u}(t, \cdot)$ and $\overline{u}(t, \cdot)$, respectively. Then it follows from (5.9) and (5.12) that

$$\underline{\phi}(t) \le R_h(t) \le \overline{\phi}(t) \quad \text{for all } t \in [0, h) \text{ and } h \in (0, h_2) \tag{5.10}$$

for all $h \in (0, h_2)$.

We estimate $\underline{\phi}$ and $\overline{\phi}$. At first, it is easily seen by (3.22) that there exists an $h_3 \in (0, \delta^2/4 \wedge h_2)$ such that

$$-t^2 U_4 = \underline{u}(t, \widetilde{\phi}(t)) < 0 = \underline{u}(t, \underline{\phi}(t)) < \underline{u}(t, \widetilde{\phi}(t) - 2t^2), \tag{5.11}$$

$$\underline{u}_r(t, r) \le -\frac{U_{0,\rho}(1)}{2\sqrt{t}} + Kt \le -\frac{U_{0,\rho}(1)}{4\sqrt{t}} < 0 \tag{5.12}$$

for all $t \in (0, h_3)$ and $r \in \mathbb{R}$ satisfying $|\widetilde{d}(t, r)| \le 2\sqrt{t} (\le \delta)$. Here we have used (3.22) to derive these estimates. Thus we can observe from (5.11) and (5.12) that

$$\widetilde{\phi}(t) - Kt^{5/2} \le \underline{\phi}(t) \le \widetilde{\phi}(t) \quad \text{for all } t \in (0, h) \text{ and } h \in (0, h_3). \tag{5.13}$$

We can also show similarly that

$$\widetilde{\phi}(t) \le \overline{\phi}(t) \le \widetilde{\phi}(t) + Kt^{5/2} \quad \text{for all } t \in (0, h) \text{ and } h \in (0, h_3). \tag{5.14}$$

Combining (5.13), (5.14) with (5.10) and setting $L_1 = K$, we obtain the result for $k = 0$.

In the case $k \ge 1$, let η be defined by (3.7) and set

$$\widetilde{d}^k(t, r) = \phi^k(t) - t^2 \phi_1^k(t) - r, \qquad z^k(t, r) = \eta(\widetilde{d}^k(t, r)).$$

We define $\underline{u}^k$ and $\overline{u}^k$ by (3.18)–(3.20) with replacing $\widetilde{d}$, z with $\widetilde{d}^k$, z^k, respectively. Then we can check that Proposition 3.2 holds for these $\underline{u}^k$ and $\overline{u}^k$ for any $h \in (0, h_2)$ and small $h_2 > 0$. Since we can easily verify that the choices of h_2 and h_3 depend only on $\delta \in (0, 1/5)$, we can apply the above argument to obtain the result. ◻

5.2. Derivation of a corrector for . In this subsection we formally calculate $R_h(t) - \phi(t)$ and find a corrector for $R_h(t)$ on each time interval $[kh, (k+1)h)$ $(k \in \mathbb{N} \cup \{0\})$. By Theorem 5.3, we see that

$$|R_h(\bar{t}) - (\phi(\bar{t}) - \bar{t}^2 \phi_1^0(\bar{t}))| \le L_1 \bar{t}^{5/2} \qquad \text{for all } \bar{t} \in [0, h] \text{ and } h \in (0, h_3). \tag{5.15}$$

Next we compute $R_h(\bar{t} + h) - \phi(\bar{t} + h)$ for $\bar{t} \in [0, h]$. Theorem 5.3 yields that $|R_h(\bar{t} + h) - (\phi^1(\bar{t}) - \bar{t}^2 \phi_1^1(\bar{t}))| \le L_1 \bar{t}^{5/2}$ for all $\bar{t} \in [0, h_3]$. From (5.15) and this estimate, we have

$$\begin{aligned} R_h(\bar{t} + h) - \phi(\bar{t} + h) &\ge \phi^1(\bar{t}) - \bar{t}^2 \phi_1^1(\bar{t}) - \phi(\bar{t} + h) - L_1 \bar{t}^{5/2} \\ &= \sqrt{(R_h(h))^2 - 2\bar{t}} - \sqrt{1 - 2(\bar{t} + h)} - \bar{t}^2 \phi_1^1(\bar{t}) + L_1 \bar{t}^{5/2} \\ &\ge \sqrt{(\phi(h) - h^2 \phi_1^0(h) - L_1 h^{5/2})^2 - 2\bar{t}} - \sqrt{(\phi(h))^2 - 2\bar{t}} - \bar{t}^2 \phi_1^1(\bar{t}) - L_1 \bar{t}^{5/2} \\ &=: I_1 - I_2 - \bar{t}^2 \phi_1^1(\bar{t}) - L_1 \bar{t}^{5/2}. \end{aligned} \tag{5.16}$$

We observe by Taylor expansion to I_1 and I_2 around $\bar{t} = 0$ that

$$
(5.17)\quad \begin{aligned} I_1 - I_2 = &-h^2\phi_1^0(h) - L_1h^{5/2} - \frac{(h^2\phi_1^0(h) + L_1h^{5/2})\bar{t}}{\phi(h)(\phi(h) - h^2\phi_1^0(h) - L_1h^{5/2})} \\ &- \int_0^{\bar{t}} \left(\frac{(\bar{t}-s)}{\{(\phi(h) - h^2\phi_1^0(h) - L_1h^{5/2})^2 - 2s\}^{3/2}} - \frac{(\bar{t}-s)}{\{(\phi(h))^2 - 2s\}^{3/2}} \right) ds. \end{aligned}
$$

It follows from

$$
(5.18)\qquad \left(\frac{1}{1-r}\right)^3 \le 1 + 8r \quad \text{for all } |r| \ll 1
$$

that

$$
\begin{aligned} &\frac{1}{\{(\phi(h) - h^2\phi_1^0(h) - L_1h^{5/2})^2 - 2s\}^{3/2}} \\ &\quad \le \frac{1}{\{(\phi(h))^2 - 2s\}^{3/2}} \left(1 + \frac{8(h^2\phi_1^0(h) + L_1h^{5/2})(2\phi(h) - h^2\phi_1^0(h) - L_1h^{5/2})}{(\phi(h))^2 - 2s}\right) \end{aligned}
$$

for any $s \in [0, h]$ and small $h > 0$. By using this inequality, we have

$$
\begin{aligned} I_1 - I_2 \ge &-h^2\phi_1^0(h) - L_1h^{5/2} - \frac{\bar{t}(h^2\phi_1^0(h) + L_1h^{5/2})}{\phi(h)(\phi(h) - h^2\phi_1^0(h) - L_1h^{5/2})} \\ &- \frac{4\bar{t}^2(h^2\phi_1^0(h) + L_1h^{5/2})(2\phi(h) - h^2\phi_1^0(h) - L_1h^{5/2})}{\{(\phi(h))^2 - 2\bar{t}\}^{5/2}}. \end{aligned}
$$

In addition, since we also see by (5.18) that

$$
\frac{1}{\phi(h) - h^2\phi_1^0(h) - L_1h^{5/2}} \le \frac{1}{\phi(h)}\left(1 + \frac{8(h^2\phi_1^0(h) + L_1h^{5/2})}{\phi(h)}\right)
$$

for any small $h > 0$, noting that $h^2\phi_1^0(h) + L_1h^{5/2} > 0$ and $(\phi(h))^2 - 2\bar{t} \ge (\phi(2h))^2$ on $[0, h]$, we get

$$
(5.19)\quad \begin{aligned} I_1 - I_2 \ge &-h^2\phi_1^0(h) - L_1h^{5/2} \\ &- \bar{t}h^2\left\{\left(\frac{1}{(\phi(h))^2} + \frac{8h\phi(h)}{(\phi(2h))^5}\right)\phi_1^0(h) + \frac{8h^2(\phi_1^0(h))^2}{(\phi(h))^3}\right\} \\ &- \bar{t}\left\{\left(\frac{1}{(\phi(h))^2} + \frac{8h\phi(h)}{(\phi(2h))^5} + \frac{16h^2\phi_1^0(h)}{(\phi(h))^3}\right)L_1h^{5/2} + \frac{8(L_1h^{5/2})^2}{(\phi(h))^3}\right\}. \end{aligned}
$$

Setting

$$
\begin{aligned} \phi_1^1 &= \left(\frac{1}{(\phi(h))^2} + \frac{8h\phi(h)}{(\phi(2h))^5}\right)\phi_1^0(h) + \frac{8h^2(\phi_1^0(h))^2}{(\phi(h))^3}, \\ \mathcal{L}^1 &= \left(\frac{1}{(\phi(h))^2} + \frac{8h\phi(h)}{(\phi(2h))^5} + \frac{16h^2\phi_1^0(h)}{(\phi(h))^3}\right)L_1h^{5/2} + \frac{8(L_1h^{5/2})^2}{(\phi(h))^3}, \end{aligned}
$$

we obtain

$$
R_h(\bar{t} + h) - \phi(\bar{t} + h) \ge -h^2\phi_1^0(h) - \bar{t}^2\phi_1^1(\bar{t}) - \bar{t}h^2\phi_2^1 - L_1h^{5/2} - L_1\bar{t}^{5/2} - \mathcal{L}^1\bar{t}
$$

for all $\bar{t} \in [0, h]$ and small $h > 0$.

To consider the case $k = 2, 3, \ldots, m$, we define

$$\psi^k = \psi_1^k + h\psi_2^k,\ \psi_1^k = \sum_{l=0}^{k} \phi_1^l(h),\ \psi_2^k = \sum_{l=0}^{k} \phi_2^l,\ L_2^k = h\sum_{l=0}^{k} \mathcal{L}^l \tag{5.20}$$

$$(\phi_2^0 = 0, \mathcal{L}^0 = 0),$$

$$\phi_2^l = \left(\frac{1}{(\phi(lh))^2} + \frac{8h\phi(lh)}{(\phi((l+1)h))^5}\right)\psi^{l-1} + \frac{8h^2(\psi^{l-1})^2}{(\phi(lh))^3}, \tag{5.21}$$

$$\begin{aligned}\mathcal{L}^l = {}& \left(\frac{1}{(\phi(lh))^2} + \frac{8h\phi(lh)}{(\phi((l+1)h))^5} + \frac{16h^2\psi^{l-1}}{(\phi(lh))^3}\right)(L_1 lh^{5/2} + L_2^{l-1}) \\ & + \frac{8(L_1 lh^{5/2} + L_2^{l-1})^2}{(\phi(lh))^3}.\end{aligned} \tag{5.22}$$

Assume that for $k \geq 2$,

$$\begin{aligned}& R_h(\bar{t} + (k-1)h) - (\phi(\bar{t} + (k-1)h) \\ & \quad \geq -h^2\psi^{k-2} - \bar{t}^2\phi_1^{k-1}(\bar{t}) - \bar{t}h^2\phi_2^{k-1} - L_1(k-1)h^{5/2} - L_1\bar{t}^{5/2} - L_2^{k-2} - \mathcal{L}^{k-1}\bar{t}\end{aligned}$$

for all $\bar{t} \in [0, h]$. Since by Theorem 5.3 we have $|R_h(\bar{t}+kh) - (\phi^k(\bar{t}) - \bar{t}^2\phi_1^k(\bar{t}))| \leq L_1\bar{t}^{5/2}$ for all $\bar{t} \in [0, h_2]$, similar calculations to (5.16) yield

$$\begin{aligned}& R_h(\bar{t} + kh) - \phi(\bar{t} + kh) \\ & \geq \sqrt{(\phi(kh) - h^2\psi^{k-1} - L_1kh^{5/2} - L_2^{k-1})^2 - 2\bar{t}} - \sqrt{(\phi(kh))^2 - 2\bar{t}} - \bar{t}^2\phi_1^k(\bar{t}) - L_1\bar{t}^{5/2} \\ & =: I_4 - \bar{t}^2\phi_1^k(\bar{t}) - L_1\bar{t}^{5/2}.\end{aligned}$$

Replacing $h^2\phi_1^0(h)$ and $L_1h^{5/2}$ with, respectively, $h^2\psi^{k-1}$ and $L_1kh^{5/2} + L_2^{k-1}$ in the case $k = 1$, we get

$$I_4 \geq -h^2\psi^{k-1} - \bar{t}^2\phi_1^k(\bar{t}) - \bar{t}h^2\phi_2^k - L_1kh^{5/2} - L_1\bar{t}^{5/2} - L_2^{k-1} - \mathcal{L}^k\bar{t}.$$

Thus we have

$$\begin{aligned}R_h(\bar{t} + kh) - \phi(\bar{t} + kh) \geq {}& -h^2\psi^{k-1} - \bar{t}^2\phi_1^k(\bar{t}) - \bar{t}h^2\phi_2^k \\ & - kL_1h^{5/2} - L_1\bar{t}^{5/2} - L_2^{k-1} - \mathcal{L}^k\bar{t}\end{aligned}$$

for all $\bar{t} \in [0, h]$ and small $h > 0$.

Similarly, we can observe that

$$\begin{aligned}R_h(\bar{t} + kh) - \phi(\bar{t} + kh) \leq {}& -h^2\psi^{k-1} - \bar{t}^2\phi_1^k(\bar{t}) - \bar{t}h^2\phi_2^k \\ & + L_1kh^{5/2} + L_1\bar{t}^{5/2} + L_2^k + \mathcal{L}^k\bar{t}\end{aligned}$$

for all $\bar{t} \in [0, h]$, $k \in \mathbb{N} \cup \{0\}$ and small $h > 0$. Therefore we obtain

$$\begin{aligned}& |R_h(\bar{t} + kh) - \{\phi(\bar{t} + kh) - (h^2\psi^{k-1} + \bar{t}^2\phi_1^k(\bar{t}) + \bar{t}h^2\phi_2^k)\}| \\ & \quad \leq L_1kh^{5/2} + L_1\bar{t}^{5/2} + L_2^k + \mathcal{L}^k\bar{t}\end{aligned} \tag{5.23}$$

for all $t \in [0, h]$, $k \in \mathbb{N} \cup \{0\}$ and small $h > 0$. This inequality shows that the term $h^2\psi^{k-1} + \bar{t}^2\phi_1^k(\bar{t}) + \bar{t}h^2\phi_2^k$ is a (formal) corrector for $R_h(t)$.

5.3. Proofs of Theorem 5.1 and Corollary 5.2. This subsection is devoted to the estimates and the limits of ψ_1^k, ψ_2^k, L_2^k, and $\mathcal{L}^k$ and the proofs of Theorem 5.1 and Corollary 5.2. Remember that we have taken $\delta \in (0, 1/5)$ and set T_{max}, T_δ, and m as in (5.1).

PROPOSITION 5.4. *Let $\varphi_1(t) = (1/\phi(t) - 1)/3$. Then there exist $h_3 > 0$ and $M_3 > 0$ such that*

$$\sup_{t\in[0,T_\delta]} |\varphi_1(t) - h\psi_1^{[t/h]}| \leq M_3 h \quad \text{for all } h \in (0, h_3).$$

Proof. We remark that $\varphi_1(t) = \frac{1}{3}\int_0^t \frac{1}{(\phi(s))^3} ds$. Set $k = [t/h]$. It is easily seen by the definition of ϕ_1^k in (5.20) and $kh \leq T_{max}$ that

$$|\varphi_1(t) - h\psi_1^{[t/h]}| \leq \frac{1}{3}\left|\int_0^{kh} \frac{1}{(\phi(s))^3} ds - h\sum_{l=0}^{k} \frac{1}{(\phi(lh))^3}\right| + \frac{h}{3(\phi(T_\delta))^3}$$
$$+ \frac{T_{max}}{3} \max_{0\leq l\leq k} \left|\frac{1}{(\phi(lh))^3} - \frac{1}{\{(R_h(lh))^2 - 2h\}^{3/2}}\right|.$$

Since $1/(\phi(t))^3$ is increasing in t, we easily observe that

$$\left|\int_0^{kh} \frac{1}{(\phi(s))^3} ds - h\sum_{l=0}^{k} \frac{1}{(\phi(lh))^3}\right| \leq \left(\frac{1}{(\phi(T_\delta))^3} - \frac{1}{(\phi(0))^3}\right) h.$$

Combining (5.2) with this inequality, we have the result. $\square$

We obtain the estimates for ψ_2^k by the following lemma.

LEMMA 5.5. *There exist $h_4 > 0$, $M_4 > 0$, and $M_5 > 0$ such that*

$$M_4(kh)^2 \leq h^2\psi_2^k \leq M_5$$

for $k = 0, 1, \ldots, m$ and $h \in (0, h_4)$.

Proof. The definition of ψ_1^k in (5.20) yields that

$$0 < \psi_1^1 \leq \cdots \leq \psi_1^m. \tag{5.24}$$

Besides we easily see that

$$\alpha := \sup_{0\leq l\leq m, h>0} \max\left\{\frac{1}{(\phi(lh))^2}, \frac{8\phi(lh)}{(\phi((l+1)h))^5}, \frac{8}{(\phi(lh))^3}\right\} < +\infty.$$

Thus, for $l = 1, 2, \ldots, m$, ϕ_2^l satisfies

$$\phi_2^l \leq \alpha\{(1+h)\psi^{l-1} + h^2(\psi^{l-1})^2\}. \tag{5.25}$$

We estimate ϕ_2^l by using (5.21) and this inequality.

First, for sufficiently small $h > 0$ we get $\psi_2^0(h) = \phi_2^0(h) \leq 1$ and

$$\phi_2^1 \leq \alpha(2 + h).$$

Fix $k = 2, 3, \ldots, m$ and let $l = 2, 3, \ldots, k$. From the fact $m = [T_\delta/h]$ we remark that

$$(1 + \alpha h(2+h))^l \leq (1 + \alpha h(2+h))^m \leq e^{3\alpha T_{max}} \tag{5.26}$$

for $l = 2, 3, \dots, m$ and $h \in (0,1)$. Taking (5.24) and this estimate into account, we choose $h_4 \in (0,1)$ such that for any $h \in (0, h_4)$,

$$h(\psi_1^m + h\psi_1^1) \le M_{5,1},\ M_{5,1}h \le M_{5,1}e^{3\alpha T_{max}}h \le 1, \tag{5.27}$$

where $M_{5,1} = \varphi_1(T_\delta) + 1$.

It follows from (5.21) with $l = 2$ and (5.25) that

$$\phi_2^2 \le \alpha\{(1+h)(\psi_1^1 + h\psi_2^1) + h^2(\psi_1^1 + h\psi_2^1)^2\}.$$

We easily see by (5.27) that $h^2(\psi_1^1 + h\psi_2^1) \le M_{5,1}h \le 1$. Thus we get

$$\phi_2^2 \le \alpha(2+h)(\psi_1^1 + h\psi_2^1). \tag{5.28}$$

In the case of $l = 3$, since $\psi^2 = \psi_1^2 + h\psi_2^1 + h\phi_2^2$, we see by (5.24), (5.25), and (5.28) that

$$\phi_2^3 \le \alpha[(1+h)(1+\alpha h(2+h))(\psi_1^2 + h\psi_2^1) + h^2\{(1+\alpha h(2+h))(\psi_1^2 + h\psi_2^1)\}^2].$$

Using (5.26) and (5.27), we get

$$h^2(1+\alpha h(2+h))(\psi_1^2 + h\psi_2^1) \le M_{5,1}e^{3\alpha T_{max}}h \le 1.$$

Therefore we have

$$\phi_2^3 \le \alpha(2+h)(1+\alpha h(2+h))(\psi_1^2 + h\psi_2^1). \tag{5.29}$$

As to the case of $l = 4$, note that $\psi^3 = \psi_1^3 + h\psi_2^1 + h(\phi_2^2 + \phi_2^3)$. Hence it is observed by (5.25)–(5.29) and a similar argument that

$$\phi_2^4 \le \alpha(2+h)(1+\alpha h(2+h))^2(\psi_1^3 + h\psi_2^1).$$

By repeating this procedure, we can show that

$$\phi_2^l \le \alpha(2+h)(1+\alpha h(2+h))^{l-2}(\psi_1^{l-1} + h\psi_2^1) \quad \text{for } l = 4, \dots, k.$$

Summing up $l = 0$ to $l = k$ and using (5.27), we get

$$h^2\psi_2^k = h^2\sum_{l=0}^{k}\phi_2^l \le 3\alpha(1 + M_{5,1}e^{3\alpha T_{max}}) \quad \text{for } h \in (0, h_4).$$

Setting $M_5 = 3\alpha(1 + M_{5,1}e^{3\alpha T_{max}})$, we have an upper bound for $h^2\psi_2^k$.

As for a lower bound for $h^2\psi_2^k$, we observe from the definition of ϕ_2^l in (5.21) and (5.8) that

$$\phi_2^l \ge \frac{\psi^l}{(\phi(lh))^2} \ge \frac{l\phi_1^0(0)}{(\phi(0))^2} \ge \frac{l}{3(1+\delta)^3}.$$

Thus, putting $M_4 = 1/6(1+\delta)^3$, we obtain $h^2\psi_2^k \ge M_4(kh)^2$. $\square$

We use this lemma to prove the following.

PROPOSITION 5.6. *Let $\varphi_2(t) = -\log\phi(t)/3\phi(t) - \varphi_1(t)$. Then there exist $h_5 > 0$ and $M_6 > 0$ such that*

$$\sup_{t\in[0,T_\delta]} |\varphi_2(t) - h^2\psi_2^{[t/h]}| \le M_6 h \quad \text{for all } h \in (0, h_5).$$

Proof. We easily see that φ_2 is a unique solution of

$$\varphi_2(t) = \int_0^t \frac{\varphi_1(s) + \varphi_2(s)}{(\phi(s))^2} ds.$$

For each $t \in [0, T_\delta]$, set $k = [t/h]$. It is easily observed from the definition of ψ_2^k in (5.20) and Lemma 5.5 that

$$\begin{aligned} |\varphi_2(t) - h^2 \psi_2^{[t/h]}| \le\ & (1 + Kh) \left\{ \left| \int_0^t \frac{\varphi_1(s)}{(\phi(s))^2} ds - h \sum_{l=0}^{k} \frac{h\psi_1^l}{(\phi(lh))^2} \right| \right. \\ & \left. + \left| \int_0^t \frac{\varphi_2(s)}{(\phi(s))^2} ds - h \sum_{l=0}^{k} \frac{h^2\psi_2^l}{(\phi(lh))^2} \right| \right\} + Kh \\ =:\ & (1 + Kh)(I_1 + I_2) + Kh. \end{aligned}$$

Since $1/(\phi(t))^2$ (resp., ψ^k) is increasing with respect to t (resp., k), we have

$$\int_{lh}^{(l+1)h} \frac{h\psi_1^{[s/h]}}{(\phi(s))^2} ds \le \frac{h^2 \psi_1^{l+1}}{(\phi((l+1)h))^2}.$$

Using Proposition 5.4 and this inequality, we calculate

$$\begin{aligned} I_1 &= \left| \int_0^{kh} \frac{\varphi_1(s) - h\psi_1^{[s/h]} + h\psi_1^{[s/h]}}{(\phi(s))^2} ds - h \sum_{l=0}^{k} \frac{h\psi_1^l}{(\phi(lh))^2} + \int_{kh}^{t} \frac{\varphi_1(s)}{(\phi(s))^2} ds \right| \\ &\le Kh \int_0^{kh} \frac{ds}{(\phi(s))^2} + h \left| \sum_{l=0}^{k} \frac{h\psi_1^{l+1}}{(\phi((l+1)h))^2} - \sum_{l=0}^{k} \frac{h\psi_1^l}{(\phi(lh))^2} \right| + Kh \\ &\le Kh. \end{aligned}$$

Similarly we can show that

$$I_2 \le \int_0^t \frac{|\varphi_2(s) - h^2\psi_2^{[s/h]}|}{(\phi(s))^2} ds + Kh.$$

Therefore we obtain

$$|\varphi_2(t) - h^2\psi_2^{[t/h]}| \le Kh + (1 + Kh) \int_0^t \frac{|\varphi_2(s) - h^2\psi_2^{[s/h]}|}{(\phi(s))^2} ds.$$

We apply the Gronwall inequality to get

$$|\varphi_2(t) - h^2\psi_2^{[t/h]}| \le Kh \exp\left((1 + Kh) \int_0^{T_\delta} \frac{ds}{(\phi(s))^2} \right)$$

for all $t \in [0, T_\delta]$ and small $h > 0$. Thus we have the result. ◻

Finally we obtain the bounds for $\mathcal{L}^k$ and L_2^k.

PROPOSITION 5.7. *There exist $h_6 > 0$ and $M_7 > 0$ such that*

$$\mathcal{L}^k \le M_7(kh)h^{3/2}, \qquad L_2^k \le M_7(kh)^2 h^{3/2}$$

for all $k = 1, 2, \ldots, m$ and $h \in (0, h_6)$.

Proof. The proof is similar to that of Lemma 5.5.

It follows from (5.2) and Proposition 5.4 that

$$\alpha := \sup_{0\le l\le m, h>0} \max\left\{\frac{1}{(\phi(lh))^2}, \frac{8\phi(lh)}{(\phi((l+1)h)^5}, \frac{16h\psi^l}{(\phi(lh))^3}, \frac{8}{(\phi(lh))^3}\right\} < +\infty.$$

Thus, for $l = 1, 2, \ldots, m$, $\mathcal{L}^l$ satisfies

$$\mathcal{L}^l \le \alpha\{(1+h)(L_1 lh^{5/2} + L_2^{l-1}) + (L_1 lh^{5/2} + L_2^{l-1})^2\}. \tag{5.30}$$

We estimate $\mathcal{L}^l$ by using (5.22) and this inequality.

First, for sufficiently small $h \in (0, 1)$, we have

$$\mathcal{L}^1 \le \alpha(2+h)L_1 h^{5/2} \le M_{7,1}h^{5/2},\ M_{7,1} = 3\alpha L_1.$$

Fix $k = 2, 3, \ldots, m$ and let $l = 2, 3, \ldots, k$. In view of

$$(1 + \alpha h(2+h))^l \le (1 + \alpha h(2+h))^m \le e^{3\alpha T_{max}} \tag{5.31}$$

for $l = 2, 3, \ldots, m$ and $h \in (0, 1)$, we can choose $h_6 \in (0, 1)$ such that for any $h \in (0, h_6)$,

$$e^{3\alpha T_{max}}(T_{max}L_1 + M_{7,1}h)h^{3/2} \le 1. \tag{5.32}$$

It is easily seen from (5.20) that $L_2^1 = h\mathcal{L}^1 \le M_{7,1}h^{7/2}$. Thus we get, by (5.30),

$$\mathcal{L}^2 \le \alpha\{(1+h)(2L_1h^{5/2} + M_{7,1}h^{7/2}) + (2L_1h^{5/2} + M_{7,1}h^{7/2})^2\}. \tag{5.33}$$

We easily observe by (5.32) that $2L_1h^{5/2} + M_{7,1}h^{7/2} \le 1$. Hence we have

$$\mathcal{L}^2 \le \alpha(2+h)(2L_1h^{5/2} + M_{7,1}h^{7/2}) \le 3\alpha(L_1 + M_{7,1}h)(2h)h^{3/2}. \tag{5.34}$$

In the case of $l = 3$, since $L_2^2 = h(\mathcal{L}^1 + \mathcal{L}^2)$, we see by (5.30) that

$$\begin{aligned}\mathcal{L}^3 \le \alpha[&(1+h)(1+\alpha h(2+h))(3L_1h^{5/2} + M_{7,1}h^{7/2})\\ &+\{(1+\alpha h(2+h))(3L_1h^{5/2} + M_{7,1}h^{7/2})\}^2].\end{aligned}$$

Using (5.31) and (5.32), we obtain

$$(1+\alpha h(2+h))(3L_1h^{5/2} + M_{7,1}h^{7/2}) \le e^{3\alpha T_{max}}(T_{max}+1)L_1h^{3/2} \le 1.$$

Thus we get

$$\mathcal{L}^3 \le \alpha(2+h)(1+\alpha h(2+h))(3L_1h^{5/2} + M_{7,1}h^{7/2}) \le 3\alpha e^{3\alpha T_{max}}(L_1 + M_{7,1}h)(3h)h^{3/2}.$$

We repeat the above arguments to obtain

$$\mathcal{L}^l \le \alpha(2+h)(1+\alpha h(2+h))^{l-2}(L_1 lh^{5/2} + M_{7,1}h^{7/2}) \le M_7(lh)h^{3/2} \tag{5.35}$$

for $M_7 = 3\alpha e^{3\alpha T_{max}}(L_1 + M_{7,1})$. From this estimate, we get

$$L_2^k = h\sum_{l=1}^{k} \mathcal{L}^l \le M_7(kh)^2h^{3/2}$$

for all $k = 0, 1, \ldots, m$ and $h \in (0, h_5)$. $\square$

We observe from Propositions 5.4–5.7 that

$$0 \le h^2\psi^{k-1} + \bar{t}^2\phi_1^k(\bar{t}) + \bar{t}h^2\phi_2^k \le Kh,\ 0 \le L_1kh^{5/2} + L_1\bar{t}^{5/2} + L_2^k + \mathcal{L}^k\bar{t} \le Kh^{3/2}$$

for all $\bar{t} \in [0, h)$, $k = 0, 1, \ldots, m$. Thus (5.23) rigorously holds for sufficiently small $h > 0$.

Proof of Theorem 5.1. In the case $k = 0$, (5.3) is obtained by Theorem 5.3. Thus we assume $k \ge 1$ and prove (5.4).

Noting that $\varphi = \varphi_1 + \varphi_2$, in view of Propositions 5.4–5.7, we can find an $h_0 > 0$ so small that

$$|R_h(t) - (\phi(t) - h\varphi(t))| \le Kh^2 + Kh^{3/2} \le Lh^{3/2} \tag{5.36}$$

for some large $L > 0$ and all $t \in [0, T_\delta]$ and $h \in (0, h_0)$. ☐

Proof of Corollary 5.2. In the case $k = 0$, we have (5.5) by (5.3). Thus we may assume $k \ge 1$ and $kh \le t < (k+1)h$. Let $h_0 > 0$ be given in Theorem 5.1.

Using (5.8), (5.23), Lemma 5.5, and Proposition 5.7, we have

$$\begin{aligned} R_h(t) - \phi(t) &\le -kKh^2 - M_4(kh)^2h - K(t-kh)^2 \\ &\qquad + kL_1h^{5/2} + L_1(t-kh)^{5/2} + M_7(kh)^2h^{3/2} + M_7(t-kh)(kh)h^{3/2} \\ &\le -(K - L_1h^{1/2})(kh^2 + (t-kh)^2) - (M_4 - 2M_7h^{1/2})(kh)^2h \\ &\le -(K - L_1h^{1/2})(kh^2 + (t-kh)^2) \end{aligned}$$

for any $h \in (0, h_0)$ satisfying $M_4 \ge 2M_7h^{1/2}$. Take $h_1 \in (0, h_0)$ such that $2M_7h^{1/2} \le M_4$ and $Lh^{1/2} \le K/2$. Since we get, from $kh \ge t - h$,

$$kh^2 + (t-kh)^2 \ge \frac{1}{2}kh^2 + \frac{1}{2}kh^2 \ge \frac{1}{2}(t-h)h + \frac{1}{2}kh^2 \ge \frac{1}{2}th,$$

setting $\underline{L} = K/4$, we obtain

$$R_h(t) - \phi(t) \le -\underline{L}th.$$

Similarly we can show that

$$R_h(t) - \phi(t) \ge -\overline{L}th \quad \text{for all } t \in [h, T_\delta] \text{ and } h \in (0, h_1)$$

for some $\overline{L} > 0$. From these two estimates, we have (5.6). ☐

6. Appendix. We give the proof of Lemma 2.2.

Proof of Lemma 2.2. Put $\widetilde{r} = vt$. Then it follows from Evans [5, Theorem 4.1] that $|v - (N-1)/R| \le Kt^{1/2}$ for any small $t > 0$. Hence we estimate v as $t \searrow 0$ more precisely. In the following we always assume that $t > 0$ is sufficiently small.

To simplify our consideration, we treat the following problem instead of (2.1):

$$\begin{cases} u_t - \Delta u = 0 & \text{in } (0, +\infty) \times \mathbb{R}^N, \\ u(0, x) = \begin{cases} 1, & x \in \overline{B(z_0, R)}, \\ 0, & x \in \mathbb{R}^N \setminus \overline{B(z_0, R)}, \end{cases} \end{cases} \tag{6.1}$$

where $z_0 = (0, \ldots, 0, R) \in \mathbb{R}^N$. We note that, in this setting, $C_t^h = \{x \in \mathbb{R}^N \mid u(t, x) \ge 1/2\}$. Then the solution u of (6.1) can be represented as

$$u(t, x) = \frac{1}{(4\pi t)^{N/2}} \int_{B(z_0, R)} e^{-|y-x|^2/4t} dy.$$

Set $x_0 = (0, \ldots, 0, vt)$ with $v \geq 0$ and assume that $u(t, x_0) = 1/2$. Then

$$\frac{1}{2} = \frac{1}{(4\pi t)^{N/2}} \int_{\overline{B(z_0,R)}} e^{-(|y'|^2+|y_N-vt|^2)/4t} dy.$$

Since the lower hemisphere of $\partial B(z_0, R)$ can be written as

$$y_N = R - \sqrt{R^2 - |y'|^2} \quad (y = (y', y_N) \in \partial B(z_0, R),\ y' \in \overline{B'(0,R)}),$$

where $B'(0,R) = \{x' \in \mathbb{R}^{N-1} \mid |x'| < R\}$, we observe that

$$\frac{1}{2} = \frac{1}{(4\pi t)^{N/2}} \int_{\overline{B'(0,R)}} e^{-|y'|^2/4t} \int_{\overline{R-\sqrt{R^2-|y'|^2}}}^{+\infty} e^{-|y_N-vt|^2/4t} dy_N dy' + O(e^{-K/t}).$$

Changing the variable by setting $z' = y'/2\sqrt{t}$, $z_N = (y_N - vt)/2\sqrt{t}$, we compute that

$$\frac{1}{2} = \frac{1}{\pi^{N/2}} \int_{\overline{B'(0,R/2\sqrt{t})}} e^{-|z'|^2} \left\{ \int_0^{+\infty} - \int_0^{\sqrt{t}(g(t,z')-v/2)} \right\} e^{-|z_N|^2} dz_N dz' + O(e^{-K/t}).$$

Here $g(t, z')$ is defined by

$$g(t,z') = \frac{1}{2t}(R - \sqrt{R^2 - 4t|z'|^2}) = \frac{|z'|^2}{R} + \frac{t|z'|^4}{R^3} + O\left(\frac{t^2|z'|^6}{R^5}\right). \tag{6.2}$$

Since

$$\frac{1}{\pi^{N/2}} \int_{\overline{B'(0,R/2\sqrt{t})}} e^{-|z'|^2} \int_0^{+\infty} e^{-|z_N|^2} dz_N dz' = \frac{1}{2} - O(e^{-K/t}),$$

we deduce that

$$\frac{1}{\pi^{N/2}} \int_{\overline{B'(0,R/2\sqrt{t})}} e^{-|z'|^2} \int_0^{\sqrt{t}(g(t,z')-v/2)} e^{-|z_N|^2} dz_N dz' = O(e^{-K/t}).$$

Using Taylor expansion to the function $\int_0^s e^{-|z_N|^2} dz_N$ around $s = 0$, we can observe that

$$\frac{1}{\pi^{N/2}} \int_{\overline{B'(0,R/2\sqrt{t})}} e^{-|z'|^2} \left\{ (2g(t,z') - v) - \frac{t}{12}(2g(t,z') - v)^3 \right\} = O(t^2). \tag{6.3}$$

By the way, lengthy calculations yield that

$$\int_{\overline{B'(0,R/2\sqrt{t})}} e^{-|z'|^2} |z'|^{2k} dz' = \pi^{(N-1)/2} \prod_{l=1}^{k} \frac{N+2l-3}{2} - O(e^{-K/t}). \tag{6.4}$$

We use this estimate to compute the left-hand side of (6.3). Combining (6.2) with (6.4) with $k = 1, 2$, we get

$$\frac{1}{\pi^{(N-1)/2}} \int_{\overline{B'(0,R/2\sqrt{t})}} e^{-|z'|^2} (2g(t,z') - v) dz' = \frac{N-1}{R} + \frac{(N^2-1)t}{2R^3} - v + O(e^{-K/t}).$$

We note that

$$(2g(t,z') - v)^3 = \frac{8|z'|^6}{R^3} - \frac{12v|z'|^4}{R^2} + \frac{6v^2|z'|^2}{R} - v^3 + tP(t, |z'|),$$

where $tP(t,|z'|)$ is the remainder term satisfying

$$\int_{\mathbb{R}^{N-1}} e^{-|z'|^2} P(t,|z'|)dz' = O(1).$$

We use (6.4) with $k = 0, 1, 2, 3$ and this estimate to obtain

$$\begin{aligned} &\frac{t}{12\pi^{(N-1)/2}} \int_{B'(0,R/2\sqrt{t})} e^{-|z'|^2} (2g(t,z') - v)^3 dz' \\ &\quad = \frac{(N+3)(N^2-1)t}{12R^3} - \frac{(N^2-1)t}{4R^2} v + \frac{(N-1)t}{4R} v^2 - \frac{t}{12} v^3 + O(t^2). \end{aligned}$$

Therefore we obtain the following:

$$\frac{N-1}{R} - \frac{(N^2-1)(N-3)t}{12R^3} - v + \frac{(N^2-1)t}{4R^2} v - \frac{(N-1)t}{4R} v^2 + \frac{t}{12} v^3 = O(t^2).$$

Let $G(v)$ be the left-hand side of this estimate. We find a root v^* of $G(v) = 0$ near $v_0 = (N-1)/R$ and consider $v^* - v_0$. We easily see that

$$G(v_0 + s) = \frac{(N-1)(3N-1)t}{6R^3} - \left(1 - \frac{(N-1)t}{2R^2}\right) s + \frac{t}{12} s^3. \tag{6.5}$$

Set

$$v_1 = v_1(t) := \frac{(N-1)(3N-1)t}{6R^3 - 3R(N-1)t}.$$

It is observed by (6.5) that

$$G(v_0 + 2v_1) < 0 < G(v_0 + v_1) \le Kt^4, \qquad -1 \le \frac{dG}{dv}(v_0 + s) \le -\frac{1}{2} \text{ for } s \in [0, 2v_1).$$

Hence there exists a root v^* of $G(v) = 0$ satisfying $0 < v^* - (v_0 + v_1) \le Kt^4$. Thus we conclude that

$$\left| v^* - \left(v_0 + \frac{(N-1)(3N-1)t}{6R^3} \right) \right| \le Kt^2.$$

In the case of $G(v) = O(t^2)$, we can obtain the result of Lemma 2.2 by slightly modifying the above argument. $\square$

REFERENCES

[1] G. Barles and C. Georgelin, *A simple proof of convergence for an approximation scheme for computing motions by mean curvature*, SIAM J. Numer. Anal., 32 (1995), pp. 484–500.

[2] J. Bence, B. Merriman, and S. Osher, *Diffusion generated motion by mean curvature*, in Computational Crystal Growers Workshop, J. Taylor, ed., Sel. Lectures Math., AMS, Providence, RI, 1992, pp. 73–83.

[3] X. Chen, *Generation and propagation of the interface for reaction-diffusion equations*, J. Differential Equations, 96 (1992), pp. 116–141.

[4] P. de Mottoni and M. Schatzman, *Geometrical evolution of developed interfaces*, Trans. Amer. Math. Soc., 347 (1995), pp. 1533–1589.

[5] L. C. Evans, *Convergence of an algorithm for mean curvature motion*, Indiana Univ. Math. J., 42 (1993), pp. 533–557.

[6] L. C. Evans and J. Spruck, *Motion of level sets by mean curvature* II, Trans. Amer. Math. Soc., 330 (1992), pp. 321–322.

[7] P. C. Fife, *Dynamics of Internal Layers and Diffusive Interfaces*, SIAM, Philadelphia, 1988.
[8] M. Gage and R. Hamilton, *The heat equation shrinking convex plane curves*, J. Differential Geometry, 23 (1986), pp. 69–95.
[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1983.
[10] Y. Goto, K. Ishii, and T. Ogawa, *Method of the distance function to the Bence–Merriman–Osher algorithm for motion by mean curvature*, Comm. Pure Appl. Anal., 4 (2005), pp. 311–339.
[11] M. A. Grayson, *The heat equation shrinks embedded plane curves to round points*, J. Differential Geom., 26 (1987), pp. 285–314.
[12] M. A. Grayson, *A short note on the evolution of surfaces via mean curvature*, Duke Math. J., 58 (1989), pp. 555–558.
[13] G. Huisken, *Flow by mean curvature of convex surfaces into spheres*, J. Differential Geom., 20 (1984), pp. 237–266.
[14] H. Ishii, *A generalization of the Bence, Merryman and Osher algorithm for motion by mean curvature*, in Curvature Flows and Related Topics, A. Damlamian, J. Spruck, and A. Visintin, eds., Gakuto Intermat. Ser. Math. Sci. Appl., Gakkōtosho, Tokyo, 1995, pp. 111–127.
[15] H. Ishii and K. Ishii, *An approximation scheme for motion by mean curvature with right-angle boundary condition*, SIAM J. Math. Anal., 33 (2001), pp. 369–389.
[16] H. Ishii, G. E. Pires, and P. E. Souganidis, *Threshold dynamics type approximation schemes for propagating fronts*, J. Math. Soc. Japan, 50 (1999), pp. 267–308.
[17] K. Ishii and T. Nakamura, *An error estimate to Bence–Merriman–Osher algorithm for motion by mean curvature*, Rev. Kobe Univ. Mercantile Marine, 51 (2003), pp. 105–115.
[18] F. Leoni, *Convergence of an approximation scheme for curvature-dependent motions of sets*, SIAM J. Numer. Anal., 39 (2001), pp. 1115–1131.
[19] P. Mascarenhas, *Diffusion Generated Motion by Mean Curvature*, Campus Report, Mathematics Department, University of California, Los Angeles, 1992.
[20] R. H. Nochetto, M. Paolini, and C. Verdi, *Optimal interface error estimates for the mean curvature flow*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 6 (1994), pp. 193–212.
[21] M. Paolini and C. Verdi, *Asymptotic and numerical analysis of the mean curvature flow with a space-dependent relaxation parameter*, Asymptotic Anal., 5 (1992), pp. 553–574.
[22] S. J. Ruuth, *Efficient Algorithms for Diffusion-Generated Motion by Mean Curvature*, Ph.D. Thesis, University of British Columbia, Vancouver, BC, Canada, 1996.
[23] L. Vivier, *Convergence of an approximation scheme for computing motions with curvature dependent velocities*, Differential Integral Equations, 13 (2000), pp. 1263–1288.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 867–888

IDENTIFICATION OF OPERATORS WITH BANDLIMITED SYMBOLS*

W. KOZEK† AND G. E. PFANDER‡

Abstract. Underspread and overspread operators are Hilbert–Schmidt operators with strictly bandlimited Kohn–Nirenberg symbols. In this paper, we prove a classical conjecture concerning the necessity of the underspread condition for the identifiability of such operator classes, and, in doing so, we exhibit a new uncertainty principle phenomenon in the time-frequency analysis of operators.

Key words. underspread and overspread operators, Gabor frame operators, bandlimited Kohn–Nirenberg symbol, spreading function

AMS subject classifications. 42C02, 47G02, 81S02

DOI. 10.1137/S0036141003433437

1. Introduction.

Identification of incompletely known linear operators based on the observation of a restricted number of input and corresponding output signals is an important goal in many applied sciences. In *communications engineering*, for instance, identifying the transmission channel can help to adjust signal synthesis at the transmitter and signal analysis at the receiver. This is possible in wired communications, since a linear *time-invariant* system is a convolution operator and—leaving numerical instability of deconvolution aside—is completely determined by its action on a single function.

Underspread and overspread operators on the other hand are *time-varying* Hilbert–Schmidt operators. They act on a space of d-dimensional signals, but the corresponding kernels of time-varying operators are essentially $2d$-dimensional so that a single observation of its action cannot uniquely determine the operator unless one has additional a priori knowledge of the operator class at hand in the form of certain constraints.

Hilbert–Schmidt operators can be represented as a weighted superposition of *translation operators* T_t, $t \in \mathbb{R}$, with $T_t f(x) = f(x-t)$, $x \in \mathbb{R}^d$, and *modulation operators* M_ν, $\nu \in \widehat{\mathbb{R}}$, with $M_\nu f(x) = f(x)e^{2\pi i \nu \cdot x}$, $x \in \mathbb{R}^d$, i.e., as an operator valued integral

$$H = \int\int \eta_H(t,\nu)\, T_t M_\nu \, dt\, d\nu\,. \tag{1.1}$$

Underspread and overspread operators are characterized by the property that the support of their *spreading function* η_H in (1.1) is contained in a rectangular parallelepiped. Such an operator is called *underspread* if the volume of the rectangular parallelepiped does not exceed one, and it is *overspread* otherwise, conditions which are intimately related to uncertainty phenomena in time-frequency analysis. The Kohn–Nirenberg symbol of a Hilbert–Schmidt operator is the symplectic Fourier transformation of the

*Received by the editors August 19, 2003; accepted for publication (in revised form) March 20, 2005; published electronically December 2, 2005. The research of this work was supported by the German Ministry of Education and Science (BMBF) under grant 01 BP 902.
http://www.siam.org/journals/sima/37-3/43343.html

†Siemens AG, Austria, Programm und Systementwicklung PSE, Erdberger Lände 26, A-1030, Vienna, Austria (werner.kozek@siemens.com).
‡School of Engineering and Science, International University Bremen, 28759 Bremen, Germany (g.pfander@iu-bremen.de).

respective spreading function, and, consequently, it is bandlimited in the case of an underspread or overspread operator.

The identification of underspread and overspread operators is important in various areas of electrical engineering and applied mathematics, including radar/sonar measurements and mobile radio communications, which we now briefly describe.

The principle of *radar/sonar measurements* is to send out a signal modulated onto an electromagnetic/acoustic wave and to deduce information about a (generally) moving target from an echo of the signal [Sko80]. In simple *range-Doppler estimation* the target is modeled as a pure time-frequency shift and distance ("range") and velocity ("Doppler-shift") are estimated. A more precise model of the physical phenomenon is the *doubly spread target model*. Here, the reflection is described as a continuous superposition of time-frequency shifts which arise since the target causes different reflections whose distance and velocity vary over a certain interval of the real-line. Unambiguous identification of the target was realized to depend on the product of the range and Doppler uncertainty, a fact that led to the terminology of underspread and overspread targets [Gre68]. Qualitatively speaking, overspread targets are those where the inherent uncertainty of the model is larger than the amount of information gathered by observing the reflected signal [VT71].

In *mobile radio communication*, the transmitted signal typically undergoes multiple reflections with different time-delay (corresponding to translation operators) and Doppler-shift (corresponding to modulation operators). The action of such channels on the signal can be modeled by underspread and overspread operators [VT71]. In order to obtain reliable communication, it is necessary to gather knowledge about channels by means of observations of transmitted and received signals to identify the channel operator (channel sounding) [MMH$^+$02, MGO03, LKS03].

Starting in the late 1950s, Thomas Kailath analyzed the identifiability of operators with restricted time and frequency spread [Kai59, Kai62, Kai63]. In engineering terms and without detailing a mathematical setup, Kailath proclaimed that a collection of communication channels which are characterized by having common maximum delay a and common maximum Doppler spread b would be identifiable by a single input signal if and only if $ab \leq 1$, i.e., if and only if the operator class is underspread. To prove the necessity of the underspread condition, Kailath provided ingenious arguments based on the comparison of the degrees of freedom of operators (which approximate underspread operators) and degrees of freedom of the output signal. To compare finite dimensions, Kailath used the theoretical construct of a bandlimited input signal with finite duration.

Being aware of the mathematical shortcomings of his approach, and understanding the work of Slepian, Landau, and Pollak on "the dimensions of the space of essentially time- and bandlimited functions" [SP61, LP61, LP62], Kailath conjectured that the underspread condition $ab \leq 1$ is necessary in general [Kai62].

We shall prove Kailath's conjecture in section 3 of this paper using the mathematical framework which is described in section 2. In section 4, we shall describe connections between the critical density in Gabor theory and the *critical spread* $ab = 1$ in the theory of operators with bandlimited symbols. We prove an identification result for Gabor frame operators in section 4.1 and relate this result and Kailath's conjecture to uncertainty principles in time-frequency analysis in section 4.2.

In section 5, we shall extend our identifiability result to higher dimensions and include classes of operators which have restricted but not necessarily rectangular spreading support. These results are based on the representation theory of the re-

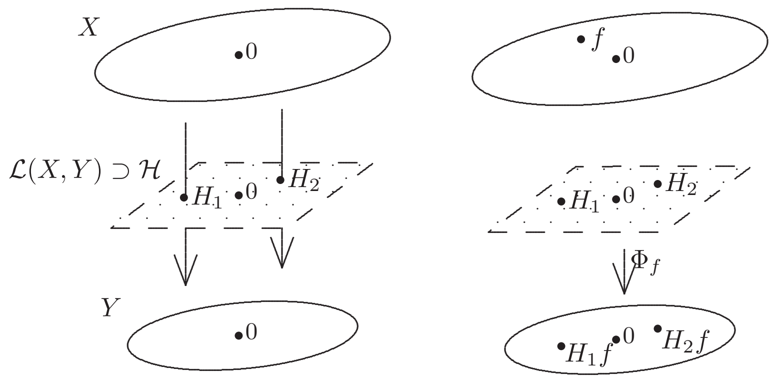


FIG. 2.1. *The goal of operator identification: Find $f \in X$ such that $\Phi_f : \mathcal{H} \longrightarrow Y$ is bounded and stable.*

duced Weyl–Heisenberg group, a fact which indicates close connections of our results to quantum mechanics.

2. Preliminaries. The goal of operator/system identification is to locate, for given normed linear spaces X and Y and a normed linear space of bounded linear operators $\mathcal{H} \subset \mathcal{L}(X, Y)$, an element $f \in X$ which induces a bounded and stable linear map $\Phi_f : \mathcal{H} \longrightarrow Y,\ H \mapsto Hf$ (see Figure 2.1). Consequently, we call $\mathcal{H}$ *identifiable by $f \in X$ if there exist $A, B > 0$ with $A\,\|H\|_{\mathcal{H}} \leq \|Hf\|_Y \leq B\,\|H\|_{\mathcal{H}}$ for all $H \in \mathcal{H}$.*

In sections 2.1 and 2.2, we shall describe the operator spaces $\mathcal{H}$, the domain spaces X, and the target spaces Y that are considered in this paper. In section 2.3, we shall present some techniques from Gabor analysis which will be used in this paper.

2.1. Hilbert–Schmidt operators with bandlimited symbols. We shall use Hilbert–Schmidt operators which act on the Hilbert space $L^2(\mathbb{R}^d)$ of complex valued and square integrable functions as a model of physical time-varying linear systems, as they appear in radar and in mobile communications [FL96, Yoo02, Str05].

A Hilbert–Schmidt operator $H : L^2(\mathbb{R}^d) \longrightarrow L^2(\mathbb{R}^d)$ is given by

$$Hf(x) = \int \kappa_H(x,t) f(t)\, dt = \int \kappa_H(x, x-t) f(x-t)\, dt \quad \text{(a.e.)},$$

with kernel $\kappa_H \in L^2(\mathbb{R}^{2d})$. The space of Hilbert–Schmidt operators $HS(L^2(\mathbb{R}^d))$ is itself a Hilbert space with inner product $\langle H_1, H_2 \rangle_{\mathrm{HS}} = \langle \kappa_{H_1}, \kappa_{H_2} \rangle_{L^2}$ [Die70, Gaa73].

Underspread and overspread operators are Hilbert–Schmidt operators which satisfy two constraints: First, they have restricted delay; i.e., $\kappa_H(x, x-t)$ vanishes for large $|t|$, say, for $|t| > \frac{a}{2} > 0$. Consequently, if f satisfies $\operatorname{supp} f \subseteq [0, T]$, then $\operatorname{supp} Hf \subseteq [-\frac{a}{2}, T+\frac{a}{2}]$. Second, underspread and overspread operators have the property that they are almost time-invariant, i.e., that their characteristics change only slowly over time. A comparison to the time-invariant convolution operators K given by $Kf(x) = \int \kappa_K(t) f(x-t)\, dt$—whose kernel κ_K is independent of the time variable x—leads us to quantify the *slow variance* of an operator H by means of a Paley–Wiener-type support condition on its spreading function which is given by

$$\eta_H(t, \nu) = \int \kappa_H(x, x-t) e^{-2\pi i \nu x}\, dx \quad \text{(a.e.)}.$$

In fact, underspread and overspread operators have the property that $\eta_H(t,\nu)$ vanishes for large $|\nu|$, say, for $|\nu| > \frac{b}{2} > 0$.

Combining the aforementioned time and frequency spread conditions on H leads to the condition

$$\operatorname{supp}\eta_H \subseteq Q_{a,b} = \left[-\frac{a}{2},\frac{a}{2}\right]^d \times \left[-\frac{b}{2},\frac{b}{2}\right]^d \tag{2.1}$$

for some $a,b>0$. An operator which satisfies (2.1) for $a,b>0$ is called underspread if $ab\le 1$ and overspread if $ab>1$.

The spreading function η_H of a Hilbert–Schmidt operator H leads to a representation of H as an operator valued integral by means of (1.1). Here and in the following, operator valued integrals shall be interpreted weakly, i.e., $\int H(z)dzf,\ f\in L^2(\mathbb{R})$, is given by means of

$$\left\langle \int H(z)dzf,\, g\right\rangle_{L^2(\mathbb{R}^d)} = \int \langle H(z)f,g\rangle_{L^2(\mathbb{R}^d)}\,dz \quad \text{for all } g\in L^2(\mathbb{R}^d).$$

Equation (1.1) illustrates that support restrictions on η_H reflect limitations on the maximal time and frequency shifts which the input signals undergo, a fact which emphasizes the usefulness of η_H in the time-frequency analysis of operators.

Note that condition (2.1) on a Hilbert–Schmidt operator H is a band-limitation on its Kohn–Nirenberg symbol σ_H which is given by

$$\sigma_H(x,\xi) = \int \kappa_H(x,x-y)\,e^{-2\pi i y\xi}\,dy = \int\!\!\int \eta_H(t,\nu)e^{2\pi i(x\nu - t\xi)}\,dt\,d\nu \quad \text{(a.e.)} \tag{2.2}$$

[KN65, Fol89].

To prove that a class of Hilbert–Schmidt operators whose spreading functions satisfy (2.1) for fixed $a,b>0$ with $ab\le 1$ is identifiable necessitates the use of *Shah distributions* (also called *combfunctions* or *delta trains*) $⧢_a = \sum_{n\in\mathbb{Z}^d}\delta_{an}$, $a>0$ as identifiers (see section 2.2). Since not all Hilbert–Schmidt operators in $\mathcal{L}(L^2(\mathbb{R}^d),L^2(\mathbb{R}^d))$ can be extended to a space of distributions containing the Shah distribution, we need to restrict ourselves to operators which satisfy a regularity condition on their kernels. Here, we choose Hilbert–Schmidt operators with kernels in the Feichtinger algebra $S_0(\mathbb{R}^{2d})$, a Banach algebra of test functions which is discussed in detail in section 2.2. In fact, if $\kappa_H\in S_0(\mathbb{R}^{2d})$, then the Hilbert–Schmidt operator H extends to $S_0'(\mathbb{R}^d)$ with $⧢_a\in S_0'(\mathbb{R}^d)$ [FZ98]. We set

$$\mathcal{H} = \left\{H\in HS(L^2(\mathbb{R}^d)):\ \kappa_H\in S_0(\mathbb{R}^{2d})\right\}, \tag{2.3}$$

and, as discussed above, we consider operator classes with restricted spreading; i.e., we consider operator classes of the form

$$\mathcal{H}_M = \{H\in\mathcal{H}:\ \operatorname{supp}\,\eta_H\subseteq M\}\,,\quad M\subset\mathbb{R}^d\times\widehat{\mathbb{R}}^d. \tag{2.4}$$

Note that $\mathcal{H}$ and $\mathcal{H}_M$, $M\subset\mathbb{R}^d\times\widehat{\mathbb{R}}^d$, are not closed as linear subspaces of the space of Hilbert–Schmidt operators, and that $\mathcal{H}_M\subseteq\mathcal{H}_{M'}$ if $M\subseteq M'$.

2.2. The Feichtinger algebra. Introduced in [Fei81], Feichtinger's Banach algebra $S_0(\mathbb{R}^d)$ of test functions gained popularity in the growing field of Gabor analysis, which is discussed in section 2.3. The usefulness of $S_0(\mathbb{R}^d)$ stems from the fact that it is the smallest Banach space allowing a meaningful time-frequency analysis, which, as a consequence, extends to its respectively large dual Banach space $S_0'(\mathbb{R}^d)$. In fact, the L^2-Fourier transform, the modulation operators M_ν, $\nu \in \widehat{\mathbb{R}}^d$, and the translation operators T_t, $t \in \mathbb{R}^d$, which are all unitary on the Hilbert space $L^2(\mathbb{R}^d)$, are isometric isomorphisms on the Feichtinger algebra $S_0(\mathbb{R}^d)$ and, therefore, on its dual $S_0'(\mathbb{R}^d)$. The Feichtinger algebra $S_0(\mathbb{R}^d)$ can be continuously embedded in any Banach space with these properties and which contains at least one, and therefore all, nontrivial Schwartz function [FZ98].

Note that we chose to work with the Banach spaces $S_0(\mathbb{R}^d)$ and $S_0'(\mathbb{R}^d)$ rather than with the Fréchet space of Schwartz functions $\mathcal{S}(\mathbb{R}^d) \subset S_0(\mathbb{R}^d)$ and its dual $\mathcal{S}'(\mathbb{R}^d) \supset S_0'(\mathbb{R}^d)$ of tempered distributions for the convenience of expressing continuity (boundedness) and openness (stability) of linear operators by means of norm inequalities. We would like to point out that the results in this paper are consequences of the structure of the identification problem at hand and not of topological subtleties.

There exist various ways of defining $S_0(\mathbb{R}^d)$, and equally many different equivalent norms for $S_0(\mathbb{R}^d)$. Here, we shall give a definition based on the space of Lebesgue measurable and integrable functions $L^1(\mathbb{R}^d)$, the space of Fourier transforms of functions in $L^1(\mathbb{R}^d)$, which is denoted by $A(\mathbb{R}^d)$ and which is equipped with the Banach-space structure of $L^1(\mathbb{R}^d)$ by means of $\|\widehat{f}\|_A = \|f\|_{L^1}$ [Kat76], and the space of absolutely summable sequences $l^1(\mathbb{Z}^d)$.

The *Feichtinger algebra* $S_0(\mathbb{R}^d)$ coincides with the *Wiener amalgam space* $W(A(\mathbb{R}^d), l^1(\mathbb{Z}^d))$. Consequently, we have $f \in S_0(\mathbb{R}^d)$ if and only if f is locally in $A(\mathbb{R}^d)$ with global decay of l^1-type; i.e., given any compactly supported $\psi \in A(\mathbb{R}^d)$ with $\sum_{n\in\mathbb{Z}^d} T_n\psi = 1$ we have $f \in S_0(\mathbb{R}^d)$ if and only if $\sum_{n\in\mathbb{Z}^d} \|f \cdot T_n\psi\|_A < \infty$, and

$$\|f\|_{S_0} = \sum_{n\in\mathbb{Z}^d} \|f \cdot T_n\psi\|_A$$

is a norm on $S_0(\mathbb{R}^d)$. Moreover, $S_0(\mathbb{R}^d)$ is a Banach algebra under convolution and pointwise multiplication.

The dual space $S_0'(\mathbb{R}^d)$ of the Feichtinger algebra satisfies $S_0'(\mathbb{R}^d) = W(A'(\mathbb{R}^d), l^\infty(\mathbb{Z}^d))$ since the class of compactly supported functions in $A(\mathbb{R}^d)$ is dense in $A(\mathbb{R}^d)$ [FG85]. Hence, $S_0'(\mathbb{R}^d)$ contains *Dirac's delta* $\delta : f \mapsto f(0)$ and *Shah distributions* $⧢_a = \sum_{n\in\mathbb{Z}^d} \delta_{an}$, where $\delta_{na} = T_{na}\delta$ and $a > 0$. We set $⧢ = ⧢_1$.

2.3. Gabor analysis. Most techniques applied in this paper originate from Gabor analysis.

Gabor introduced the concept of coherent states to electrical engineering independently of quantum theory [Gab46, Grö01]. Hence, we shall simply call the family

$$(g, a, b) = \{M_{kb}T_{la}g\}_{k,l\in\mathbb{Z}^d}$$

of coherent states a *Gabor system*.

One of the basic results of Gabor analysis is the fact that there exists $g \in L^2(\mathbb{R}^d)$ such that (g, a, b) is an orthonormal basis for $L^2(\mathbb{R}^d)$ if and only if $ab = 1$. For example, the Gabor system $(\mathbf{1}_{[0,a)}, a, b)$ is an orthonormal basis for $L^2(\mathbb{R}^d)$, where $\mathbf{1}_A(x) = 1$ for $x \in A$ and $\mathbf{1}_A(x) = 0$ else.

If $ab > 1$, the system (g, a, b) is not complete. However, if $ab > 1$, then there exists $g \in L^2(\mathbb{R})$ such that the (g, a, b)-synthesis map $D_g : l^2(\mathbb{Z}^2) \longrightarrow L^2(\mathbb{R})$, $\{c_{k,l}\} \mapsto \sum c_{k,l}M_{kb}T_{la}g$ is well-defined, bounded, and stable; i.e., (g, a, b) is a Riesz basis for its closed linear span, $\overline{\mathrm{span}(g,a,b)}$, in $L^2(\mathbb{R}^d)$; hence, there exist $A, B > 0$ such that

$$A\|\{c_{k,l}\}\|_{l^2} \leq \| \sum_{k,l\in\mathbb{Z}^d} c_{k,l} M_{kb} T_{la} g \|_{L^2} \leq B\|\{c_{k,l}\}\|_{l^2} \quad \text{for all } \{c_{k,l}\} \in l^2(\mathbb{Z}^{2d}). \tag{2.5}$$

For $ab < 1$, the system (g,a,b), $g \in L^2(\mathbb{R})$, is overcomplete; i.e., there exists a nontrivial coefficient sequence $\{c_{k,l}\} \in l^2(\mathbb{Z}^{2d}) \setminus \{0\}$ such that $\sum c_{k,l} M_{kb} T_{la} g = 0$ in $L^2(\mathbb{R}^d)$. Nevertheless, for an appropriate choice of g, e.g., g being a Gaussian, the (g,a,b)-analysis operator $C_g = D_g^* : L^2(\mathbb{R}) \longrightarrow l^2(\mathbb{Z}^2)$, $f \mapsto \{\langle f, M_{kb} T_{la} g\rangle\}$ is well-defined, bounded, and stable; i.e., (g,a,b) forms a frame for $L^2(\mathbb{R}^d)$; hence, there exist $A, B > 0$ such that

$$A\|f\|_{L^2}^2 \leq \sum |\langle f, M_{kb} T_{la} g\rangle|^2 \leq B\|f\|_{L^2}^2 \quad \text{for all } f \in L^2(\mathbb{R}^d)\,. \tag{2.6}$$

As a consequence of (2.6), every $f \in L^2(\mathbb{R}^d)$ has a stable representation

$$f = \sum_k \sum_l c_{k,l} M_{kb} T_{la} g \;\text{ in } L^2(\mathbb{R}^d)$$

in terms of the frame (g,a,b), where the coefficients $\{c_{k,l}\} \in l^2(\mathbb{Z}^2)$ can be chosen by means of inner products, i.e., $c_{k,l} = \langle f, M_{kb} T_{la} \gamma\rangle$, where (γ, a, b) is a so-called dual frame of (g,a,b).

More details on time-frequency analysis with some relevance to this paper can be found in [Grö01].

Operator-theoretic applications of Gabor theory as presented in this paper have drawn increasing interest in applied harmonic analysis; see, for example, [Dau88, HRT97, FK98, Koz98, RT98, Lab01, FN03, CG03, Hei03, GLM04].

3. Identification of underspread and overspread operators. We shall first prove Kailath's conjecture for operators acting on functions defined on the real-line, i.e., we choose $d = 1$. The identification problem is given by the operator space $\mathcal{H}_{Q_{a,b}}$, $a, b > 0$, which is defined in (2.3) and (2.4), where $M = Q_{a,b} = [-\frac{a}{2}, \frac{a}{2}]\times[-\frac{b}{2}, \frac{b}{2}]$. The linear space $\mathcal{H}_{Q_{a,b}}$ is equipped with the Hilbert–Schmidt norm and its elements map $X = S_0'(\mathbb{R})$ to $Y = L^2(\mathbb{R})$ [FK98].

The Lebesgue measure $a{\cdot}b$ of the set $Q_{a,b}$ plays a crucial role in determining the identifiability of $\mathcal{H}_{Q_{a,b}}$. The main result of our paper is the following.

THEOREM 3.1. *The set $\mathcal{H}_{Q_{a,b}}$ is identifiable; i.e., there is $f \in S_0'(\mathbb{R})$ such that $\Phi_f : \mathcal{H}_{Q_{a,b}} \longrightarrow L^2(\mathbb{R})$ is bounded and stable, where $\mathcal{H}_{Q_{a,b}}$ is equipped with the Hilbert–Schmidt norm if and only if $ab \leq 1$.*

First, we shall give a proof of the long-understood identifiability of $\mathcal{H}_{Q_{a,b}}$ for $ab \leq 1$.

3.1. Sufficiency of $\quad \leq 1$ for the identifiability of $\mathcal{H}_{\quad a,b}$. Our proof of the sufficiency of the underspread condition is based on the unitarity of the Zak transformations $Z_c : L^2(\mathbb{R}) \longrightarrow L^2(Q_{c,\frac{1}{c}})$, $c > 0$, which are defined by

$$Z_c f(t,\nu) = c^{\frac{1}{2}} \sum_{n\in\mathbb{Z}} f(t - cn) e^{2\pi i c n \nu} \quad \text{(for almost every) } (t,\nu) \in Q_{c,\frac{1}{c}}$$

and the following lemma.

Lemma 3.2. *For $H \in \mathcal{H}$ we have*

$$Z_c \circ H\text{Ш}_c(t,\nu) = c^{-\frac{1}{2}} \sum_{n\in\mathbb{Z}} \sum_{m\in\mathbb{Z}} \eta_H(t - cn, \nu - \tfrac{m}{c}) e^{2\pi i(\nu - \frac{m}{c})t}, \quad (t,\nu) \in Q_{c,\frac{1}{c}}\,.$$

Proof. For $x \in \mathbb{R}$ we have $H\text{Ш}_c(x) = \langle \text{Ш}_c, \kappa_H(x,\cdot)\rangle = \sum_{k\in\mathbb{Z}} \kappa_H(x, ck)$. Using in succession the Tonelli–Fubini theorem, the formula

$$\kappa_H(x,y) = \int \eta_H(x - y, \nu) e^{2\pi i \nu x}\, d\nu, \quad (x,y) \in \mathbb{R}^2,$$

two substitutions, and the Poisson summation formula [Grö01, p. 250], we obtain for $(t,\nu) \in Q_{c,\frac{1}{c}}$ that

$$\begin{aligned} Z_c \circ H\text{Ш}_c(t,\nu) &= c^{\frac{1}{2}} \sum_{l\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \kappa_H(t - cl, ck) e^{2\pi i cl\nu} \\ &= c^{\frac{1}{2}} \sum_{l\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} \int \eta_H(t - cl - ck, \omega) e^{2\pi i(cl\nu + \omega(t - cl))}\, d\omega \\ &\overset{\substack{\xi=\nu+\omega \\ n=k+l}}{=} c^{\frac{1}{2}} \sum_{n\in\mathbb{Z}} \sum_{m\in\mathbb{Z}} \int \eta_H(t - cn, \xi + \nu) e^{2\pi i(\xi+\nu)t} e^{-2\pi i cl\xi}\, d\xi \\ &= c^{-\frac{1}{2}} \sum_{n\in\mathbb{Z}} \sum_{m\in\mathbb{Z}} \eta_H(t - cn, \nu - \tfrac{m}{c}) e^{2\pi i(\nu - \frac{m}{c})t}. \qquad \square \end{aligned}$$

A standard periodization argument leads to the sufficiency of $ab \leq 1$ for the identifiability of $\mathcal{H}_{Q_{a,b}}$. In fact, the following theorem shows that for $f = \text{Ш}_a \in S_0'(\mathbb{R})$ we have $\Phi_f : \mathcal{H}_{Q_{a,b}} \longrightarrow L^2(\mathbb{R})$, where $\mathcal{H}_{Q_{a,b}}$ is equipped with the Hilbert–Schmidt norm, is bounded, and is stable whenever $ab \leq 1$.

Theorem 3.3. *The operator family $\mathcal{H}_M = \{H \in \mathcal{H} : \ \operatorname{supp} \eta_H \subseteq M\}$ can be identified with the identifier Ш_c if and only if the interior M° of M satisfies*

$$M^\circ \ \cap \bigcup_{(m,n)\in\mathbb{Z}^2\setminus\{(0,0)\}} \Big(M^\circ + (cn, \tfrac{m}{c}) \Big) = \emptyset, \tag{3.1}$$

i.e., if and only if M° is contained in a fundamental domain of the lattice $c\mathbb{Z} \times \frac{1}{c}\mathbb{Z}$. In particular, $\mathcal{H}_{Q_{a,b}}$, $a, b > 0$, is identifiable with Ш_c if and only if $a \leq c$ and $ab \leq 1$.

Note that Theorem 3.3 classifies all sets M with the property that $\mathcal{H}_M$ can be identified using the tempered distribution Ш_c, $c > 0$. No result regarding the necessity of the underspread condition $ab \leq 1$ for the identifiability of $\mathcal{H}_{Q_{a,b}}$ by any other $f \in S_0'(\mathbb{R})$ has been obtained.

Figure 3.1 is a picture proof of Theorem 3.3 for $M = Q_{c,\frac{1}{c}}$, $c > 0$. Details in the case $c = 1$ are given below.

Proof of Theorem 3.3. For ease of notation, we shall only provide a proof of Theorem 3.3 for $c = 1$. The general case follows from Theorem 5.4.

First, we show that if (3.1) holds, then Ш identifies $\mathcal{H}_M$. Set $Q = Q_{1,1}$ and let $A_{m,n} = M^\circ \cap \big(Q + (m,n)\big)$ and $B_{m,n} = A_{m,n} - (m,n) \subseteq Q$. Then $B_{m,n} \cap B_{m',n'} = \emptyset$ for $(m,n) \neq (m',n')$, since else $M^\circ \cap \big(M^\circ + (m - m', n - n')\big) \neq \emptyset$. Further, the spreading function η_H of each $H \in \mathcal{H}_M$ is continuous, and, therefore, $\eta_H(t,\nu) = 0$ for all $(t,\nu) \notin \bigcup_{m,n} A_{m,n}$. We conclude that $\{(t,\nu) \in Q : \ Z \circ H\text{Ш}(t,\nu) \neq 0\} \subseteq$

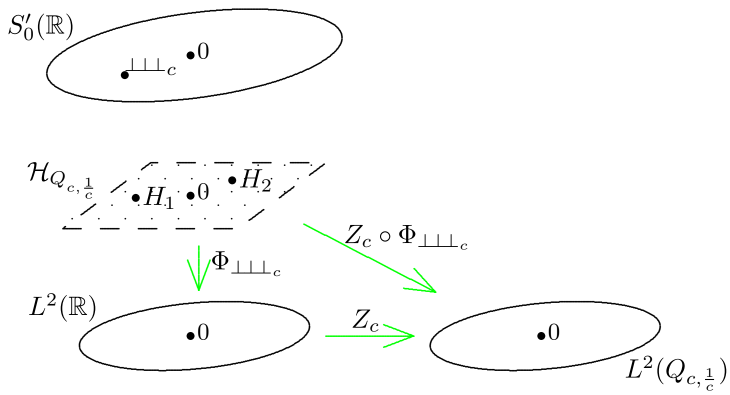


FIG. 3.1. *Sketch of the proof of the identifiability of $\mathcal{H}_{Q_{c,\frac{1}{c}}}$, $c > 0$, using as identifier Ш_c. The Zak transform Z_c is unitary and, therefore, bounded and stable, and $Z_c \circ \Phi_{\text{Ш}_c}$ maps $\mathcal{H}_{Q_{c,\frac{1}{c}}}$ into $L^2(Q_{c,\frac{1}{c}})$ and is bounded and stable as well. We conclude that $\Phi_{\text{Ш}_c}$ is bounded and stable on $\mathcal{H}_{Q_{c,\frac{1}{c}}}$, i.e., Ш_c identifies $\mathcal{H}_{Q_{c,\frac{1}{c}}}$.*

$\bigcup_{m,n} B_{m,n} \subseteq Q$. For $H \in \mathcal{H}_M$ we calculate

$$\begin{aligned}\|H\|_{HS} &= \|\eta_H\|_{L^2(\mathbb{R}\times\widehat{\mathbb{R}})} = \sum_{m,n} \|\eta_H\|_{L^2(A_{m,n})} = \sum_{m,n} \|T_{(-m,-n)}\eta_H\|_{L^2(B_{m,n})} \\ &= \sum_{m,n} \|Z \circ H\text{Ш}\|_{L^2(B_{m,n})} = \|Z \circ H\text{Ш}\|_{L^2(Q)} = \|H\text{Ш}\|_{L^2(\mathbb{R})} \\ &= \|\Phi_{\text{Ш}} H\|_{L^2(\mathbb{R})}\,; \qquad (3.2)\end{aligned}$$

and, by definition, $\mathcal{H}_M$ allows identification with identifier Ш.

Let us now assume that $M^\circ \cap \bigcup_{(m,n)\neq 0} M^\circ{+}(m,n) \neq \emptyset$ and show that $\mathcal{H}_M$ is not identifiable. In this case, there exists $(t_0, \nu_0) \in M^\circ \cap \bigcup_{(m,n)\neq 0} \big(M^\circ{+}(m,n)\big)$, $\frac{1}{2} > \epsilon > 0$, and $(n_0, m_0) \in \mathbb{Z}^{2d}$ with $B_\epsilon(t_0,\nu_0) \subset M^\circ \cap \big(M^\circ{+}(n_0,m_0)\big)$, where $B_\epsilon(t_0,\nu_0) = \{(t,\nu) : \|(t,\nu) - (t_0,\nu_0)\|_\infty < \epsilon\}$. Hence $B_\epsilon(t_0 - n_0, \nu_0 - m_0) \subset M^\circ$. Choose $0 \neq \widetilde{\eta} \in A(\mathbb{R}^{2d}) \subset S_0(\mathbb{R}^{2d})$ with supp $\widetilde{\eta} \subset B_\epsilon(t_0,\nu_0)$, and define $H \in \mathcal{H}_M$ by means of $\eta(t,\nu) = \widetilde{\eta}(t,\nu) - \widetilde{\eta}(t+n_0, \nu+m_0)e^{2\pi i t m_0} \not\equiv 0$, $(t,\nu) \in \mathbb{R}\times\widehat{\mathbb{R}}$. We obtain

$$\begin{aligned}Z\circ H\text{Ш}(t,\nu) &= \sum_{m,n\in\mathbb{Z}} \eta(t{-}n, \nu{-}m)e^{2\pi i(\nu-m)t} \\ &= \sum_{m,n\in\mathbb{Z}} \Big(\widetilde{\eta}(t{-}n, \nu{-}m) - \widetilde{\eta}(t{-}n{+}n_0, \nu{-}m{+}m_0)e^{2\pi i(t-n)m_0}\Big)\, e^{2\pi i(\nu-m)t} \\ &= \Big(\sum_{m,n\in\mathbb{Z}} \widetilde{\eta}(t{-}n, \nu{-}m)e^{2\pi i(\nu-m)t}\Big) \\ &\quad -\Big(\sum_{m,n\in\mathbb{Z}} \widetilde{\eta}(t{+}n_0{-}n, \nu{+}m_0{-}m)e^{2\pi i(t+n_0)m_0}e^{2\pi i(\nu-m)t}\Big) \;=\; 0\,.\end{aligned}$$

The injectivity of the Zak transformation implies $H\text{Ш} = 0$, contradicting the injectivity of $\Phi_{\text{Ш}}$ and therefore the identifiability of $\mathcal{H}_M$ by Ш. ∎

Note that equation (3.2) implies that $\Phi_{\text{Ш}}$, which is a priori defined on $\mathcal{H}_M = \{H \in \mathcal{H} : \text{ supp}\, \eta_H \subseteq M\} \subset \mathcal{H} \subset HS(L^2(\mathbb{R}))$, where M is a fundamental do-

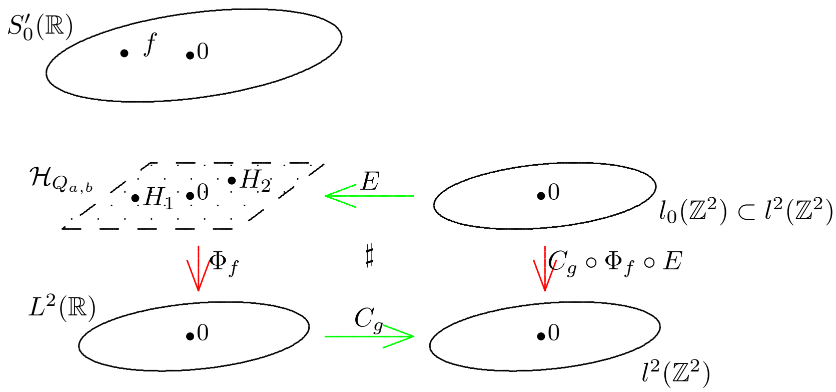


FIG. 3.2. *Sketch of the proof that $\mathcal{H}_{Q_{a,b}}$ is not identifiable if $ab > 1$. We show that for all $f \in S_0'(\mathbb{R})$, the bounded operator $C_g \circ \Phi_f \circ E$ is not stable. The synthesis operator E and the analysis operator C_g are stable, hence, stability of $C_g \circ \Phi_f \circ E$ must fail at Φ_f.*

main of $\mathbb{Z}\times\mathbb{Z}$, can be isometrically extended to its HS-closure $\overline{\mathcal{H}_M} = \{H \in HS(L^2(\mathbb{R})) : \operatorname{supp} \eta_H \subseteq M\}$. Certainly, not all $H \in HS(L^2(\mathbb{R}))$ extend in this fashion to $S_0'(\mathbb{R})$, and, hence, we must continue to focus our attention on operators with kernels in the Feichtinger algebra, i.e., on operator classes contained in $\mathcal{H} = \big\{H \in HS(L^2(\mathbb{R})) : \kappa_H \in S_0(\mathbb{R}^2)\big\}$.

3.2. Necessity of ≤ 1 for the identifiability of $\mathcal{H}_{a,b}$. We shall show that for $ab > 1$ and *every* $f \in S_0'(\mathbb{R})$, the well-defined operator $\Phi_f : \mathcal{H}_{Q_{a,b}} \longrightarrow L^2(\mathbb{R})$ is *not* stable.

To obtain this result, we shall equip $l_0(\mathbb{Z}^2)$ with the topology induced by the l^2-norm and use the fact that $ab > 1$ to construct a bounded and stable synthesis operator $E:\ l_0(\mathbb{Z}^2) \to \mathcal{H}_M$ in Lemma 3.4, and a bounded and stable (g, a', b')-analysis operator $C_g : L^2(\mathbb{R}) \longrightarrow l^2(\mathbb{Z}^2)$ in the proof of Theorem 3.6, with the property that the compositions

$$C_g \circ \Phi_f \circ E:\ l_0(\mathbb{Z}^2) \longrightarrow l^2(\mathbb{Z}^2), \quad f \in S_0'(\mathbb{R}),$$

are not stable. The stability of E and C_g implies that all operators $\Phi_f : \mathcal{H}_{Q_{a,b}} \longrightarrow L^2(\mathbb{R})$, $f \in S_0'(\mathbb{R})$, must not be stable, showing that $\mathcal{H}_{Q_{a,b}}$ is not identifiable for $ab > 1$ (see Figure 3.2).

We shall now construct the aforementioned synthesis operator E. For $ab > 1$, we choose $\lambda \in \mathbb{R}$ with $1 < \lambda^4 < a\,b$. Using a product-convolution operator $P : f \mapsto (f * \eta_1)\,\check{\eta}_2$ as prototype operator, we define the embedding operator E by means of

$$E:\ l_0(\mathbb{Z}^2) \to \mathcal{H}_M, \quad \{\sigma_{k,l}\} \mapsto \sum_{k,l} \sigma_{k,l}\ M_{k\lambda\alpha} T_{l\lambda\beta} P T_{-l\lambda\beta} M_{-k\lambda\alpha}\,,$$

where we chose $\alpha = \frac{1}{a}$ and $\beta = \frac{1}{b}$ for simplicity of notation. The choice of λ allows us to construct $P \in \mathcal{H}_{Q_{a,b}}$ in Lemma 3.4 such that $\{M_{k\lambda\alpha} T_{l\lambda\beta} P T_{-l\lambda\beta} M_{-k\lambda\alpha}\}_{k,l\in\mathbb{Z}}$ is a Riesz basis for its closed linear space in the Hilbert space of Hilbert–Schmidt operators, and, as consequence of (2.5), E is stable. In addition to the Riesz property, P is designed in Lemma 3.4 to satisfy a time-frequency localization property which will play a central role in the proof of our main result.

LEMMA 3.4. *Fix $\lambda > 1$ with $1 < \lambda^4 < a\,b$ and choose $\eta_1, \eta_2 \in \mathcal{S}(\mathbb{R})$ with values in $[0,1]$ and*

$$\eta_1(t) = \begin{cases} 1 & \text{for } |t| \le \frac{a}{2\lambda}, \\ 0 & \text{for } |t| \ge \frac{a}{2} \end{cases} \quad \text{and} \quad \eta_2(\nu) = \begin{cases} 1 & \text{for } |\nu| \le \frac{b}{2\lambda}, \\ 0 & \text{for } |\nu| \ge \frac{b}{2}. \end{cases}$$

The operator $P \in \mathcal{H}_{Q_{a,b}}$ defined by $\eta_P = \eta_1 \otimes \eta_2$ has the following properties:

(a) *The synthesis operator*

$$E: \ l_0(\mathbb{Z}^2) \to \mathcal{H}_M, \quad \{\sigma_{k,l}\} \mapsto \sum_{k,l} \sigma_{k,l}\ M_{k\lambda\alpha} T_{l\lambda\beta} P T_{-l\lambda\beta} M_{-k\lambda\alpha} \tag{3.3}$$

is well-defined, bounded, and stable.

(b) *The operator $P \in \mathcal{H}_M$ is a time-frequency localization operator in the following sense: There exist functions $d_1, d_2 : \mathbb{R} \to \mathbb{R}_0^+$ which decay rapidly at infinity and which have the property that for all $f \in S_0'(\mathbb{R})$ we have $|Pf(x)| \le \|f\|_{S_0'}\, d_1(x)$, $x \in \mathbb{R}$, and $|\widehat{Pf}(\xi)| \le \|f\|_{S_0'}\, d_2(\xi)$, $\xi \in \widehat{\mathbb{R}}$.*

Proof. (a) Observe that for any $(s,\omega) \in \mathbb{R}\times\widehat{\mathbb{R}}$ and $f \in S_0(\mathbb{R})$ we have

$$\begin{aligned} M_\omega T_s P T_{-s} M_{-\omega} f &= \int\!\!\int \eta_P(t,\nu) M_\omega T_s\, T_t M_\nu\, T_{-s} M_{-\omega} f \, dt\, d\nu \\ &= \int\!\!\int \eta_P(t,\nu) e^{2\pi i(\omega t - s\nu)} T_t M_\nu f\, dt\, d\nu. \end{aligned}$$

Hence, for E defined in (3.3) and any $\{\sigma_{k,l}\} \in l_0(\mathbb{Z}^2)$, we have $E\{\sigma_{k,l}\} \in \mathcal{H}_{Q_{a,b}}$ with

$$\eta_{E\{\sigma_{k,l}\}}(t,\nu) = \eta_P(t,\nu) \sum_{k,l\in\mathbb{Z}} \sigma_{k,l}\, e^{2\pi i(k\lambda\alpha t - l\lambda\beta\nu)}, \quad (t,\nu) \in \mathbb{R}\times\widehat{\mathbb{R}}. \tag{3.4}$$

We consider $l_0(\mathbb{Z}^2)$ as a subspace of $l^2(\mathbb{Z}^2)$ and observe that E is stable, since

$$\|E\{\sigma_{k,l}\}\|_{HS} = \|\eta_{E\{\sigma_{k,l}\}}\|_{L^2} \ge \|\eta_{E\{\sigma_{k,l}\}}\, \mathbf{1}_{[-\frac{a}{2\lambda},\frac{a}{2\lambda}]\times[-\frac{b}{2\lambda},\frac{b}{2\lambda}]}\|_{L^2} = \frac{ab}{\lambda^2}\, \|\{\sigma_{k,l}\}\|_{l^2}.$$

The boundedness of E follows from a similar calculation.

(b) For $f \in S_0(\mathbb{R})$ and $x \in \mathbb{R}$ we have

$$\begin{aligned} |Pf(x)| &= \left| \int\!\!\int \eta_P(t,\nu) e^{2\pi i\nu x} f(x-t)\, dt\, d\nu \right| \\ &= \left| \int \eta_2(\nu) e^{2\pi i\nu x}\, d\nu \right| \ \left| \int \eta_1(t) f(x-t)\, dt \right| \\ &\le |\widehat{\eta}_2(-x)|\ \ \|f\|_{S_0'} \|\eta_1\|_{S_0}. \end{aligned} \tag{3.5}$$

The function $d_1(x) = |\widehat{\eta}_2(-x)|\, \|\eta_1\|_{S_0}$ decays rapidly at infinity, i.e., $d_1(x) \to 0$ as $|x| \to \infty$ faster than any power of $\frac{1}{x}$, since $\widehat{\eta}_2 \in \mathcal{S}(\mathbb{R})$. Further, the inequality $|Pf(x)| \le \|f\|_{S_0'}\, d_1(x)$, $x \in \mathbb{R}$, extends to general $f \in S_0'(\mathbb{R})$, since $S_0(\mathbb{R})$ is w*-dense in $S_0'(\mathbb{R})$.

To establish a rapidly decaying bound on $|\widehat{Pf}|$, $f \in S_0'(\mathbb{R}^d)$, we first assume $f \in S_0(\mathbb{R})$ and calculate for $\xi \in \widehat{\mathbb{R}}$

$$\begin{aligned}
|\widehat{Pf}(\xi)| &= \left| \int \widehat{\eta}_2(-x) \int \eta_1(t) f(x-t)\, dt\, e^{-2\pi i \xi x}\, dx \right| \\
&= \left| \int \widehat{\eta}_2(-x) \int \widehat{\eta}_1(\gamma) \widehat{f}(\gamma) e^{-2\pi i x(\xi-\gamma)}\, d\gamma\, dx \right| \\
&= \left| \int \eta_2(\xi-\gamma) \widehat{\eta}_1(\gamma) \widehat{f}(\gamma)\, d\gamma \right| \qquad (3.6) \\
&\leq \|f\|_{S_0'} \|\eta_2(\xi - \cdot) \widehat{\eta}_1(\cdot)\|_{S_0}. \qquad (3.7)
\end{aligned}$$

The application of the theorem of Tonelli and Fubini to obtain (3.6) is valid for $f \in S_0(\mathbb{R})$, and the validity of (3.7) extends once more to general $f \in S_0'(\mathbb{R})$.

We claim that $d_2(\xi) = \|\eta_2(\xi - \cdot)\widehat{\eta}_1(\cdot)\|_{S_0}$ is rapidly decaying. Since the Feichtinger algebra $S_0(\mathbb{R})$ equals the Wiener amalgam space $W(A(\mathbb{R}), l^1(\mathbb{Z}))$, we choose $\widehat{\varphi} \in \mathcal{S}(\mathbb{R}) \subset A(\mathbb{R})$ with supp $\widehat{\varphi} \subseteq [-1,+1]$, and $\sum_{n\in\mathbb{Z}} \mathrm{T}_n \widehat{\varphi} \equiv 1$, and observe that $(\|\widehat{\eta}_1 \cdot \mathrm{T}_n \widehat{\varphi}\,\|_A)_{n\in\mathbb{Z}}$ decays rapidly, i.e., for any $k \in \mathbb{N}$ there exists $C_k > 0$ such that

$$\left\|g \cdot \mathrm{T}_n \widehat{\varphi}\right\|_A = \int \left| \int \eta_1(x) e^{-2\pi i x n} \varphi(t-x)\, dx \right| dt \leq C_k (1+n^2)^{-k/2},\ n \in \mathbb{Z} \qquad (3.8)$$

[Grö01, p. 228]. For $k \in \mathbb{N}$ we choose C_k satisfying (3.8) and calculate

$$\begin{aligned}
d_2(\xi) = \|\eta_2(\xi - \cdot)\widehat{\eta}_1(\cdot)\|_{S_0} &\leq C \sum_{n\in\mathbb{Z}} \|\,\mathrm{T}_n \widehat{\varphi}(\cdot)\ \eta_2(\xi - \cdot)\widehat{\eta}_1(\cdot)\|_A \\
&= C \sum_{\xi-1-\frac{b}{2}<n<\xi+1+\frac{b}{2}} \|\,\mathrm{T}_n \widehat{\varphi}(\cdot)\ \eta_2(\xi - \cdot)\widehat{\eta}_1(\cdot)\|_A \\
&\leq C\, \|\eta_2\|_A \sum_{\xi-1-\frac{b}{2}<n<\xi+1+\frac{b}{2}} \|\,\mathrm{T}_n \widehat{\varphi}(\cdot)\ \widehat{\eta}_1(\cdot)\|_A \\
&\leq C\, C_k\, \|\eta_2\|_A \left\lceil 2+b \right\rceil \left(1 + \min\left\{ \left\lceil \xi - 1 - \frac{b}{2} \right\rceil^2, \left\lfloor \xi + 1 + \frac{b}{2} \right\rfloor^2 \right\} \right)^{-k/2} \\
&\leq \widetilde{C} (1+\xi^2)^{-k/2}. \qquad \square
\end{aligned}$$

Lemma 3.5 is technical but of upmost importance in the proof of Theorem 3.6. It generalizes the fact that $m \times n$ matrices with $m < n$ have a nontrivial kernel and, therefore, are not stable, to operators acting on $l^2(\mathbb{Z}^2)$. In fact, the bi-infinite matrices $M = (m_{j',j})_{j',j\in\mathbb{Z}^2}$ considered in Lemma 3.5 are not dominated by its diagonal $m_{j,j}$—which would correspond to square matrices—but by a skewdiagonal $m_{j,\lambda j}$, with $\lambda > 1$.

LEMMA 3.5. *Given $M = (m_{j',j}) : l^2(\mathbb{Z}^2) \to l^2(\mathbb{Z}^2)$. If there exists a monotonically decreasing function $w : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ with $w(x) = O\left(x^{-2-\delta}\right)$, $\delta > 0$, and constants $\lambda > 1$ and $K_0 > 0$ with $|m_{i,j}| < w(\|\lambda j' - j\|_\infty)$ for $\|\lambda j' - j\|_\infty > K_0$, then M is not stable.*

Proof. First, we show that if $w : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ with $w(x) = O\left(x^{-2-\delta}\right)$ is monotonically decreasing, then

$$\sum_{K\geq 1} K \sum_{k\geq K} k\, w(k)^2 < \infty\,. \qquad (3.9)$$

Inequality (3.9) is proven using the Riemann integral criterium for sums. To this end, we pick continuous $v \in L^\infty(\mathbb{R}^+)$ with $w(x) \leq v(x) x^{-2-\delta}$ and observe that

$$\sum_{K\geq 1} K \sum_{k\geq K} k\, w(k)^2 \leq \int_0^\infty x \int_x^\infty y\, w(y)^2\, dy\, dx \leq \int_0^\infty x \int_x^\infty v(y)^2 y^{-3-2\delta}\, dy\, dx$$
$$\leq \frac{\|v\|_\infty^2}{2+2\delta} \int_0^\infty x^{-1-2\delta}\, dx < \infty\,.$$

Now, we shall use (3.9) to show that $\inf_{x\in l^2(\mathbb{Z}^2)}\{\frac{\|Mx\|_{l^2}}{\|x\|_{l^2}}\} = 0$. To this end, fix $\epsilon > 0$ and pick $K_1 > K_0$ with

$$\sum_{K\geq K_1} K \left(\sum_{k\geq K} k\, w(k)^2 \right) \leq 2^{-6}\epsilon^2\,.$$

Pick $N \in \mathbb{N}$ with $\widetilde{N} := \lceil \frac{N}{\lambda} \rceil + K_1 < N$ and define

$$\widetilde{M} = (m_{j',j})_{\|j'\|_\infty \leq \widetilde{N}, \|j\| \leq N} : \mathbb{C}^{(2N+1)^2} \to \mathbb{C}^{(2\widetilde{N}+1)^2}.$$

The matrix $\widetilde{M}$ has a nontrivial kernel since $(2\widetilde{N}+1)^2 < (2N+1)^2$, so we can choose $\widetilde{x} \in \mathbb{C}^{(2N+1)}$ with $\|\widetilde{x}\|_2 = 1$ and $\widetilde{M}\widetilde{x} = 0$. Define $x \in l^2(\mathbb{Z}^2)$ according to $x_j = \widetilde{x}_j$ if $\|j\|_\infty \leq N$ and $x_j = 0$ otherwise.

By construction we have $\|x\|_{l^2} = 1$, and $(Mx)_{j'} = 0$ for $\|j'\|_\infty \leq \widetilde{N}$.

To estimate $(Mx)_{j'}$ for $\|j'\|_\infty > \widetilde{N}$, we fix $K > K_1$ and $j' \in \mathbb{Z}^d$ with $\|j'\|_\infty = \lceil \frac{N}{\lambda} \rceil + K$. We have $\|\lambda j'\|_\infty \geq N + K\lambda$ and $\|\lambda j' - j\|_\infty \geq K\lambda \geq K$ for all $j \in \mathbb{Z}^d$ with $\|j\|_\infty \leq N$, and, therefore,

$$|(Mx)_j'|^2 = \left| \sum_{\|j\|_\infty \leq N} m_{j',j} x_j \right|^2 \leq \|x\|_2^2 \sum_{\|j\|_\infty \leq N} |m_{j',j}|^2$$
$$\leq \sum_{\|j\|_\infty \leq N} w(\|\lambda j' - j\|_\infty)^2 \leq \sum_{\|j\|_\infty \geq K} w(\|j\|_\infty)^2$$
$$= 2^2 \sum_{k\geq K} 2k\, w(k)^2 = 2^3 \sum_{k\geq K} k\, w(k)^2.$$

Finally, we can compute

$$\|Mx\|_{l^2}^2 = \sum_{j'\in\mathbb{Z}^d} |(Mx)_j'|^2 \quad = \sum_{\|j'\|_\infty \geq \lceil \frac{N}{\lambda} \rceil + K_1} |(Mx)_{j'}|^2$$
$$= \; 2^3 \sum_{\|j'\|_\infty \geq \lceil \frac{N}{\lambda} \rceil + K_1} \sum_{k \geq \|j'\|_\infty} k\, w(k)^2 \leq 2^6 \sum_{K \geq \lceil \frac{N}{\lambda} \rceil + K_1} K \sum_{k\geq K} k\, w(k)^2 \leq \epsilon^2$$

and obtain $\|Mx\|_{l^2} \leq \epsilon$. Since ϵ was chosen arbitrarily and $\|x\|_{l^2} = 1$, we have $\inf_{x\in l^2(\mathbb{Z}^2)}\{\frac{\|Mx\|_{l^2}}{\|x\|_{l^2}}\} = 0$ and M is not stable. $\square$

Now all pieces are in place to state and prove the main contribution of this paper.

THEOREM 3.6. *For $a, b > 0$ with $ab > 1$, $\mathcal{H}_{Q_{a,b}}$ is not identifiable.*

Proof. Fix $a, b > 0$ with $ab > 1$ and choose λ, η_1, η_2, P, and E as in Lemma 3.4.

To construct the aforementioned stable (g, a', b')-analysis operator C_g, we choose as Gabor atom the Gaussian $g_0 : \mathbb{R} \to \mathbb{R}^+, \quad x \mapsto e^{-\pi x^2}$. Lyubarskii [Lyu92], and Seip [Sei92], and Seip and Wallstén [SW92] have shown that $(g_0, a', b') = \{M_{ka'}T_{lb'}g_0\}$ is a frame for any $a', b' > 0$ with $a'b' < 1$, and, hence, we conclude that the analysis map given by

$$C_{g_0} : L^2(\mathbb{R}) \to l^2(\mathbb{Z}^2), \quad f \mapsto \big\{\langle f, M_{k\lambda^2\alpha}T_{l\lambda^2\beta}g_0\rangle\big\}_{k,l}$$

is bounded and stable since $\lambda^2\beta \cdot \lambda^2\alpha = \frac{\lambda^4}{ab} < 1$.

Let us now fix $f \in S_0'(\mathbb{R})$ and consider the composition

$$\begin{array}{ccccccc} l_0(\mathbb{Z}^2) & \overset{E}{\to} & \mathcal{H}_M & \overset{\Phi_f}{\to} & L^2 & \overset{C_{g_0}}{\to} & l^2(\mathbb{Z}^2) \\ \{\sigma_{k,l}\} & \mapsto & E\{\sigma_{k,l}\} & \mapsto & E\{\sigma_{k,l}\}\, f & \mapsto & \big\{\, \langle\, E\{\sigma_{k,l}\}\, f,\ M_{k'\lambda^2\alpha}T_{l'\lambda^2\beta}g_0 \,\rangle\, \big\}_{k',l'} \, . \end{array}$$

The bi-infinite matrix

$$M = \Big(m_{k',l',k,l}\Big) = \Big(\langle\, M_{k\lambda\alpha}T_{l\lambda\beta}PT_{-l\lambda\beta}M_{-k\lambda\alpha}f,\ M_{k'\lambda^2\alpha}T_{l'\lambda^2\beta}g_0 \,\rangle\Big)$$

represents the operator $C_{g_0} \circ \Phi_f \circ E$ with respect to the canonical basis of $l^2(\mathbb{Z}^2)$, since

$$\begin{aligned} \Big(C_{g_0} \circ \Phi_f \circ E\ \{\sigma_{k,l}\}\Big)_{k',l'} &= \Big\langle \sum_{k,l} \sigma_{k,l} M_{k\lambda\alpha}T_{l\lambda\beta}PT_{-l\lambda\beta}M_{-k\lambda\alpha}f,\ M_{k'\lambda^2\alpha}T_{l'\lambda^2\beta}g_0 \Big\rangle \\ &= \sum_{k,l} \langle\, M_{k\lambda\alpha}T_{l\lambda\beta}PT_{-l\lambda\beta}M_{-k\lambda\alpha}f,\ M_{k'\lambda^2\alpha}T_{l'\lambda^2\beta}g_0 \,\rangle\, \sigma_{k,l} \\ &= \sum_{k,l} m_{k',l',k,l}\ \sigma_{k,l}\, . \end{aligned}$$

In order to use Lemma 3.5 to show that M, and, therefore, $C_{g_0} \circ \Phi_f \circ E$, is not stable, we have to obtain bounds on the matrix entries of M. Lemma 3.4(b) will provide us with these bounds. In fact, for $k, l, k', l' \in \mathbb{Z}$, we have

$$\begin{aligned} |m_{k',l',k,l}| &= \big|\langle\, M_{k\lambda\alpha}T_{l\lambda\beta}PT_{-l\lambda\beta}M_{-k\lambda\alpha}f,\ M_{k'\lambda^2\alpha}T_{l'\lambda^2\beta}\, g_0 \,\rangle\big| \\ &\le \langle\, T_{l\lambda\beta}\, |PT_{-l\lambda\beta}M_{-k\lambda\alpha}f|\, ,\ T_{l'\lambda^2\beta}\, |g_0|\, \rangle \\ &\le \|f\|_{S_0'}\ d_1 * g_0\, (\lambda\beta(\lambda l' - l))\, , \end{aligned}$$

and

$$\begin{aligned} |m_{k',l',k,l}| &= \big|\langle\, T_{k\lambda\alpha}M_{-l\lambda\beta}\big(PT_{-l\lambda\beta}M_{-k\lambda\alpha}f\big)\widehat{\ },\ T_{k'\lambda^2\alpha}M_{-l'\lambda^2\beta}\, \widehat{g_0} \,\rangle\big| \\ &\le \langle\, T_{k\lambda\alpha}\, \big|\big(PT_{-l\lambda\beta}M_{-k\lambda\alpha}f\big)\widehat{\ }\big|\ ,\ T_{k'\lambda^2\alpha}\, |g_0|\, \rangle \\ &\le \|f\|_{S_0'}\ d_2 * g_0(\lambda\alpha(\lambda k' - k)). \end{aligned}$$

In these calculations, we used that $g_0 \ge 0$, $\widehat{g_0} = g_0$, and $g_0(-x) = g_0(x)$, as well as the Parseval–Plancherel identity. Since d_1, d_2, and g_0 decay rapidly, so do $d_1 * g_0$ and $d_2 * g_0$. We set

$$w(x) = \|f\|_{S_0'}\ \max\big\{d_1 * g_0(\lambda\beta x),\ d_1 * g_0(-\lambda\beta x),\ d_2 * g_0(\lambda\alpha x),\ d_2 * g_0(-\lambda\alpha x)\big\}$$

and obtain $|m_{k',l',k,l}| \leq w\big(\max\{|\lambda k' - k|,\ |\lambda l' - l|\}\big)$ with $w(x) = O\,(x^{-n})$ for $n \in \mathbb{N}$. Lemma 3.5 implies that M is not stable, and, by construction, we can conclude that $C_{g_0} \circ \Phi_f \circ E$, and thus Φ_f is not stable. ∎

Note that Lemma 3.5 is crucial for the understanding of Theorem 3.6: For any $f \in S_0'$, the operator $C_g \circ \Phi_f \circ E : l_0(\mathbb{Z}^2) \longrightarrow l^2(\mathbb{Z}^2)$, and, therefore, the operator $\Phi_f : \mathcal{H}_{Q_{a,b}} \longrightarrow L^2(\mathbb{R})$, is not stable as a result of the nonquadratic structure of the canonical matrix representation of $C_g \circ \Phi_f \circ E$. The validity of Lemma 3.5 does not depend on the choice of (reasonable) topologies on domain and range; in fact, a more general version of Lemma 3.5 can be found in [Pfa05].

4. Gabor frame operators, underspread operators, and uncertainty. The proof of Kailath's conjecture in section 3 relies strongly on the existence of a Schwartz function $g \in \mathcal{S}(\mathbb{R})$ such that (g, a, b) is a Gabor frame for given $a, b > 0$ with $ab < 1$. In section 4.1 we shall discuss the role of the critical density $ab = 1$ in the identification of Gabor frame operators and analogies of underspread and Gabor frame operators. Interpretations of the results in sections 3 and 4.1 as consequences of uncertainty in time-frequency analysis are given in section 4.2.

As in section 3, we choose to work in section 4 in the one-dimensional setting.

4.1. Identification of Gabor frame operators. For appropriate $g, h \in L^2(\mathbb{R})$, e.g., for $g, h \in S_0(\mathbb{R})$, and $a, b > 0$, the *Gabor frame operator* $S_{g,h}^{a,b} : L^2(\mathbb{R}) \longrightarrow L^2(\mathbb{R})$ is given by

$$S_{g,h}^{a,b} f = D_h \circ C_g\ f = \sum_{k,l \in \mathbb{Z}} \langle f, M_{kb} T_{la} g \rangle\, M_{kb} T_{la} h\,, \quad f \in L^2(\mathbb{R})\,.$$

Let us compare the spreading function representation of Hilbert–Schmidt operators given in (1.1) with *Janssen's representation* of the Gabor frame operator, which is

$$S_{g,h}^{a,b} f = (ab)^{-1} \sum_{m,n \in \mathbb{Z}} \big\langle h, M_{\frac{m}{a}} T_{\frac{n}{b}} g \big\rangle\, M_{\frac{m}{a}} T_{\frac{n}{b}} f\,, \quad f \in L^2(\mathbb{R})$$

[Jan95, Grö01]. Both types of operators are superpositions of time-frequency shifts, and, hence, we shall refer to the tempered distribution

$$(ab)^{-1} \sum_{m,n \in \mathbb{Z}} \big\langle h, M_{\frac{m}{a}} T_{\frac{n}{b}} g \big\rangle\ \delta_{\frac{n}{b}} \otimes \delta_{\frac{m}{a}} \in S_0'(\mathbb{R} \times \widehat{\mathbb{R}})$$

as spreading function of the Gabor frame operator $S_{g,h}^{a,b}$.

On a formal level, the relationship between Gabor frame operators and underspread and overspread operators is striking: the spreading functions of Gabor frame operators are supported (as distributions) on a full rank lattice $\frac{1}{b}\mathbb{Z} \times \frac{1}{a}\mathbb{Z}$ in the time-frequency plane $\mathbb{R} \times \widehat{\mathbb{R}}$, whereas the spreading functions of underspread and overspread operators are supported on a fundamental domain of such a lattice (see Figure 4.1). The duality of *compact* and *discrete* locally compact abelian groups suggests that results in the theory of underspread and overspread operators might lead to analogous results in Gabor analysis, and vice versa.

The correspondence of underspread and overspread operators to Gabor frame operators has not yet been fully explored. To initiate research in this direction, we shall show in Theorem 4.1 that identifiability of a canonically defined class of Gabor frame operators with fixed lattice $a\mathbb{Z} \times b\mathbb{Z}$ is equivalent to the existence of $f \in L^2(\mathbb{R})$ such that (f, a, b) is a Gabor frame for $L^2(\mathbb{R})$. As in section 3, we need to define a

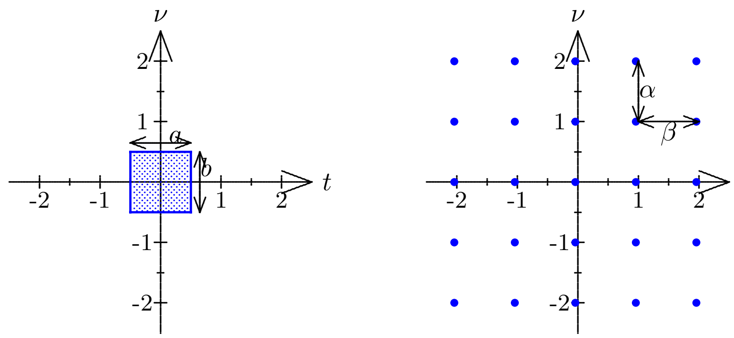


FIG. 4.1. *Support of the spreading symbol of an underspread or overspread operator and distributional support of the spreading symbol of a Gabor frame operator.*

domain X and classes of Gabor frame operators $\mathcal{S}^{a,b}$ with some care in order to have X sufficiently large to allow identification for $ab \leq 1$, and X small enough to allow for an easy proof of the nonidentifiability in case of $ab > 1$.

We choose as domain the Wiener space $W(\mathbb{R})$, i.e.,

$$X = W(\mathbb{R}) = W(L^\infty(\mathbb{R}), l^1(\mathbb{Z})) = \left\{ f \in L^2(\mathbb{R}) : \|f\|_W = \sum_{k\in\mathbb{Z}} \|f \cdot 1_{[k,k+1)}\|_\infty < \infty \right\},$$

as range, once more, $Y = L^2(\mathbb{R})$, and, for $a, b > 0$, we consider the operator class

$$\mathcal{S}^{a,b} = \left\{ S^{a,b}_{g,h} : \ g \in L^2(\mathbb{R}), \ h \in W(\mathbb{R}) \right\} \quad \text{with} \quad \|S^{a,b}_{g,h}\|_{\mathcal{S}^{a,b}} = \left\| \left\{ \left\langle h, M_{\frac{m}{a}} T_{\frac{n}{b}} g \right\rangle \right\} \right\|_{l^2}.$$

We have

$$\|S^{a,b}_{g,h} f\|_{L^2} \leq \sqrt{(a+1)(b+1)} \ \|f\|_W \ \|S^{a,b}_{g,h}\|_{\mathcal{S}^{a,b}}$$

[Grö01, p. 107], and, therefore, $\mathcal{S}^{a,b} \subset \mathcal{L}\left(W(\mathbb{R}), L^2(\mathbb{R})\right)$ and $\{\Phi_f : \ f \in W(\mathbb{R})\} \subset \mathcal{L}\left(\mathcal{S}^{a,b}, L^2(\mathbb{R}^d)\right)$, where $\Phi_f : S^{a,b}_{g,h} \mapsto S^{a,b}_{g,h} f$.

THEOREM 4.1. *$\mathcal{S}^{a,b}$ is identifiable if and only if $ab \leq 1$. Moreover, for any a, b with $ab > 1$ and any $f \in W(\mathbb{R})$ there exist $g \in L^2(\mathbb{R})$ and $h \in W(\mathbb{R})$ such that $S^{a,b}_{g,h} f = 0$.*

Note that identification of $\mathcal{S}^{a,b}$ does not require us to uncover g and h in $S^{a,b}_{g,h}$, but only to obtain the coefficients $\left\{ \left\langle h, M_{\frac{n}{a}} T_{\frac{m}{b}} g \right\rangle \right\}$ in Janssen's representation of the Gabor frame operator $S^{a,b}_{g,h}$.

Proof of Theorem 4.1. To show the identifiability of $\mathcal{S}^{a,b}$ for $ab \leq 1$, we use the fact that for any $ab \leq 1$ there exists $f \in W(\mathbb{R})$ such that $(f, a, b) = \{M_{kb} T_{la} f\}$ is a frame for $L^2(\mathbb{R})$. For example, if $ab < 1$, we may choose the Gaussian $f = g_0 : \mathbb{R} \to \mathbb{R}^+$, $x \mapsto e^{-\pi x^2}$, with $g_0 \in \mathcal{S}(\mathbb{R}) \subset W(\mathbb{R})$ and for $ab = 1$ we could choose $f = \mathbf{1}_{[0,a)} \in W(\mathbb{R})$. The Ron–Shen duality principle implies that (f, a, b) is a frame for $L^2(\mathbb{R})$ if and only if $(f, \frac{1}{b}, \frac{1}{a})$ is a Riesz basis for its closed linear span in $L^2(\mathbb{R})$, i.e., if and only if there exist $A, B > 0$ such that for all $\{d_{m,n}\} \in l^2(\mathbb{Z}^2)$ we have

$$A\|\{d_{m,n}\}\|_{l^2} \leq \| \sum_{m,n\in\mathbb{Z}} d_{m,n} M_{\frac{m}{a}} T_{\frac{n}{b}} f\|_{L^2} \leq B\|\{d_{m,n}\}\|_{l^2} \tag{4.1}$$

[RS97, Grö01]. Replacing $\{d_{m,n}\}$ by $\{\langle h, M_{\frac{m}{a}} T_{\frac{n}{b}} g\rangle\} \in l^2(\mathbb{Z}^2)$ in (4.1) shows that any f with (f, a, b) is a frame for $L^2(\mathbb{R})$ identifies $\mathcal{S}^{a,b}$.

We shall now show that for any $a, b > 0$ with $ab > 1$ and any $f \in W(\mathbb{R})$ there exist $g \in L^2(\mathbb{R})$ and $h \in W(\mathbb{R})$ such that $S^{a,b}_{g,h} \in \mathcal{S}^{a,b} \setminus \{0\}$ and $S^{a,b}_{g,h} f = 0$, contradicting that f identifies $\mathcal{S}^{a,b}$. Fix $a, b > 0$ with $ab > 1$ and $f \in W(\mathbb{R})$ and pick $g \in L^2(\mathbb{R})$ such that $g \perp \overline{\text{span}(f, a, b)}$, and, therefore, $f \perp \overline{\text{span}(g, a, b)}$. Let $h = g_0 \in W(\mathbb{R})$ be the Gaussian defined above and observe that $(g_0, \frac{1}{b}, \frac{1}{a})$ is a frame for $L^2(\mathbb{R})$ since $\frac{1}{ab} < 1$. Hence, $\{\langle h, M_{\frac{m}{a}} T_{\frac{n}{b}} g\rangle\} = \{e^{2\pi i \frac{mn}{ab}} \overline{\langle g, M_{-\frac{m}{a}} T_{-\frac{n}{b}} h\rangle}\} \in l^2(\mathbb{Z}^2) \setminus \{0\}$, i.e., $S^{a,b}_{g,h} \in \mathcal{S}^{a,b}$, $\|S^{a,b}_{g,h}\| \neq 0$, but $S^{a,b}_{g,h} f = \sum \langle f, T_{am} M_{bn} g\rangle\, T_{am} M_{bn} h = 0$. □

4.2. Uncertainty. Theorem 4.1 illustrates a strong relationship of critical density in Gabor analysis to the identification of canonically defined classes of Gabor frame operators. The critical density phenomenon in Gabor analysis is well known to be rooted in uncertainty in time-frequency analysis:

- functions cannot be arbitrarily well localized simultaneously in time and frequency, i.e., in phase space, and we can therefore exclude the possibility that there exist Gabor systems (g, a, b) which are Riesz bases for $L^2(\mathbb{R})$ if $ab < 1$, and
- functions cannot represent an area in phase space of volume larger than one in the sense that one cannot construct complete Gabor frames (g, a, b) for $L^2(\mathbb{R})$ if $ab > 1$.

Due to the first of the two limitations described above, Folland refers to a rectangle of volume one in phase space as a "minimal rectangle in phase space" [FS97].

Theorem 3.1 describes a new interpretation of minimal rectangles which plays a role in the time-frequency analysis of operators: an operator, whose spreading symbol is known to be supported in a rectangle in the time-frequency plane, can be identified if the rectangle has volume less than or equal to one, and cannot be identified if the rectangle has volume greater than one. Note that this phenomenon is not a direct consequence of the fact that we cannot construct functions which are arbitrarily well localized in phase space, since, in fact, there exist no support restrictions for the construction of operator symbols or spreading functions in phase space.

Theorems 3.1 and 4.1 can also be viewed as pull-backs of the critical density phenomenon of "phase space expansions" as described in [Lan93] to operator theory. Any operator output signal can carry only a restricted amount of time-frequency structured information, and therefore any output signal can be used only to resolve a limited amount of information from an operator. Theorem 3.1 illustrates that this amount of information corresponds to a *minimal rectangle in the spreading domain*. Theorem 4.1 shows that the resolvable amount of information of operators, whose spreading functions have discrete distributional support contained in a lattice $\frac{1}{b}\mathbb{Z}\times\frac{1}{a}\mathbb{Z}$, is connected to the sparsity of the lattice. In fact, all information inherent in such an operator can be resolved using a single test signal if and only if $ab \leq 1$. Note that in the latter case, the Kohn–Nirenberg symbol, which is the symplectic Fourier transformation of the spreading function, is $a\times b$ periodic, i.e., is the periodization of a function supported on a *minimal rectangle of size* $ab \leq 1$ *in phase space*.

We would like to add that the physical interpretation of the uncertainty principle as a limit to the achievable precision when measuring position and momentum of quantum mechanical objects parallels the identifiability result for underspread and overspread operators nicely, since the latter tells us that we will not be able to identify an overspread operator no matter how smartly a signal is chosen to test the system.

The uncertainty principle phenomena discussed above, among others, can be found in [Fef83, Dau92, Lan93, BHW98, Grö01, Grö03].

5. Generalized spreading constraints. We shall now extend the results stated in section 3 to higher dimensions and to nonrectangular spreading support sets.

Similarly to the one-dimensional case, we have

$$\mathcal{H} = \mathcal{H}(\mathbb{R}^d) = \left\{H \in HS(L^2(\mathbb{R}^d)) :\ \kappa_H \in S_0(\mathbb{R}^{2d})\right\} \subset \mathcal{L}\left(S_0'(\mathbb{R}^d), L^2(\mathbb{R}^d)\right).$$

Once more, we shall use a Zak transformation, namely, $Z : L^2(\mathbb{R}^d) \longrightarrow L^2(\mathbb{R}^d \times \widehat{R}^d)$, $Zf(t,\nu) = \sum_{n\in\mathbb{Z}^d} f(t-n)e^{2\pi i n\cdot\nu}$ for almost every $(t,\nu) \in Q = Q_{1,1}$, and the Shah distribution $⊥⊥⊥ = ⊥⊥⊥_1$. Adjusting Lemma 3.2 accordingly, we obtain

$$Z \circ H⊥⊥⊥(t,\nu) = \sum_{n,m\in\mathbb{Z}^d} \eta_H(t-n,\nu-m)e^{2\pi i(\nu-m)\cdot t} \quad \text{for all } (t,\nu) \in \left[-\frac{1}{2},\frac{1}{2}\right]^{2d}, \tag{5.1}$$

an identity which leads immediately to the following.

THEOREM 5.1. *$\mathcal{H}_M = \{H \in \mathcal{H} :\ \operatorname{supp} \eta_H \subseteq M\}$ is identifiable with identifier $⊥⊥⊥$ if and only if $M^\circ \cap \bigcup_{(m,n)\in\mathbb{Z}^{2d}\setminus(0,0)} \left(M^\circ + (m,n)\right) = \emptyset$.*

The proof of Theorem 5.1 is similar to the proof of Theorem 3.3 and is therefore omitted.

A straightforward generalization of either Theorem 3.3 or Theorem 5.1 leads to the identifiability of $\mathcal{H}_{\mathcal{D}Q}$, $Q = Q_{1,1} = [-\frac{1}{2},\frac{1}{2}]^{2d}$ in the case that $\mathcal{D}$ is a diagonal matrix with diagonal $(a_1,\ldots,a_d,\frac{1}{a_1},\ldots,\frac{1}{a_d}) \in (\mathbb{R}^+)^{2d}$. This observation leads us to the question for which general diagonal or nondiagonal, volume preserving matrices $\mathcal{A} \in SL(2d,\mathbb{R})$ is the operator space $\mathcal{H}_{\mathcal{A}Q}$ identifiable.

The underlying idea of obtaining identifiability results on $\mathcal{H}_{\mathcal{A}Q}$ for nondiagonal matrices $\mathcal{A} \in SL(2d,\mathbb{R})$ is to use the canonical correspondence of elements in $\mathcal{H}_{\mathcal{A}Q}$ with elements in $\mathcal{H}_Q$ given by a coordinate transformation in the spreading domain. Theorem 5.3 states that for symplectic $\mathcal{A}$, there exist unitary operators $O_\mathcal{A}$ on $L^2(\mathbb{R}^d)$, such that the following formal calculation of operator valued integrals holds for all $H \in \mathcal{H}_{\mathcal{A}Q}$. Note that here we set $\mu(t,\nu) = M_\nu T_t$ to obtain

$$\begin{aligned}
H &= \int\int \eta_H(t,\nu)\, M_\nu T_t\ dt\, d\nu = \int\int \eta_H(t,\nu)\ \mu(t,\nu)\, dt\, d\nu \\
&= \int\int \eta_H(\mathcal{A}(t,\nu))\ \mu(\mathcal{A}(t,\nu))\ dt\, d\nu = \int\int \eta_H(\mathcal{A}(t,\nu))\ O_\mathcal{A}\mu(t,\nu)O_\mathcal{A}{}^*\ dt\, d\nu \\
&= O_\mathcal{A} \int\int \eta_{H_\mathcal{A}}(t,\nu)\ \mu(t,\nu)\ dt\, d\nu\ O_\mathcal{A}{}^* = O_\mathcal{A}\ H_\mathcal{A}\ O_\mathcal{A}{}^*,
\end{aligned} \tag{5.2}$$

where $\eta_{H_\mathcal{A}} = \eta_H \circ \mathcal{A}$ and $H_\mathcal{A} \in \mathcal{H}_Q$. We shall see that the *intertwining* operators $O_\mathcal{A} \in U\big(L^2(\mathbb{R}^d)\big)$ in (5.2) extend to $S_0'(\mathbb{R}^d)$ and act isomorphically on $S_0(\mathbb{R}^d)$. The identifiability of $\mathcal{H}_Q$ leads therefore to the identifiability of $\mathcal{H}_{\mathcal{A}Q}$ using as identifier the tempered distribution $O_\mathcal{A}⊥⊥⊥ \in S_0'(\mathbb{R}^d)$. See Figure 5.1 for an illustration of this approach.

To gather all $\mathcal{A} \in SL(2d,\mathbb{R})$ which allow for calculations similar to those in (5.2), we turn to the representation theory of the reduced Weyl–Heisenberg group $\mathbb{H}_d^{red}$, which is identical to $\mathbb{R}^d \times \widehat{\mathbb{R}}^d \times \mathbb{T}$ in topology and Haar measure. The group operation on the reduced Weyl–Heisenberg group is

$$\big(t,\nu,e^{2\pi i s}\big) \cdot \big(t',\nu',e^{2\pi i s'}\big) = \big(t+t',\nu+\nu',e^{2\pi i(s+s'+\frac{1}{2}(t'\cdot\nu - t\cdot\nu'))}\big),$$

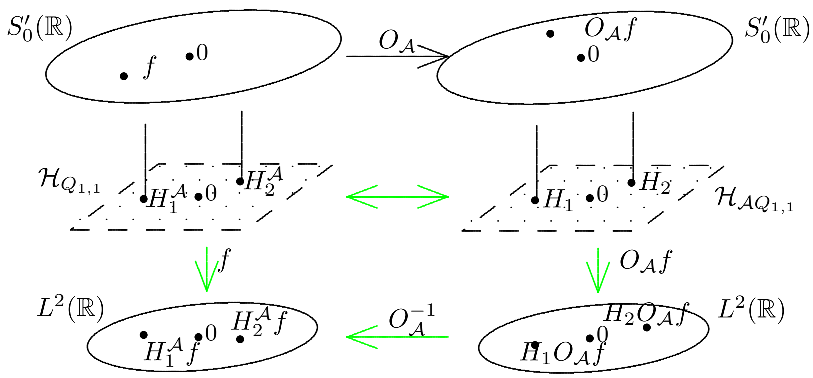


FIG. 5.1. *Identifiability of $\mathcal{H}_{\mathcal{A}Q}$, $\mathcal{A} \in Sp(d,\mathbb{R})$, based on the existence of an intertwining operator $O_{\mathcal{A}}$.*

and its Schrödinger representation on the space of unitary operators on $L^2(\mathbb{R}^d)$ is given by

$$\begin{array}{lllllllll}
\rho: & \mathbb{H}_d^{red} & \to & U\big(L^2(\mathbb{R}^d)\big) & & & & & \\
& (t,\nu,s) & \mapsto & \rho(t,\nu,s): & L^2(\mathbb{R}) & \to & L^2(\mathbb{R}) & & \\
& & & & f & \mapsto & \rho(t,\nu,e^{2\pi i s})f: & \mathbb{R}^d & \to \quad \mathbb{R}^d \\
& & & & & & & x & \mapsto \quad e^{2\pi i(\nu\cdot x)+s} f(x+t)\ .
\end{array}$$

Representing H once more as operator valued integral, we obtain

$$\begin{aligned}
H = \int\!\!\int \eta_H(t,\nu) M_\nu T_t \, dt\, d\nu \ &= \ \int_{\mathbb{R}^d}\int_{\widehat{\mathbb{R}}^d}\int_0^1 \eta_H(t,\nu)\rho(-t,\nu,0)\, dt\, d\nu\, ds \\
= \int_{\mathbb{H}_d^{red}} e^{-2\pi i s}\eta_H(-t,\nu)\rho(t,\nu,e^{2\pi i s})\, d\mu(t,\nu,s) \ &= \ \rho(\eta_H^\circ),
\end{aligned}$$

where $\eta_H^\circ(t,\nu,e^{2\pi i s}) = e^{-2\pi i s}\eta_H(-t,\nu)$. In other words, a Hilbert–Schmidt operator H with $\eta_H \in L^1(\mathbb{R}^{2d})$ is the integrated Schrödinger representations of η_H° with respect to the reduced Weyl–Heisenberg group $\mathbb{H}_d^{red}$ [Fol89, Grö01].

Before listing the relevant results from representation theory in Theorem 5.3, it is now time to define the symplectic group.

DEFINITION 5.2. *The symplectic group $Sp(d,\mathbb{R})$ consists of those matrices $\mathcal{A} \in SL(2d,\mathbb{R})$ that satisfy $\mathcal{A}^\star \left(\begin{smallmatrix} 0 & -I_d \\ I_d & 0 \end{smallmatrix}\right) \mathcal{A} = \left(\begin{smallmatrix} 0 & -I_d \\ I_d & 0 \end{smallmatrix}\right)$, where I_d is the $d\times d$ identity matrix.*

Theorem 5.3(a) outlines the scope of our approach [Fol89]. Part (c) delivers intertwining operators for equivalent representations $\rho \circ \mathcal{A}$ and ρ. Parts (d), (e), (f), and (g) describe these operators as products of some elementary operators. This characterization shows us that the a priori Hilbert space theory applies to the Feichtinger algebra setup used in this paper (see part (h)). Part (i) covers shifts of the spreading support which allow us to extend Theorem 5.4 to affine linear coordinate transformations.

For ease of notation we shall not distinguish between the matrix $\mathcal{A}$ and the corresponding linear map; i.e., we have $\mathcal{A}(t,\nu) = \big((t,\nu)\cdot\mathcal{A}^t\big)^t$.

THEOREM 5.3.

(a) *Let S operate on $\mathbb{H}_d^{red}$. The induced map $\rho_S = \rho\circ S : \mathbb{H}_d^{red} \longrightarrow U\left(L^2(\mathbb{R}^d)\right)$ is an unitary representation of $\mathbb{H}_d^{red}$ which is unitarily equivalent to the irre-*

ducible Schrödinger representation ρ; i.e., there exists an unitary intertwining operator O such that $O\rho(g)O^* = \rho_S(g)$ for all $g \in HW_d^{red}$ if and only if there exists $\mathcal{A} \in Sp(d,\mathbb{R})$ with $S = \widetilde{\mathcal{A}}$, where $\widetilde{\mathcal{A}}$ is given by $\widetilde{\mathcal{A}} : \mathbb{H}_d^{red} \longrightarrow \mathbb{H}_d^{red}, (t,\nu,e^{2\pi is}) \mapsto (\mathcal{A}(t,\nu), e^{2\pi is})$.

(b) *Let $\mathcal{A} \in Sp(d,\mathbb{R})$ and let $\rho_{\mathcal{A}} = \rho \circ \widetilde{\mathcal{A}}$. Then $\rho_{\mathcal{A}}(f) = \rho(f \circ \mathcal{A}^{-1})$ for $f \in L^1(\mathbb{H}_d^{red})$.*

(c) *For $\mathcal{A} \in Sp(d,\mathbb{R})$ there exists a unitary operator $O_{\mathcal{A}}$ on $L^2(\mathbb{R}^d)$, with $O_{\mathcal{A}}HO_{\mathcal{A}}^* = \rho(\eta(H)^\circ \circ \widetilde{\mathcal{A}}^{-1})$ for all $H \in HS(L^2(\mathbb{R}^d))$ with $\eta(H) \in L^1(\mathbb{R}^{2d})$.*

(d) *The matrix $\mathcal{I} = \left(\begin{smallmatrix} 0 & I_d \\ -I_d & 0 \end{smallmatrix}\right)$ together with the subgroups*

$$N = \left\{ \left(\begin{smallmatrix} I_d & 0 \\ A & I_d \end{smallmatrix}\right), A = A^* \right\} \quad \text{and} \quad D = \left\{ \left(\begin{smallmatrix} A & 0 \\ 0 & A^{*-1} \end{smallmatrix}\right), A \in GL(n,\mathbb{R}) \right\}$$

of $Sp(d,\mathbb{R})$ generate $Sp(d,\mathbb{R})$.

(e) *For $\mathcal{A} = \left(\begin{smallmatrix} 0 & I_d \\ -I_d & 0 \end{smallmatrix}\right)$ we have $\mu \circ \mathcal{A}(t,\nu) = \mu(\mathcal{A}(t,\nu)) = \mathcal{F}^{-1}\mu(t,\nu)\mathcal{F}$.*

(f) *For $\mathcal{A} = \left(\begin{smallmatrix} I_d & 0 \\ A & I_d \end{smallmatrix}\right)$ with $A = A^*$ define C_A through $C_Af(x) = e^{-\pi i x^TAx}f(x)$. Then we have $\mu \circ \mathcal{A}(t,\nu) = \mu(\mathcal{A}(t,\nu)) = C_A \circ \mu(t,\nu) \circ C_A^*$.*

(g) *For $\mathcal{A} = \left(\begin{smallmatrix} A & 0 \\ 0 & A^{*-1} \end{smallmatrix}\right)$ let U_A be defined by setting $U_Af(x) = |\det A|^{-\frac{1}{2}}f(A^{-1}x)$. Then $\mu \circ \mathcal{A}(t,\nu) = \mu(\mathcal{A}(t,\nu)) = U_A \circ \mu(t,\nu) \circ U_A^*$.*

(h) *The unitary operators $\mathcal{F}$, C_A, and U_A restrict and extend to $S_0(\mathbb{R})$ and $S_0'(\mathbb{R})$, respectively.*

(i) *Set $L_{(a,b)} : \mathbb{R}^d \times \widehat{\mathbb{R}}^d \longrightarrow \mathbb{R}^d \times \widehat{\mathbb{R}}^d, (t,\nu) \mapsto (a+t, b+\nu)$. Then*

$$\mu \circ L_{(a,b)}(t,\nu) = e^{2\pi i\nu a}\mu(a,b)\mu(t,\nu) = e^{2\pi ibt}\mu(t,\nu)\mu(a,b).$$

For details on representation theoretic background, see [Fol89, FK98, Grö01]. Using Theorem 5.3, we obtain the following.

THEOREM 5.4. *Let $S = L_{(a,b)} \circ \mathcal{A}$, $\mathcal{A} \in Sp(d,\mathbb{R})$. Then $\mathcal{H}_M$ is identifiable if and only if $\mathcal{H}_{SM}$ is identifiable.*

Proof. Assume that $\mathcal{H}_M$ is identifiable with $f_M \in S_0'(\mathbb{R}^d)$. Theorem 5.3 provides us with an unitary operator $O_{\mathcal{A}}$ on $L^2(\mathbb{R}^d)$ which extends to $S_0'(\mathbb{R}^d)$. We claim that $O_{\mathcal{A}}f_M \in S_0'(\mathbb{R}^d)$ identifies $\mathcal{H}_{SM}$. To see this, observe that for all $H \in \mathcal{H}_{SM}$ we have

$$\begin{aligned}
H &= \int\int \eta_H(t,\nu)\ \mu(t,\nu)\,dt\,d\nu \\
&= \int\int \eta_H\big(\mathcal{A}(t,\nu)+(a,b)\big)\ \mu\big(\mathcal{A}(t,\nu)+(a,b)\big)\,dt\,d\nu \\
&= \int\int \eta_H\big(\mathcal{A}(t,\nu)+(a,b)\big)\ e^{2\pi ia(Ct+D\nu)}\mu(a,b)\mu(\mathcal{A}(t,\nu))\ dt\,d\nu \\
&= \int\int \eta_H\big(\mathcal{A}(t,\nu)+(a,b)\big)e^{2\pi ia(Ct+D\nu)}\ \mu(a,b)O_{\mathcal{A}}\mu(t,\nu){O_{\mathcal{A}}}^*\ dt\,d\nu \\
&= \mu(a,b)O_{\mathcal{A}}\int\int \eta_{H_{\mathcal{A},(a,b)}}(t,\nu)\ \mu(t,\nu)\ dt\,d\nu\ {O_{\mathcal{A}}}^* \\
&= \mu(a,b)O_{\mathcal{A}}\ H_{\mathcal{A},(a,b)}\ {O_{\mathcal{A}}}^*,
\end{aligned}$$

and

$$\begin{aligned}
&\|H\,O_{\mathcal{A}}f_M\|_{L^2(\mathbb{R}^d)} = \|{O_{\mathcal{A}}}^*\mu(a,b)^*HO_{\mathcal{A}}f_M\|_{L^2(\mathbb{R}^d)} \\
&= \|H_{\mathcal{A},(a,b)}f_M\|_{L^2(\mathbb{R}^d)} = \|\eta_{H_{\mathcal{A},(a,b)}}\|_{L^2(\mathbb{R}^{2d})} = \|\eta_H\|_{L^2(\mathbb{R}^{2d})} \equiv \|H\|_{HS}. \qquad \square
\end{aligned}$$

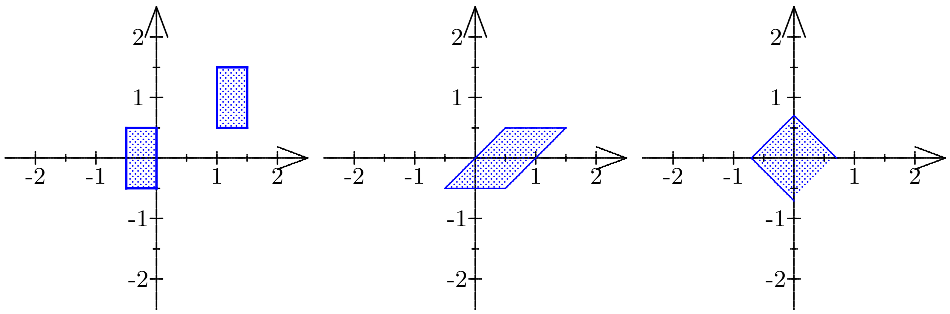


FIG. 5.2. *Examples of sets M such that H_M is identifiable.*

Theorems 3.3 and 5.4 imply that the exemplary spreading support sets M given Figure 5.2 define identifiable operator classes H_M.

For $M = Q$, we can identify $\mathcal{H}_{SQ}$ using the identity in the following.

COROLLARY 5.5. *Let $S = L_{(a,b)} \circ \mathcal{A}$ for some $(a,b) \in \mathbb{R}^d \times \widehat{\mathbb{R}}^d$ and $\mathcal{A} = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in Sp(d,\mathbb{R})$. Then for $H \in \mathcal{H}_{SQ}$ and for $(t,\nu) \in \operatorname{supp}(\eta_H)$*

$$e^{2\pi i a\cdot(Ct+Dv)+\nu t} Z \circ O_S \circ H \circ O^*_{S}{\perp\!\perp\!\perp}(t,\nu) = \eta_H(\mathcal{A}^{-1}(t-a,\nu-b)).$$

We have shown that identifiability is robust with respect to symplectic coordinate transformations in the spreading domain. This result is rooted in the representation theory of the Weyl–Heisenberg group. Theorem 5.3(i) shows that this approach can not be extended to obtain insights on nonsymplectic coordinate transformations.

Nevertheless, we should note that for $\mathcal{A} \in SL(2d,\mathbb{R})$ the condition $\mathcal{A} \in Sp(d,\mathbb{R})$ is not necessary for $\mathcal{H}_{\mathcal{A}Q}$ to be identifiable. In fact, the diagonal matrix $\mathcal{D}$ with diagonal $(2,\, \frac{1}{2},\, 1,\, 1)$ has the property $\mathcal{D} \in SL(4,\mathbb{R}) \setminus Sp(2,\mathbb{R})$, but $\mathcal{H}_{\mathcal{D}Q}$ is identifiable since $\mathcal{D}Q$ is a fundamental domain of the symplectic lattice $\left(\begin{smallmatrix} A & 0 \\ 0 & A^{*-1} \end{smallmatrix}\right)\mathbb{Z}^4$ with $A = \left(\begin{smallmatrix} 1 & 0 \\ \frac{1}{2} & 1 \end{smallmatrix}\right)$, and therefore an application of Theorems 5.1 and 5.4 is permissible.

For similar results on nonsymplectic lattices in Gabor theory see [Bek04, HW01, HW04].

Acknowledgments. We would like to thank Georg Zimmermann for many fruitful discussions regarding Kailath's conjecture and slowly time-varying operators in general. Further, we thank the anonymous reviewer for many suggestions that helped to improve the clarity and readability of this paper.

REFERENCES

[Bek04] B. BEKKA, *Square integrable representations, von Neumann algebras, and an application to Gabor analysis*, J. Fourier Anal. Appl., 10 (2004), pp. 325–349.

[BHW98] J. J. BENEDETTO, C. HEIL, AND D. F. WALNUT, *Gabor systems and the Balian-Low theorem*, in Gabor Analysis and Algorithms: Theory and Applications, H. G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Boston, MA, 1998, pp. 85–122.

[CG03] E. CORDERO AND K. GRÖCHENIG, *Time-frequency analysis of localization operators*, J. Funct. Anal., 205 (2003), pp. 107–131.

[Dau88] I. DAUBECHIES, *Time-frequency localization operators: A geometric phase space approach*, IEEE Trans. Inform. Theory, 34 (1988), pp. 605–612.

[Dau92] I. DAUBECHIES, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, 1992.

[Die70] J. DIEUDONNÉ, *Treatise on Analysis.* Vol. II; translated from the French by I. G. MacDonald. Pure Appl. Math. 10-II, Academic Press, New York, 1970.

[Fef83] C. L. Fefferman, *The uncertainty principle*, Bull. Amer. Math. Soc. (N.S.), 9 (1983), pp. 129–206.

[Fei81] H. G. Feichtinger, *On a new Segal algebra*, Monatsh. Math., 92 (1981), pp. 269–289.

[FG85] H. G. Feichtinger and P. Gröbner, *Banach spaces of distributions defined by decomposition methods.* I, Math. Nachr., 123 (1985), pp. 97–120.

[FK98] H. G. Feichtinger and W. Kozek, *Quantization of TF-lattice invariant operators on elementary LCA groups*, in Gabor Analysis and Algorithms: Theory and Applications, H. G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Boston, MA, 1998, pp. 233–266.

[FL96] B. H. Fleury and P. E. Leuthold, *Radiowave propagation in mobile communications: An overview of European research*, IEEE Commun. Mag., 34 (1996), pp. 70–81.

[FN03] H. G. Feichtinger and K. Nowak, *A first survey of Gabor multipliers*, in Advances in Gabor Analysis, H. G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Boston, MA, 2003, pp. 99–128.

[Fol89] G. B. Folland, *Harmonic Analysis in Phase Space*, Ann. of Math. Stud., Princeton University Press, Princeton, NJ, 1989.

[FS97] G. B. Folland and A. Sitaram, *The uncertainty principle: A mathematical survey*, J. Fourier Anal. Appl., 3 (1997), pp. 207–238.

[FZ98] H. G. Feichtinger and G. Zimmermann, *A space of test functions for Gabor analysis*, in Gabor Analysis and Algorithms: Theory and Applications, H. G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Boston, MA, 1998, pp 123–170.

[Gaa73] S. A. Gaal, *Linear Analysis and Representation Theory*, Springer-Verlag, New York, 1973.

[Gab46] D. Gabor, *Theory of communication*, J. IEE London, 93 (1946), pp. 429–457.

[GLM04] P. C. Gibson, M. P. Lamoureux, and G. F. Margrave, *Representation of linear operators by Gabor multipliers*, preprint.

[Gre68] P. E. Green, *Radar measurements of target scattering properties*, in Radar Astronomy, McGraw–Hill, New York, 1968, pp. 1–78.

[Grö01] K. Gröchenig, *Foundations of Time-Frequency Analysis*, Appl. Numer. Harmon. Anal., Birkhäuser Boston, Boston, MA, 2001.

[Grö03] K. Gröchenig, *Uncertainty principles for time-frequency representations*, in Advances in Gabor Analysis, H. G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Boston, MA, 2003, pp. 11–30.

[Hei03] C. Heil, *Integral operators, pseudodifferential operators, and Gabor frames*, in Advances in Gabor Analysis, H. G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Boston, MA, 2003, pp. 153–169.

[HRT97] C. Heil, J. Ramanathan, and P. Topiwala, *Singular values of compact pseudodifferential operators*, J. Funct. Anal., 150 (1997), pp. 426–452.

[HW01] D. Han and Y. Wang, *Lattice tiling and the Weyl–Heisenberg frames*, Geom. Funct. Anal., 11 (2001), pp. 742–758.

[HW04] D. Han and Y. Wang, *The existence of Gabor bases and frames*, in Wavelets, Frames, and Operator Theory, Contemp. Math. 35, AMS, Providence, RI, 2004, pp. 183–192.

[Jan95] A. J. E. M. Janssen, *Duality and biorthogonality for Weyl–Heisenberg frames*, J. Fourier Anal. Appl., 1 (1995), pp. 403–436.

[Kai59] T. Kailath, *Sampling Models for Linear Time-Variant Filters*, Technical report 352, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, 1959.

[Kai62] T. Kailath, *Measurements on time-variant communication channels*, IEEE Trans. Inform. Theory, 8 (1962), pp. 229–236.

[Kai63] T. Kailath, *Report on progress in information theory in the USA* 1960–1963*. IV: Time-variant communication channels*, IEEE Trans. Inform. Theory, IT-9 (1963), pp. 223–264.

[Kat76] Y. Katznelson, *An Introduction to Harmonic Analysis*, Dover, New York, 1976.

[KN65] J. J. Kohn and L. Nirenberg, *An algebra of pseudo-differential operators*, Comm. Pure Appl. Math., 18 (1965), pp. 269–305.

[Koz98] W. Kozek, *Adaptation of Weyl–Heisenberg frames to underspread environments*, in Gabor Analysis and Algorithms: Theory and Applications, H. G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Boston, MA, 1998, pp. 323–352.

[Lab01] D. Labate, *Time-frequency analysis of pseudodifferential operators*, Monatsh. Math., 133 (2001), pp. 143–156.

[Lan93] H. J. Landau, *On the density of phase-space expansions*, IEEE Trans. Inform. Theory, 39 (1993), pp. 1152–1156.

[LKS03] K. Liu, T. A. Kadous, and A. M. Sayeed, *Orthogonal time-frequency signaling over doubly dispersive channels*, IEEE Trans. Inform. Theory, 50 (2004), pp. 2583–2603.

[LP61] H. J. Landau and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis, and uncertainty.* II, Bell System Tech. J., 40 (1961), pp. 65–84.

[LP62] H. J. Landau and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis, and uncertainty.* III. *The dimension of the space of essentially time- and band-limited signals*, Bell System Tech. J., 41 (1962), pp. 1295–1336.

[Lyu92] Y. I. Lyubarskii, *Frames in the Bargmann space of entire functions*, Adv. Soviet Math., 429 (1992), pp. 107–113.

[MGO03] X. Ma, G. B. Giannakis, and S. Ohno, *Optimal training for block transmissions over doubly selective wireless fading channels*, IEEE Trans. Signal Process., 51 (2003), pp. 1351–1366.

[MMH$^+$02] G. Matz, A. F. Molisch, F. Hlawatsch, M. Steinbauer, and I. Gaspard, *On the systematic measurement errors of correlative mobile radio channel sounders*, IEEE Trans. Comm., 50 (2002), pp. 808–821.

[Pfa05] G. E. Pfander, *Two Results on the Invertibility of Bi-infinite Matrices with Applications to Gabor Analysis*, preprint, 2005.

[RS97] A. Ron and Z. Shen, *Weyl–Heisenberg frames and Riesz bases in* $L_2(\mathbf{R}^d)$, Duke Math. J., 89 (1997), pp. 237–282.

[RT98] R. Rochberg and K. Tachizawa, *Pseudodifferential operators, Gabor frames, and local trigonometric bases*, in Gabor Analysis and Algorithms: Theory and Applications, H. G. Feichtinger and T. Strohmer, eds., Birkhäuser Boston, Boston, MA, 1998, pp. 171–192.

[Sei92] K. Seip, *Density theorems for sampling and interpolation in the Bargmann–Fock space.* I, J. Reine Angew. Math., 429 (1992), pp. 91–106.

[Sko80] M. I. Skolnik, *Introduction to Radar Systems*, McGraw–Hill, New York, 1980.

[SP61] D. Slepian and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis, and uncertainty.* I, Bell System Tech. J., 40 (1961), pp. 43–63.

[Str05] T. Strohmer, *Pseudodifferential Operators and Banach Algebras in Mobile Communications*, Appl. Comp. Harm. Anal., to appear.

[SW92] K. Seip and R. Wallstén, *Density theorems for sampling and interpolation in the Bargmann–Fock space.* II, J. Reine Angew. Math., 429 (1992), pp. 107–113.

[VT71] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, Wiley, New York, 1971.

[Yoo02] D.-S. Yoo, *Channel Characterization and System Design in Wireless Communications*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 2002.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 889–913

STABILITY OF LARGE-AMPLITUDE SHOCK PROFILES OF GENERAL RELAXATION SYSTEMS*

CORRADO MASCIA† AND KEVIN ZUMBRUN‡

Abstract. Building on previous analyses carried out in [Mascia and Zumbrun, *Indiana Univ. Math. J.*, 51 (2002), pp. 773–904] and [Mascia and Zumbrun, *Arch. Ration. Mech. Anal.*, 172 (2004), pp. 93–131], we establish $L^1 \cap H^2 \to L^p$ nonlinear orbital stability, $1 \leq p \leq \infty$, with sharp rates of decay, of large-amplitude Lax-type shock profiles for a general class of relaxation systems that includes most models in common use, under the necessary conditions of strong spectral stability, i.e., stable point spectrum of the linearized operator about the wave, transversality of the profile, and hyperbolic stability of the associated ideal shock. In particular, our results apply to standard moment-closure systems, answering a question left open in Mascia and Zumbrun (2002). The argument combines the basic nonlinear stability argument introduced in Mascia and Zumbrun (2002) with an improved "Goodman-style" weighted energy estimate similar to but substantially more delicate than that used in Mascia and Zumbrun (2004) to treat large-amplitude profiles of systems with real viscosity.

Key words. relaxation systems, stability of traveling waves

AMS subject classifications. 35L60, 35B30, 35B45, 35L65, 76L05

DOI. 10.1137/S0036141004435844

1. Introduction. In [24], a detailed study was carried out on linearized and nonlinear stability of traveling front solutions, or shock profiles

$$(u,v)(x,t) = (\bar{u},\bar{v})(x-st), \qquad \lim_{z\to\pm\infty}(\bar{u},\bar{v}) = (u_\pm, v_\pm) = (u_\pm, v^*(u_\pm)), \tag{1.1}$$

of relaxation systems

$$\begin{cases} u_t + f(u,v)_x = 0, \\ v_t + g(u,v)_x = q(u,v), \end{cases} \tag{1.2}$$

$u,\ f \in \mathbb{R}^n$, $v,\ g,\ q \in \mathbb{R}^r$, where

$$\operatorname{Re}\sigma(q_v(u,v^*(u))) < 0$$

along a smooth *equilibrium manifold* $v = v^*(u)$ defined by $q(u,v^*(u)) = 0$.

The linearized results obtained in [24] are extremely general and appear to be optimal. However, the nonlinear results are restricted to arbitrary-amplitude profiles of special, discrete kinetic models, defined as systems (1.2) for which f and g are linear, and small-amplitude profiles of general, simultaneously symmetrizable models, defined as systems (1.2) for which there exists a smooth, symmetric positive definite matrix function $A^0 = A^0(u,v)$ for which A^0A and A^0Q are symmetric, where $A := (df^t, dg^t)^t$ and $Q := (0, dq^t)^t$.

*Received by the editors June 11, 2004; accepted for publication (in revised form) November 30, 2004; published electronically December 6, 2005.
http://www.siam.org/journals/sima/37-3/43584.html

†Dipartimento di Matematica "G. Castelnuovo," Università di Roma "La Sapienza," P.le Aldo Moro, 2, 00185 Roma, Italy (mascia@mat.uniroma1.it). The research of this author was supported in part by European TMR project "Hyperbolic Systems of Conservation Laws."

‡Department of Mathematics, Indiana University, Bloomington, IN 47405-4301 (kzumbrun@indiana.edu). The research of this author was supported in part by the National Science Foundation under grants DMS-0070765 and DMS-0300487.

The first result is quite satisfactory and has important applications to the physically interesting class of systems obtained by discretization of kinetic models such as Boltzmann or Vlasov–Poisson equations. On the other hand, the equally important class of systems obtained by moment-closure approximation of kinetic models is not contained in the class of discrete kinetic models, and so the question of large-amplitude stability for these models was left open in [24]; indeed, as discussed in Remark 1.16 of [24], it was not at all clear from the analysis of [24] whether this was a technical artifact or represented a true qualitative difference in behavior between these two types of approximation. Moreover, even the requirement of simultaneous symmetrizability appears to be overly restrictive in the large-amplitude case. As described in [33], many systems of physical interest are simultaneously symmetrizable along the equilibrium manifold $v = v^*(u)$, including (standard versions of) both discrete and moment-closure approximations of kinetic models; however, so far as we can see, simultaneous symmetrizability does not typically hold away from equilibrium, in particular along a shock profile, as assumed in [24].[1] This distinction is unimportant in the small-amplitude case, for which the argument of [24] goes through by continuity assuming only simultaneous symmetrizability at the endstates $(u_-, v^*(u_-))$;[2] however, it becomes significant in the large-amplitude case, for which profiles may feature arbitrarily large excursions from equilibrium.

The difficulty in the analysis of the general case was control of higher-derivative source terms arising in the nonlinear iteration through Taylor expansion of the variable coefficient matrix $A = (df^t, dg^t)^t$. Such terms do not arise in the case of discrete kinetic models, for which A is constant, and this made it possible to carry out the entire nonlinear stability analysis using linearized (i.e., Green function) estimates alone. In the general case, we found it necessary to augment these bounds with coupled energy estimates in order to close the iteration, and these estimates, as implemented in [24], used both global symmetrizability and the small-amplitude assumption in important ways. In particular, these assumptions were used to guarantee that the perturbation equations be locally dissipative everywhere along the shock profile, whereas, in the present, large-amplitude case, the perturbation equations are in general dissipative only near plus or minus spatial infinity.

Similar difficulties arose in the closely related study [25] of stability of shock profiles for real viscosity models, initially limiting this analysis also to the small-amplitude, globally symmetrizable case. Recently, however, these obstacles were overcome and the corresponding restrictions removed in [27] by the introduction of a modified energy estimate incorporating "Goodman-type" weighted norms in the style of [12]. As discussed in [39, 40], these quantify the observation that transverse convection relative to the shock profile already yields a complementary type of dissipation near the inner shock layer by rapidly sweeping signals to plus or minus spatial infinity, where they then decay under the effects of the local dissipativity guaranteed by (A1)–(A2).

In this paper, we show that a similar approach can be applied in the relaxation case to yield a satisfactory nonlinear stability theory applying to large-amplitude profiles of the physically correct class of systems that are simultaneously symmetrizable *at equilibrium*—indeed, to the considerably more general class of equations that was considered in the linearized analysis of [24].

---

[1] This corrects a misstatement in [24, Remark 1.15], where the calculations of [33] were misquoted as asserting global simultaneous symmetrizability.

[2] Specifically, in the notation of [24, Lemma 7.4], the key property $\mathrm{Re}\,(kA - A^0Q) > 0$ persists under perturbation.

This brings the analysis of [24] to a satisfying conclusion, putting under a common framework most of the relaxation models typically studied, and in particular resolving in the negative the question posed in [24] of whether there might be a true qualitative difference in behavior between discrete and moment-closure approximations of kinetic models. We point out that our results yield new information even in the case of (non-globally-symmetrizable) discrete kinetic models, since we require significantly less regularity on the data ($L^1 \cap H^2$ vs. $W^{3,1} \cap W^{3,\infty}$). Also, the low norm decay rates, L^p for $1 \le p < 2$, are new for general models even in the small-amplitude case.

A substantial new difficulty in the relaxation as compared to the real viscosity case is that the hyperbolic characteristic speeds corresponding to eigenvalues of $(df^t, dg^t)^t$ are by the subcharacteristic condition (a necessary condition for dissipativity condition (A2); see [24, 32]) necessarily of both positive and negative signs, whereas in the case of real viscosity, hyperbolic modes were assumed to be all of one sign, i.e., strictly upwind or strictly downwind. Since weights are chosen to decay exponentially in the direction of propagation, this means that a single scalar weight no longer suffices in the relaxation case, and the introduction of a matrix of distinct diagonal weights leads to off-diagonal error terms that grow exponentially in both the amplitude of the shock and in an arbitrary constant C_* determining the amount of dissipation on the inner layer with respect to the chosen weight.

At first sight, it is hard to see how such an argument could ever close, since good terms of order C_* generate bad terms of exponential order in C_*. Remarkably, the energy estimates *can* close by a refinement of the hyperbolic compensation argument of [17, 30]: namely, the observation (Lemma 2.3 below) that the same "compensating matrix" K used to complete the partial dissipation provided by the semiparabolic matrix B may be used at the same time to eliminate off-diagonal terms of essentially arbitrary size. The details of this argument may be found in sections 2 and 3.

We now describe our results in more detail. Of system (1.2), we assume the following structural properties:

(A1) *Symmetrizability of* $A(u,v) = (df^t, dg^t)^t$*:* there exists a C^3 positive definite matrix-valued function $A^0(u,v)$ such that A^0A is symmetric.

(A2) *Dissipativity at the equilibrium states* $(u_\pm, v^*(u_\pm))$*:* for some $\theta > 0$,

$$\operatorname{Re}\sigma(i\xi A_\pm + Q_\pm) \le \frac{-\theta|\xi|^2}{1+|\xi|^2} \qquad \forall\, \xi \in \mathbb{R},$$

where $A_\pm := A(u_\pm, v^*(u_\pm))$, $Q_\pm := Q(u_\pm, v^*(u_\pm))$, and $Q(u,v) := (0, dq(u,v))^t$.

Condition (A1) is connected with well-posedness of (1.2), while (A2) is connected with time-asymptotic stability of equilibrium states $(u_\pm, v^*(u_\pm))$ [17]. As discussed variously in [5, 17, 33, 35], sufficient conditions for (A2) are either simultaneous symmetrizability, A^0Q symmetric, or else weak dissipativity, $Q(A^0)^{-1} \le 0$ and $(Q(A^0)^{-1})_{22} < 0$,[3] together with *genuine coupling*:

$$\text{No eigenvector of } A_\pm \text{ lies in the kernel of } Q_\pm. \tag{1.3}$$

Hereafter, set $s = 0$ (changing to coordinates moving with the shock), so that (1.1) becomes a standing-wave solution. Regarding the profile $(\bar u, \bar v)$, we assume the following:

[3] That is, $\tilde Q := Q(A^0)^{-1} =$ block-diag $\{0, \tilde q\}$ with $\tilde q < 0$, whence we may replace $\tilde Q$ with $\operatorname{Re}\tilde Q$ and proceed as in the simultaneously symmetrizable case, noting that $\ker A^0 \operatorname{Re}\tilde Q = \ker A^0\tilde Q = (I_{n-r}, 0)$, whence (1.3) is preserved. The example $\tilde q = \begin{pmatrix} 0 & a \\ -a & 1 \end{pmatrix}$, $a \ne 0$, shows that $\tilde q > 0$ is necessary, since $\operatorname{Re}\sigma\tilde q < 0$ but $\operatorname{Re}\tilde q =$ block-diag $\{0, \tilde q\}$ is only semidefinite.

(H0) $f,\, g,\, q \in C^3$.

(H1) The eigenvalues of $A(x) := (df^t, dg^t)^t(\bar u, \bar v)(x)$ are (i) different from 0 and (ii) of constant multiplicity.

(H2) (i) The eigenvalues $a_j^{*\pm}$ of $A_\pm^* := df^*(u_\pm)$, $f^*(u) := f(u, v^*(u))$, are real and different from 0;[4] moreover, (ii) when ordered with increasing size, they satisfy the *strict Lax characteristic conditions* [19]

$$a_{p-1}^{*-} < 0 < a_p^{*-}, \quad a_p^{*+} < 0 < a_{p+1}^{*+}$$

for some $1 \le p \le n$ (the principal characteristic field of the shock).

(H3) *Dynamical stability:* the *Liu–Majda determinant condition* [20, 21, 22, 23]

$$\Delta := \det(r_1^{*-}, \dots, r_{p-1}^{*-}, [u], r_{p+1}^{*+}, r_n^{*+}) \neq 0 \tag{1.4}$$

is satisfied, where $r_j^{*\pm}$ denote the eigenvectors of $A_\pm^*$ associated with $a_j^{*\pm}$, and $[u] := u_+ - u_-$ denotes the jump in u across the shock.

(H4) *Structural stability:* the profile $(\bar u, \bar v)(\cdot)$ is a transverse connection of the associated standing-wave ODE, in particular, therefore, locally unique up to translation.

(H5) *Strong spectral stability:* the point spectrum of the linearized operator $\mathcal{L}$ about the wave is contained in $\{\lambda : \operatorname{Re}\lambda < 0\} \cup \{0\}$.

Condition (H0) gives the regularity needed both for our analysis here and in order to apply the linearized bounds of [24]. Condition (H1)(i) is a standard assumption [34, 38, 24] ensuring that the standing-wave ODE be of nondegenerate type [24]. It is not clearly necessary, however, and at least for discrete kinetic models it can be relaxed, as we discuss in section 4; indeed, in that setting it is rather unnatural. Condition (H1)(ii) is a technical assumption that was used in the pointwise Green function analysis of [24]; at the expense of some detail in the pointwise description of linearized behavior, it may be removed altogether [39]. Constant multiplicity holds automatically for discrete kinetic models, but for moment-closure models may be difficult to verify. Together, (A1) and (H1)(ii) are equivalent to semisimplicity plus constant multiplicity of $\sigma(A)$. Condition (H2)(i) expresses hyperbolicity and noncharacteristicity of the associated "equilibrium system"

$$u_t + f^*(u)_x = 0 \tag{1.5}$$

obtained from (1.2) by formal Chapman–Enskog expansion at the endstates u_-, u_+ with respect to the corresponding ideal shock (u_-, u_+) of (1.5). Note that hyperbolicity of the equilibrium system is not required *along the profile*, thus allowing applications to interesting nonhyperbolic situations as discussed in [16, 1, 2]. Condition (H2)(ii) restricts attention for simplicity to the standard case of a classical, Lax-type

[4] This differs from hypothesis (H2) of [24], in which the eigenvalues were also required to be distinct. However, it was noted in Remark 1.12 of [24] that this requirement may be dropped when $A_\pm^*$ and $Q_\pm$ are simultaneously symmetrizable, with essentially no change in either results or notation; the same argument shows that this requirement may likewise be dropped in the general case, at the expense of further complications (specifically, the matrix-valued diffusion waves $e^{(x-a_j^{*\pm})^2(4\pi\beta_j^{*\pm}t)^{-1}}$ of the remark must be replaced by the fundamental solution of $v_t + a_j^{*\pm}v_x = \beta_j^{*\pm}v_{xx}$, where $\beta_j^{*\pm} := l_j^{*\pm t}B^{*\pm}r_j^{*\pm}$ are no longer necessarily diagonal, and the matrix-valued error-functions appearing in excited term E by its spatial integral; finally, though the precise form of scattering term S in [24, Proposition 1.10] is no longer clear, it is easily verified that it satisfies pointwise bounds yielding the same rates of $L^q \to L^p$ decay as needed for subsequent stability arguments).

shock (u_-, u_+) of (1.5); the treatment of nonclassical over- and undercompressive shocks we leave for the future.[5]

Conditions (H3)–(H5) are together equivalent to the *Evans function condition*:

($\mathcal{D}$) The Evans function $D(\cdot)$ associated with $\mathcal{L}$ has precisely one zero on $\{\lambda : \operatorname{Re}\lambda \geq 0\}$ (necessarily at $\lambda = 0$).[6]

The generalized spectral stability condition ($\mathcal{D}$) was shown in [24] to be necessary and sufficient for linearized stability under assumptions (A1)–(A2) and (H0)–(H2); the main point of this paper is to show that these conditions are sufficient also for nonlinear stability. The conditions (H3) and (H4) correspond to the classical physical notions of dynamical and structural stability (see, e.g., [3]), whereas (H5) encodes heretofore neglected relaxation effects; for further discussion in the closely related viscous case, see [38] and especially [27]. All three of conditions (H3)–(H5) hold automatically in the small-amplitude case; see [21, 22, 23, 34, 24, 28], respectively.

DEFINITION 1.1. *For a profile $\bar{U} = (\bar{u}, \bar{v})$ that is* (*as in the Lax case*) *unique up to translation, we define* nonlinear orbital stability *as convergence of $U = (u, v)(\cdot, t)$ as $t \to \infty$ to a translate $\bar{U}(\cdot - \delta(t))$, where $\delta(\cdot)$ is an appropriately chosen function describing shock location, for any solution U of* (1.2) *with initial data sufficiently close in some norm to the original profile $\bar{U}$.*

Then the main result of this paper is as follows.

THEOREM 1.2. *Let $\bar{U} = (\bar{u}, \bar{v})$ be a profile* (1.1) *of a relaxation system* (1.2), *under assumptions* (A1)–(A2) *and* (H0)–(H5). *Then $\bar{U}$ is nonlinearly orbitally stable from $L^1 \cap H^2$ to L^p for all $p \geq 2$.*

More precisely, for initial perturbations $U_0 := \tilde{U}_0 - \bar{U}$ with $|U_0|_{L^1 \cap H^2}$ sufficiently small, the solution $\tilde{U} = (\tilde{u}, \tilde{v})(x, t)$ of (1.2) *with initial data $\tilde{U}_0$ satisfies*

$$|\tilde{U}(x,t) - \bar{U}(x - \delta(t))|_{L^p} \leq C|U_0|_{L^1 \cap H^2} (1+t)^{-\frac{1}{2}(1-\frac{1}{p})} \tag{1.6}$$

for all $1 \leq p \leq \infty$, for some $\delta(t)$ satisfying

$$|\dot{\delta}(t)| \leq C|U_0|_{L^1 \cap H^2} (1+t)^{-\frac{1}{2}} \qquad \textit{and} \qquad |\delta(t)| \leq C|U_0|_{L^1 \cap H^2}.$$

Remark 1.3. Useful geometric necessary conditions for viscous stability have been obtained in [36, 38, 11] in the simultaneously symmetrizable case A^0Q symmetric using the stability index of [10, 4]. Strengthened, *signed* versions $\Delta > 0$ (under appropriate normalization) of the dynamical stability condition (1.4), these readily yield examples of spectrally unstable large-amplitude profiles, similarly to the strictly parabolic case (see, e.g., [7, 10, 38, 42]). This shows that the stability conditions assumed in Theorem 1.2 are not vacuous in the large-amplitude case. Moreover, as discussed in [38, section 6.2], the signed version of the Majda condition can serve as a physical selection principle in situations, for neither of the classical criteria of structural or dynamical stability suffice. As discussed further in [27, 38], it is an extremely interesting open problem which of the stability conditions (H3) and (H5) is in practice most restrictive.

Similarly, as in the small-amplitude case, Theorem 1.2 is obtained by a nonlinear iteration combining the linearized decay rates of [24] with an appropriate auxiliary

[5]See [24] for a linearized analysis and [41, 38, 13] for related nonlinear analyses in the viscous or viscous-dispersive case.

[6]For a precise definition of the Evans function and a proof of the equivalence of ($\mathcal{D}$) and (H3)–(H5), see [24] or [38, Appendix A2].

energy estimate controlling higher derivatives. Following [24], define the nonlinear perturbation $U = (u, v)$ by

$$U(x,t) := \tilde{U}(x + \delta(t), t) - \bar{U}(x), \tag{1.7}$$

where $\tilde{U} = (\tilde{u}, \tilde{v})$ is a solution of (1.2) and the "shock location" δ is to be determined later. Evidently, decay of U is equivalent to nonlinear orbital stability, as described in (1.6). Then the key energy estimate, and the main technical contribution of the paper, is as follows.

Proposition 1.4. *Under the hypotheses of Theorem* 1.2*, let* $U_0 \in H^2$*, and suppose that, for* $0 \le t \le T$*, both the supremum of* $|\dot{\delta}|$ *and the* $H^1 \cap W^{1,\infty}$ *norm of the perturbation* $U = (u, v)^t$ *defined by* (1.7) *remain bounded by a sufficiently small constant* $\zeta > 0$*. Then, for all* $0 \le t \le T$ *and some* $\theta > 0$*,*

$$|U|_{H^2}^2(t) \le Ce^{-\theta t}|U_0|_{H^2}^2 + C\int_0^t e^{-\theta(t-s)}(|U|_{L^2}^2 + \dot{\delta}^2)(s)ds. \tag{1.8}$$

Inequality (1.8), expressing exponential damping of high frequencies, improves the weaker bound

$$|U|_{H^2}^2(t) + \int_0^t |U_x|_{H^1}^2(s)ds \le C(\zeta^2 + |U_0|_{H^2}^2) + C\int_0^t (|U|_{H^1}^2 + \dot{\delta}^2)(s)ds \tag{1.9}$$

stated for the small-amplitude case in [24]. As discussed in [39, 40] in the context of real viscosity systems, both bounds follow from the same string of energy estimates; similar inequalities hold in the real viscosity case.

With estimate (1.9), Theorem 1.2 follows for high norms L^p, $2 \le p \le \infty$, by exactly the same argument used in [24] to treat the small-amplitude case, for the proof of (1.9) was the single place in [24] where the small-amplitude assumption was actually used. See [24, section 7], or [27, section 4] in the real viscosity case. For completeness, we give in section 4 a simplified version of this argument based on the improved estimate (1.8) of Proposition 1.4, which suffices for low norms L^p, $1 \le p < 2$, as well.

Remark 1.5. We have here restricted our attention for simplicity to the study of Lax-type shocks. Nonclassical over- and undercompressive shocks may be treated similarly under further restrictions on the initial data; see [14, 29] for analyses in the parabolic (resp., hyperbolic–parabolic) case.

The paper is outlined as follows. In section 2, we give the preliminary lemmas needed in the analysis. In section 3, we carry out the proof of Proposition 1.4, and in section 4 we prove Theorem 1.2. Finally, in section 5, we discuss the complementary characteristic case.

2. Preliminaries. As in [27], our starting point consists of the following two lemmas.

Lemma 2.1 (see [26]). *Under assumptions* (A1)–(A2)*,* (H0)–(H2)*, standing wave solutions* (1.1) *satisfy*

$$|(d/dx)^k \bar{U} - U_\pm| \le C|\bar{U}_x| \le Ce^{-\theta|x|}, \quad k = 0, \ldots, 4, \tag{2.1}$$

as $x \to \pm\infty$*, for some* $\theta > 0$*,* $U_\pm = (u_\pm, v_\pm) = (u_\pm, v^*(u_\pm))$*.*

Proof. Equivalently, the standing-wave equations may be expressed as a nondegenerate ODE with hyperbolic rest points; see [24, proof of Lemma 1.2]. ∎

LEMMA 2.2 (see [30]). *Let A and Q denote simultaneously symmetrizable matrices and A^0 their symmetrizer, with $A^0Q \le 0$.*

Then genuine coupling (1.3) *is equivalent to one of the following conditions:*

(K0) *There exists $\theta > 0$ such that $\operatorname{Re}\sigma(i\xi A + Q) \le \frac{-\theta|\xi|^2}{1+|\xi|^2}$ for all $\xi \in \mathbb{R}$;*

(K1) *There exists a smooth skew-symmetric matrix-valued function $K(A, Q, A^0)$ such that $\operatorname{Re}\left(A^0Q - KA\right) < 0$;*

(K2) block-diag $LQR < 0$, *where $L := O^t(A^0)^{\frac{1}{2}}$ and $R := L^{-1} = (A^0)^{-\frac{1}{2}}O$ are matrices of left and right eigenvectors of A block-diagonalizing LAR, with O orthonormal. Here,* block-diag M *denotes the matrix formed from the diagonal blocks of M, with each block of dimension equal to the multiplicity of corresponding eigenvalues of LAR.*

Note that strictly dissipativity assumption (A2) corresponds to condition (K0) at the asymptotic states $(u_\pm, v^*(u_\pm))$ with respect to the matrices $A_\pm$ and $Q_\pm$.

Proof. These and other useful equivalent formulations are established in [30]. The main implication for our purposes, (K2) $\Rightarrow$ (K1), follows readily from Lemma 2.3, below, by first converting to the case of symmetric A, Q by the transformations $A \to ((A^0)^{\frac{1}{2}}AA^0)^{-\frac{1}{2}}$, $Q \to ((A^0)^{\frac{1}{2}}QA^0)^{-\frac{1}{2}}$, from which the original result follows by the fact that $M > 0 \Leftrightarrow (A^0)^{\frac{1}{2}}M(A^0)^{\frac{1}{2}} > 0$, then converting by an orthonormal change of coordinates to the case that A is diagonal and Q symmetric. Variable multiplicity eigenvalues may be handled by partition of unity/interpolation, noting that $\operatorname{Re}(Q - KA) < 0$ persists under perturbation.

Under the assumed symmetry of LQR, (K0) $\Rightarrow$ (K2) follows by Taylor expansion at infinity of the spectrum of the symbol $i\xi A + Q$, from which we may deduce

$$\operatorname{Re}\sigma(\text{ block-diag}\, LQR) < 0;$$

see, e.g., [24, Appendix B]. That (K1) $\Rightarrow$ (K0) follows upon rearrangement of energy estimate

$$\langle (A^0 + |\xi|^2 A^0 - i\xi K)w, (\lambda + i\xi A + |\xi|^2 B)w\rangle = 0.$$

Finally, (K2) $\Leftrightarrow$ (1.3) is clear. ∎

For our purposes, we shall require the following slight extension of Lemma 2.2, whose proof gives at the same time an explicit description of K of which we shall later make important use. We note that an equivalent version of this result was obtained independently and previously to ours by Humpherys [15].

LEMMA 2.3. *Let D be diagonal, with real entries appearing with prescribed multiplicity in order of increasing size, and let Q be arbitrary. Then there exists a smooth skew-symmetric matrix-valued function $K(D, Q)$ such that*

$$\operatorname{Re}\,(Q - KD) = \operatorname{Re}\text{block-diag}\, Q,$$

where block-diag Q *denotes the block-diagonal part of Q, with blocks of dimension equal to the multiplicity of the corresponding eigenvalues of D.*

Proof. It is straightforward to check that the symmetric matrix $\operatorname{Re} KD = (\frac{1}{2})(KD - D^tK)$ may be prescribed arbitrarily on off-diagonal blocks by setting $K_{ij} := (a_i - a_j)^{-1}M_{ij}$, where M_{ij} is the desired block, $i \ne j$. Choosing $M = \operatorname{Re} Q$, we obtain $\operatorname{Re}(Q - KD) = \operatorname{Re}$ block-diag (Q) as claimed. ∎

We shall also need the following two elementary results.

LEMMA 2.4. *Given* (A2)*, there exist block-diagonalizing matrices* $L_\pm$*,* $R_\pm$*,* $LAR_\pm$ *block-diagonal,* $LR_\pm = I$*, such that*

$$\text{block-diag}\, LQR_\pm < 0.$$

(Note: A and Q are not assumed simultaneously symmetrizable as in Lemma 2.2.)

Proof. Again, $\operatorname{Re}\sigma(\text{block-diag}\, \tilde{L}Q\tilde{R}_\pm) < 0$ follows from (A2) by Taylor expansion at infinity of the spectrum of the symbol $i\xi A_\pm + Q_\pm$ for any block-diagonalizing transformations $\tilde{L}_\pm$, $\tilde{R}_\pm$; see [24, Appendix B]. By a standard linear-algebraic lemma (see, e.g., [31, Proposition A.9, p. 361]), $\text{block-diag}\, S^{-1}\tilde{L}Q\tilde{R}S_\pm < 0$, $S_\pm := \text{block-diag}\,\{S_1, \dots, S_k\}_\pm$ for some choice of nonsingular $S_j^\pm$. Taking $L_\pm := S^{-1}\tilde{L}_\pm$, $R_\pm := \tilde{R}S_\pm$, and we are done. □

LEMMA 2.5. *There is a correspondence between symmetric positive definite symmetrizers* A^0*,* A^0A *symmetric, and diagonalizing transformations* L*,* R*,* LAR *diagonal, given by* $A^0 = L^*L$*, or equivalently* $L = O^*(A^0)^{\frac{1}{2}}$*, where* O *is an orthonormal matrix diagonalizing the symmetric matrix* $(A^0)^{\frac{1}{2}}A(A^0)^{-\frac{1}{2}}$*.*

Moreover, the matrix O (*or equivalently* L) *may be chosen with the same degree of smoothness as* A^0 *on any simply connected domain.*

Proof. The first assertion follows by direct calculation. The second is clear in the strictly hyperbolic case, for which the correspondence is also one-to-one; in the general (constant-multiplicity) case, it follows by a standard lemma of Kato [18]. □

Remark 2.6. Lemma 2.5 hints at the strategy we shall follow in carrying out energy estimates, which is to "effectively diagonalize" by the use of a symmetrizer. That is, rather than working with LAR as we should like, we work with $A^0A = L^*LA = L^*(LAR)L$, thereby avoiding the problem that there may exist a nonlinear change of coordinates with Jacobian L. Conjugation by L of course does not affect the energy estimates.

Finally, for convenience of the reader, we recall the standard relations

$$\langle W, SW_x\rangle = -\frac{1}{2}\langle W, S_xW\rangle \tag{2.2}$$

and

$$\frac{1}{2}\langle W_x, KW\rangle_t = \langle W_x, KW_t\rangle + \frac{1}{2}\langle W_x, K_tW\rangle + \frac{1}{2}\langle W, K_xW_t\rangle, \tag{2.3}$$

valid, respectively, for symmetric S and skew-symmetric K.

3. Energy estimates. In this section, we carry out the main work of the paper, establishing Proposition 1.4.

Perturbation equation. Define the nonlinear perturbation $U(x,t) := \tilde{U}(x+\delta(t), t) - \bar{U}(x)$ as in (1.7), where $\delta(t)$ (estimating shock location) is to be determined later; for definiteness, fix $\delta(0) = 0$. Substituting (1.7) into (1.2), we obtain

$$\tilde{U}_t + \tilde{A}\tilde{U}_x - \begin{pmatrix} 0 \\ q \end{pmatrix}(\tilde{U}) = \dot{\delta}\tilde{U}_x,$$

where $\tilde{A} := (df, dg)^t(\tilde{u}, \tilde{v})$, and thereby

$$(\tilde{U} - \bar{U})_t + (\tilde{A}\tilde{U}_x - \bar{A}\bar{U}_x) - \left(\begin{pmatrix} 0 \\ q \end{pmatrix}(\tilde{U}) - \begin{pmatrix} 0 \\ q \end{pmatrix}(\bar{U})\right) = \dot{\delta}(t)\tilde{U}_x,$$

where $\tilde{U}$ now denotes $\tilde{U}(x + \delta(t), t)$ and $\bar{U}$ denotes $\bar{U}(x)$.

Expanding $(\tilde{A}\tilde{U}_x - \bar{A}\bar{U}_x)$ using the quadratic Leibniz relation

$$\tilde{A}\tilde{U}_x - \bar{A}\bar{U}_x = \tilde{A}(\tilde{U}_x - \bar{U}_x) + (\tilde{A} - \bar{A})\bar{U}_x,$$

and Taylor expanding $(\binom{0}{q}(\tilde{U}) - \binom{0}{q}(\bar{U}))$ about $\tilde{U}$, we obtain the basic nonlinear perturbation equation

$$U_t - \tilde{A}U_x - \tilde{Q}U = M_1(U)\bar{U}_x + (0, I_r)^t M_2(U) + \dot{\delta}(t)(\bar{U}_x + U_x), \tag{3.1}$$

where $\tilde{Q} := (0, dq)^t(\tilde{u}, \tilde{v})$ and

$$\begin{aligned} M_1(U) &= \mathcal{O}(|U|) := \tilde{A}(x,t) - \bar{A}(x), \\ M_2(U) &= \mathcal{O}(|U|^2) := \begin{pmatrix} 0 \\ q \end{pmatrix}(\tilde{U}) - \begin{pmatrix} 0 \\ q \end{pmatrix}(\bar{U}) - \tilde{Q}(\tilde{U} - \bar{U}). \end{aligned}$$

Weighting matrix. Let $\tilde{A}^0 := A^0(\tilde{U})$ denote the symmetrizer of $\tilde{A}$ guaranteed by (A1), and factor $\tilde{A}^0\tilde{A} = (\tilde{A}^0)^{\frac{1}{2}}\tilde{O}\tilde{D}\tilde{O}^t(\tilde{A}^0)^{\frac{1}{2}}$, or, equivalently,

$$\tilde{A} = (\tilde{A}^0)^{-\frac{1}{2}}\tilde{O}\tilde{D}\tilde{O}^t(\tilde{A}^0)^{\frac{1}{2}}, \tag{3.2}$$

where $\tilde{O}$ is orthogonal, $\tilde{O}^t = \tilde{O}^{-1}$, and C^3 is a function of (u, v) (see Lemma 2.5) and $\tilde{D} =$ block-diag$\{\tilde{a}_1, \ldots, \tilde{a}_l\}$, where $\tilde{a}_j$ denote the eigenvalues of $\tilde{A}$, indexed in increasing order

$$\tilde{a}_1 \leq \cdots \leq \tilde{a}_k < 0 < \tilde{a}_{k+1} \leq \cdots \leq \tilde{a}_l.$$

Define the "Goodman-type" [12] weighting matrix $\alpha(x) :=$ block-diag$\{\alpha_1, \ldots, \alpha_l\}$, where $\alpha_j > 0$ are defined by ODE

$$\alpha_x = C_* \mathrm{sgn}\, a_j |\bar{U}_x|\, \alpha, \qquad \alpha(0) = 1,$$

with $C_* > 0$ a sufficiently large constant to be determined later. This definition, together with $|a_j| > 0$, (H2)(i), gives the key inequality

$$\alpha_x \bar{D} \geq \theta_1 C_* |\bar{U}_x| \alpha, \tag{3.3}$$

where $\bar{D} =$ block-diag$\{\bar{a}_1, \ldots, \bar{a}_l\}$ and $\bar{a}_j$ are the eigenvalues of $\bar{A}$.

Setting

$$\tilde{A}^0_\alpha := (\tilde{A}^0)^{\frac{1}{2}}\tilde{O}\,\alpha\,\tilde{O}^t(\tilde{A}^0)^{\frac{1}{2}}, \tag{3.4}$$

we have by factorization (3.2) that

$$\tilde{A}^0_\alpha\tilde{A} = [(\tilde{A}^0)^{\frac{1}{2}}\tilde{O}\,\alpha\,\tilde{O}^t(\tilde{A}^0)^{\frac{1}{2}}]\tilde{A} = (\tilde{A}^0)^{\frac{1}{2}}\tilde{O}(\alpha\tilde{D})\tilde{O}^t(\tilde{A}^0)^{\frac{1}{2}}. \tag{3.5}$$

Hence, $\tilde{A}^0_\alpha\tilde{A}$ is symmetric and the symmetric positive definite matrix $\tilde{A}^0_\alpha$ is also a viable symmetrizer for $\tilde{A}$.

Moreover, setting $L := \tilde{O}^t(\tilde{A}^0)^{\frac{1}{2}}$ and $R := (\tilde{A}^0)^{-\frac{1}{2}}\tilde{O}$, by Lemma 2.5 and constant multiplicity of eigenvalues of A, we have the freedom to smoothly (C^3) redefine L and R so that they take on prescribed values at $x \to \pm\infty$. Thus, by Lemma 2.4, we may assume without loss of generality that

$$\mathrm{Re}\ \text{block-diag}\left(\alpha\tilde{O}^t(\tilde{A}^0)^{\frac{1}{2}}\tilde{Q}(\tilde{A}^0)^{-\frac{1}{2}}\tilde{O}\right)_\pm \leq -\theta(C_*) < 0,$$

and thereby, appealing to (2.1), $k=0$,

$$\mathrm{Re}\ \text{block-diag}\big(\alpha\tilde O^t(\tilde A^0)^{\frac12}\tilde Q(\tilde A^0)^{-\frac12}\tilde O\big)\le -\theta(C_*)+C(C_*)\zeta+C|\bar U_x|\alpha, \tag{3.6}$$

where

$$\tilde A^0_\alpha\tilde Q=(\tilde A^0)^{\frac12}\tilde O\big[\alpha\,\tilde O^t(\tilde A^0)^{\frac12}\tilde Q(\tilde A^0)^{-\frac12}\tilde O\big]\tilde O^t(\tilde A^0)^{\frac12},$$

and the exponent θ now refers to the minimum of the constants used elsewhere in the argument and that appear in (2.1). The constants $\theta(C_*)$ and $C(C_*)$ measure the conditioning of matrices α and in fact decay (resp., grow) exponentially with respect to C_*; however, this is unimportant for our argument.

Define

$$K_1:=K\big(\tilde D,\alpha\,\tilde O^t(\tilde A^0)^{\frac12}\tilde Q(\tilde A^0)^{-\frac12}\tilde O+N\big),$$

where $K(\cdot)$ is as in Lemma 2.3, and N is *an arbitrary matrix with* $|N|_{C^1_{x,t}}\le C(C_*)$ *and vanishing on diagonal blocks*, to be determined later. Moreover, let $\tilde K_\alpha$ be the skew-symmetric matrix obtained from K_1 after conjugation by $(\tilde A^0)^{\frac12}\tilde O$, i.e.,

$$\tilde K_\alpha:=(\tilde A^0)^{\frac12}\tilde O K_1\tilde O^t(\tilde A^0)^{\frac12}.$$

For later use, note that, through smooth dependence on $\tilde U=\bar U+U$ and N,

$$|\tilde K_{\alpha,x}|,|\tilde K_{\alpha,t}|\le C(C_*). \tag{3.7}$$

We have, therefore,

$$\begin{aligned}&\mathrm{Re}(-\tilde K_\alpha\tilde A+\tilde A^0_\alpha\tilde Q+(\tilde A^0)^{\frac12}\tilde O N\tilde O^t(\tilde A^0)^{\frac12})\\&\quad=\mathrm{Re}\,(\tilde A^0)^{\frac12}\tilde O\big(-K_1\tilde D+\alpha\,\tilde O^t(\tilde A^0)^{\frac12}\tilde Q(\tilde A^0)^{-\frac12}\tilde O+N\big)\tilde O^t(\tilde A^0)^{\frac12}\\&\quad=\mathrm{Re}\,(\tilde A^0)^{\frac12}\tilde O\,\text{block-diag}\big(\alpha\,\tilde O^t(\tilde A^0)^{\frac12}\tilde Q(\tilde A^0)^{-\frac12}\tilde O\big)\tilde O^t(\tilde A^0)^{\frac12}\\&\quad\le-\theta(C_*)+C(C_*)\zeta+C|\bar U_x|\,\tilde A^0_\alpha\end{aligned} \tag{3.8}$$

by (3.6) together with Lemma 2.3. By (3.3), we have also

$$\begin{aligned}(\tilde A^0)^{\frac12}\tilde O(\alpha_x\tilde D)\tilde O^t(\tilde A^0)^{\frac12}&\ge(\tilde A^0)^{\frac12}\tilde O(\alpha_x\bar D)\tilde O^t(\tilde A^0)^{\frac12}-C(C_*)\,\zeta\\&\ge\theta C_*|\bar U_x|\tilde A^0_\alpha-C(C_*)\,\zeta\end{aligned} \tag{3.9}$$

for possibly still smaller $\theta>0$ (with ζ as defined in the statement of Proposition 1.4).

Friedrichs-type estimate. As in [24], we first perform a standard "Friedrichs-type" estimate for symmetric hyperbolic systems (see [8, 9]), now incorporating the weight α.

Differentiating (3.1) twice with respect to x, we obtain

$$U_{xxt}-(\tilde AU_x)_{xx}-(\tilde QU)_{xx}=(M(U)\bar U_x)_{xx}+(0,I_r)^tM_2(U)_{xx}+\dot\delta(t)(\bar U_{xxx}+U_{xxx}). \tag{3.10}$$

Taking the L^2 inner product $\tilde A^0_\alpha U_{xx}$ against U_{xxt}, we get

$$\frac12\langle\tilde A^0_\alpha U_{xx},U_{xx}\rangle_t=\langle\tilde A^0_\alpha U_{xx},U_{xxt}\rangle+\frac12\langle(\tilde A^0_\alpha)_tU_{xx},U_{xx}\rangle, \tag{3.11}$$

where $\tilde A^0_\alpha$ is defined in (3.4).

The second term can be easily bounded: indeed, using (3.1),

$$|(\tilde{A}^0_\alpha)_t|_{L^\infty} = \left|\frac{dA^0_\alpha}{d\tilde{U}}\right| |\tilde{U}_t| = \left|\frac{dA^0_\alpha}{d\tilde{U}}\right| |U_t| \le C(C_*)[|U|_{W^{1,\infty}} + |\dot{\delta}(t)|(|\bar{U}_x|_{L^\infty} + |U_x|_{L^\infty})],$$

and hence

$$\frac{1}{2}\langle (\tilde{A}^0_\alpha)_t U_{xx}, U_{xx}\rangle \le C(C_*)\,\zeta\,|U_{xx}|^2_{L^2} \tag{3.12}$$

(with ζ as defined in the statement of Proposition 1.4).

Let us consider the first term on the right-hand side of (3.11)

$$\begin{aligned}\langle \tilde{A}^0_\alpha U_{xx}, U_{xxt}\rangle &= \langle \tilde{A}^0_\alpha U_{xx}, (\tilde{A}U_x)_{xx}\rangle + \langle \tilde{A}^0_\alpha U_{xx}, (\tilde{Q}U)_{xx}\rangle \\ &\quad + \langle \tilde{A}^0_\alpha U_{xx}, (M_1(U)\bar{U}_x)_{xx}\rangle + \langle \tilde{A}^0_\alpha U_{xx}, (0, I_r)^t M_2(U)_{xx}\rangle \\ &\quad + \dot{\delta}(t)\,\langle \tilde{A}^0_\alpha U_{xx}, \bar{U}_{xxx} + U_{xxx}\rangle.\end{aligned} \tag{3.13}$$

Differentiating the first of the terms in (3.13), we get

$$\langle \tilde{A}^0_\alpha U_{xx}, (\tilde{A}U_x)_{xx}\rangle = \langle \tilde{A}^0_\alpha U_{xx}, \tilde{A}_{xx}U_x\rangle + 2\langle \tilde{A}^0_\alpha U_{xx}, \tilde{A}_x U_{xx}\rangle + \langle \tilde{A}^0_\alpha U_{xx}, \tilde{A}U_{xxx}\rangle. \tag{3.14}$$

Since

$$\begin{aligned}\tilde{A}_x &= \frac{d\tilde{A}}{d\tilde{U}}(\bar{U}_x + U_x),\\ \tilde{A}_{xx} &= \frac{d^2\tilde{A}}{d\tilde{U}^2}\,(\bar{U}_x\,\bar{U}_x + \bar{U}_x\,U_x + U_x\,\bar{U}_x + U_x\,U_x) + \frac{d\tilde{A}}{d\tilde{U}}\,(\bar{U}_{xx} + U_{xx}),\end{aligned}$$

the first two terms of (3.14) are bounded by

$$\begin{aligned}&\langle \tilde{A}^0_\alpha U_{xx}, \tilde{A}_{xx}U_x\rangle + 2\langle \tilde{A}^0_\alpha U_{xx}, \tilde{A}_x U_{xx}\rangle \le C(C_*)\,(\zeta + \bar{\zeta})\,|U_{xx}|^2_{L^2} \\ &\quad + C(C_*, \bar{\zeta})\,|U_x|^2_{L^2} + \left\langle \tilde{A}^0_\alpha U_{xx}, \frac{d\tilde{A}}{d\tilde{U}}\,\bar{U}_x\,U_{xx}\right\rangle;\end{aligned} \tag{3.15}$$

here, we have used Young's inequality to bound $\langle U_x, U_{xx}\rangle$ with $\bar{\zeta}|U_{xx}|^2_{L^2} + C|U_x|^2_{L^2}$, with $\bar{\zeta} > 0$ chosen such that $\zeta \ll \bar{\zeta} \ll 1$.

Using the symmetry of $\tilde{A}^0_\alpha$ and (2.2) with $S = \tilde{A}^0_\alpha\tilde{A}$, we find that the last term of (3.14) takes the form

$$\langle \tilde{A}^0_\alpha U_{xx}, \tilde{A}U_{xxx}\rangle = \langle U_{xx}, \tilde{A}^0_\alpha\tilde{A}\,U_{xxx}\rangle = -\frac{1}{2}\,\langle U_{xx}, (\tilde{A}^0_\alpha\tilde{A})_x U_{xx}\rangle. \tag{3.16}$$

Recalling (3.5), we have

$$(\tilde{A}^0_\alpha\tilde{A})_x = \frac{d\,(\tilde{A}^0_\alpha\tilde{A})}{d\tilde{U}}(\bar{U}_x + U_x) + (\tilde{A}^0)^{\frac{1}{2}}\tilde{O}\,(\alpha_x\tilde{D})\,\tilde{O}^t(\tilde{A}^0)^{\frac{1}{2}},$$

where, with slight abuse of notation,

$$\frac{d\,(\tilde{A}^0_\alpha\tilde{A})}{d\tilde{U}}\,W = \frac{d\,((\tilde{A}^0)^{\frac{1}{2}}\tilde{O})}{d\tilde{U}}\,W\,\alpha\,\tilde{D}\,\tilde{O}^t(\tilde{A}^0)^{\frac{1}{2}} + (\tilde{A}^0)^{\frac{1}{2}}\tilde{O}\,\alpha\,\frac{d\,(\tilde{D}\,\tilde{O}^t(\tilde{A}^0)^{\frac{1}{2}})}{d\tilde{U}}\,W. \tag{3.17}$$

Hence, we get

$$\begin{aligned}\langle \tilde A^0_\alpha U_{xx}, \tilde A U_{xxx}\rangle &\le -\frac{1}{2}\,\langle U_{xx}, (\tilde A^0)^{\frac12}\,\tilde O\,(\alpha_x \tilde D)\,\tilde O^t (\tilde A^0)^{\frac12} U_{xx}\rangle \\ &\quad + C(C_*)\,\zeta\,|U_{xx}|^2_{L^2} - \frac{1}{2}\left\langle U_{xx}, \frac{d\,(\tilde A^0_\alpha \tilde A)}{d\tilde U}\,\bar U_x\, U_{xx}\right\rangle.\end{aligned} \tag{3.18}$$

Summarizing, the first term on the right-hand side of (3.13) can be estimated by

$$\begin{aligned}\langle \tilde A^0_\alpha U_{xx}, (\tilde A U_x)_{xx}\rangle &\le C(C_*)\,(\zeta+\bar\zeta)\,|U_{xx}|^2_{L^2} + C(C_*,\bar\zeta)\,|U_x|^2_{L^2} \\ &\quad + \left\langle \tilde A^0_\alpha U_{xx}, \frac{d\tilde A}{d\tilde U}\,\bar U_x\, U_{xx}\right\rangle - \frac{1}{2}\,\langle U_{xx}, (\tilde A^0)^{\frac12}\,\tilde O\,(\alpha_x \tilde D)\,\tilde O^t (\tilde A^0)^{\frac12} U_{xx}\rangle \\ &\quad - \frac{1}{2}\left\langle U_{xx}, \frac{d\,(\tilde A^0_\alpha \tilde A)}{d\tilde U}\,\bar U_x\, U_{xx}\right\rangle.\end{aligned} \tag{3.19}$$

The second term in (3.13) can be dealt with similarly: since

$$\tilde Q_x = \frac{d\tilde Q}{d\tilde U}\,(\bar U_x + U_x), \quad \tilde Q_{xx} = \frac{d^2\tilde Q}{d\tilde U^2}\,(\bar U_x\,\bar U_x + \bar U_x\,U_x + U_x\,\bar U_x + U_x\,U_x) + \frac{d\tilde Q}{d\tilde U}\,(\bar U_{xx} + U_{xx}),$$

we have

$$\begin{aligned}\langle \tilde A^0_\alpha U_{xx}, (\tilde Q U)_{xx}\rangle &= \langle \tilde A^0_\alpha U_{xx}, \tilde Q_{xx} U\rangle + 2\langle \tilde A^0_\alpha U_{xx}, \tilde Q_x U_x\rangle + \langle \tilde A^0_\alpha U_{xx}, \tilde Q U_{xx}\rangle \\ &\le C(C_*)\,(\zeta+\bar\zeta)\,|U_{xx}|^2_{L^2} + C(C_*,\bar\zeta)\,|U|^2_{H^1} + \langle \tilde A^0_\alpha U_{xx}, \tilde Q U_{xx}\rangle,\end{aligned} \tag{3.20}$$

with $\bar\zeta$ as in the previous case.

The third term in (3.13) can be estimated by

$$\begin{aligned}\langle \tilde A^0_\alpha U_{xx}, (M_1(U)\bar U_x)_{xx}\rangle &\le C(C_*)\,\bar\zeta\,|U_{xx}|^2_{L^2} + C(C_*,\bar\zeta)\,|U|^2_{H^1} \\ &\qquad + \left\langle \tilde A^0_\alpha U_{xx}, \frac{dM_1}{dU}\,U_{xx}\,\bar U_x\right\rangle.\end{aligned} \tag{3.21}$$

The fourth term in (3.13) is easier: since

$$M_2(U)_{xx} = \frac{d^2 M_2}{dU^2}\,U_x\,U_x + \frac{dM_2}{dU}\,U_{xx},$$

we have (recall that $M_2(U) = \mathcal{O}(|U|^2)$)

$$\langle \tilde A^0_\alpha U_{xx}, (0, I_r)^t\, M_2(U)_{xx}\rangle \le C(C_*)\,\bar\zeta\,|U_{xx}|^2_{L^2} + C(C_*,\bar\zeta)\,|U_x|^2_{L^2}. \tag{3.22}$$

Finally, we estimate the last term in (3.13): $\dot\delta(t)\,\langle \tilde A^0_\alpha U_{xx}, \bar U_{xxx} + U_{xxx}\rangle$. Since

$$\dot\delta(t)\,\langle \tilde A^0_\alpha U_{xx}, \bar U_{xxx}\rangle \le C(C_*)|\dot\delta(t)|\,|U_{xx}|_{L^2} \le C(C_*)\,\bar\zeta\,|U_{xx}|^2_{L^2} + C(C_*,\bar\zeta)|\dot\delta(t)|^2$$

and, using (2.2) with $S = \tilde A^0_\alpha$,

$$\dot\delta(t)\,\langle \tilde A^0_\alpha U_{xx}, U_{xxx}\rangle = -\frac{1}{2}\dot\delta(t)\,\langle U_{xx}, (\tilde A^0_\alpha)_x U_{xx}\rangle \le C(C_*)\,\zeta\,|U_{xx}|^2_{L^2},$$

we have

$$\dot\delta(t)\,\langle \tilde A^0_\alpha U_{xx}, \bar U_{xxx}+U_{xxx}\rangle \le C(C_*)\,\zeta\,|U_{xx}|^2_{L^2}+C(C_*)|\dot\delta(t)|. \tag{3.23}$$

Collecting (3.12), (3.19), (3.20), (3.21), (3.22), and (3.23), we get

$$\begin{aligned}\frac12\langle \tilde A^0_\alpha U_{xx},U_{xx}\rangle_t \le{}& -\frac12\,\langle U_{xx},(\tilde A^0)^{\frac12}\tilde O\,(\alpha_x\tilde D)\,\tilde O^t(\tilde A^0)^{\frac12}U_{xx}\rangle+\langle U_{xx},\tilde A^0_\alpha\tilde Q U_{xx}\rangle\\ &-\frac12\left\langle U_{xx},\frac{d\,(\tilde A^0_\alpha\tilde A)}{d\tilde U}\,\bar U_x\,U_{xx}\right\rangle+\left\langle U_{xx},\tilde A^0_\alpha\left(\frac{d\tilde A}{d\tilde U}\,\bar U_x\,U_{xx}+\frac{dM_1}{dU}\,U_{xx}\,\bar U_x\right)\right\rangle\\ &+C(C_*)\,(\zeta+\bar\zeta)\,|U_{xx}|^2_{L^2}+C(C_*,\bar\zeta)\,(|U|^2_{H^1}+|\dot\delta(t)|^2).\end{aligned}\tag{3.24}$$

Let us consider the term in (3.24) containing $\frac{d\,(\tilde A^0_\alpha\tilde A)}{d\tilde U}$. By (3.17),

$$\begin{aligned}\left\langle U_{xx},\frac{d\,(\tilde A^0_\alpha\tilde A)}{d\tilde U}\,\bar U_x\,U_{xx}\right\rangle &= s\left\langle U_{xx},\frac{d\,((\tilde A^0)^{\frac12}\tilde O)}{d\tilde U}\,\bar U_x\,\alpha\,\tilde D\,\tilde O^t(\tilde A^0)^{\frac12}\,U_{xx}\right\rangle\\ &\quad+\left\langle U_{xx},(\tilde A^0)^{\frac12}\tilde O\,\alpha\,\frac{d\,(\tilde D\,\tilde O^t(\tilde A^0)^{\frac12})}{d\tilde U}\,\bar U_x\,U_{xx}\right\rangle\\ &=\langle U_{xx},(\tilde A^0)^{\frac12}\tilde O\,P_1\,\alpha\,\tilde O^t(\tilde A^0)^{\frac12}\,U_{xx}\rangle\\ &\quad+\langle U_{xx},(\tilde A^0)^{\frac12}\tilde O\,\alpha\,P_2\,\tilde O^t(\tilde A^0)^{\frac12}\,U_{xx}\rangle,\end{aligned}$$

where

$$P_1=\tilde O^t\,(\tilde A^0)^{-\frac12}\,\frac{d\,((\tilde A^0)^{\frac12}\tilde O)}{d\tilde U}\,\bar U_x\,\tilde D\quad\text{and}\quad P_2=\frac{d\,(\tilde D\,\tilde O^t(\tilde A^0)^{\frac12})}{d\tilde U}\,\bar U_x\,(\tilde A^0)^{-\frac12}\,\tilde O.$$

Similarly,

$$\left\langle U_{xx},\tilde A^0_\alpha\left(\frac{d\tilde A}{d\tilde U}\,\bar U_x\,U_{xx}+\frac{dM_1}{dU}\,U_{xx}\,\bar U_x\right)\right\rangle=\langle U_{xx},(\tilde A^0)^{\frac12}\tilde O\,\alpha\,P_3\tilde O^t(\tilde A^0)^{\frac12}\,U_{xx}\rangle,$$

where

$$P_3=\tilde O^t(\tilde A^0)^{\frac12}\,\left(\frac{d\tilde A}{d\tilde U}+\frac{dM_1}{dU}\right)\,\bar U_x.$$

The terms

$$-\frac12\langle U_{xx},(\tilde A^0)^{\frac12}\tilde O\,P_1\,\alpha\,\tilde O^t(\tilde A^0)^{\frac12}\,U_{xx}\rangle+\left\langle U_{xx},(\tilde A^0)^{\frac12}\tilde O\,\alpha\,\left(P_3-\frac12\,P_2\right)\,\tilde O^t(\tilde A^0)^{\frac12}\,U_{xx}\right\rangle$$

can be rewritten as the sum of two terms, one taking into account "off-block-diagonal" parts (meaning off-block-diagonal after conjugation by $(\tilde A^0)^{\frac12}\tilde O$) and the other taking into account "block-diagonal" parts. We denote the first one as

$$\langle U_{xx},(\tilde A^0)^{\frac12}\tilde O N\tilde O^t(\tilde A^0)^{\frac12}U_{xx}\rangle,$$

where $|N|_{C^1(x,t}\le C(C_*)$ indeed holds, since N by definition is of form $J(\tilde U,\alpha,\bar U_x)$ with $J(\cdot)$ smooth, and $|\tilde U|_{C^1(x,t)}\le C\zeta$, $|\alpha|_{C^1(x,t)}\le C_*\sup|\alpha|\le C(C_*)$. The "block-diagonal" parts (meaning block-diagonal after conjugation by $(\tilde A^0)^{\frac12}\tilde O$) of these error terms may be estimated by $C\,\langle U_{xx}, |\bar U_x|\,\tilde A^0_\alpha\, U_{xx}\rangle$ for some constant $C>0$ (independent of α, C_*).

Hence, we get the final form of the Friedrichs-type estimate:

$$\begin{aligned}\frac12\langle \tilde A^0_\alpha U_{xx}, U_{xx}\rangle_t \le &-\frac12\,\langle U_{xx}, (\tilde A^0)^{\frac12}\,\tilde O\,(\alpha_x \tilde D)\,\tilde O^t(\tilde A^0)^{\frac12} U_{xx}\rangle + \langle U_{xx}, \tilde A^0_\alpha \tilde Q U_{xx}\rangle\\ &+\langle U_{xx}, (\tilde A^0)^{\frac12}\tilde O N \tilde O^t (\tilde A^0)^{\frac12} U_{xx}\rangle + C\langle U_{xx}, |\bar U_x|\,\tilde A^0_\alpha U_{xx}\rangle\\ &+C(C_*)\,(\zeta+\bar\zeta)\,|U_{xx}|^2_{L^2} + C(C_*,\bar\zeta)\,(|U|^2_{H^1}+|\dot\delta(t)|^2).\end{aligned}\tag{3.25}$$

Remark 3.1. The treatment of error term N, above, we regard as the most delicate and novel aspect of our argument. Without complete cancellation of off-diagonal terms, we see no way that such a "Goodman-type" estimate can close, due to exponential growth in C_* of $\sup|\alpha|$.

Kawashima-type estimate. Next, we perform a "Kawashima-type" estimate of the type formalized in [17]. Applying (2.3) to $W = U_x$ and $K = \tilde K_\alpha$,

$$\frac12\langle U_{xx}, \tilde K_\alpha U_x\rangle_t = \langle U_{xx}, \tilde K_\alpha U_{xt}\rangle + \frac12\langle U_{xx}, (\tilde K_\alpha)_t U_x\rangle + \frac12\langle U_x, (\tilde K_\alpha)_x U_{xt}\rangle.\tag{3.26}$$

Thanks to (3.7) and to Young's inequality, the second term is easily bounded by

$$\frac12\langle U_{xx}, (\tilde K_\alpha)_t U_x\rangle \le C\,\bar\zeta\,|U_{xx}|^2_{L^2} + C(\bar\zeta, C_*)\,|U_x|^2_{L^2},\tag{3.27}$$

with $\bar\zeta$ as previously chosen.

Differentiating (3.1) with respect to x, we obtain

$$U_{xt} + (\tilde A U_x)_x - (\tilde Q U)_x = (M_1(U)\bar U_x)_x + (0, I_r)^t M_2(U)_x + \dot\delta(t)(\bar U_{xx} + U_{xx}),\tag{3.28}$$

and hence (twice) the last term in (3.26) can be rewritten as

$$\begin{aligned}\langle U_x, (\tilde K_\alpha)_x U_{xt}\rangle = &-\left\langle U_x, (\tilde K_\alpha)_x \frac{d\tilde A}{d\tilde U}(\bar U_x + U_x)\,U_x\right\rangle - \langle U_x, (\tilde K_\alpha)_x \tilde A U_{xx}\rangle\\ &+\left\langle U_x, (\tilde K_\alpha)_x \frac{d\tilde Q}{d\tilde U}(\bar U_x + U_x)\,U_x\right\rangle + \langle U_x, (\tilde K_\alpha)_x \tilde Q\,U_x\rangle\\ &+\left\langle U_x, (\tilde K_\alpha)_x \frac{dM_1}{dU}\,U_x\,\bar U_x\right\rangle + \langle U_x, (\tilde K_\alpha)_x M_1(U)\bar U_{xx}\rangle\\ &+\left\langle U_x, (\tilde K_\alpha)_x (0, I_r)^t\frac{dM_2}{dU}\,U_x\right\rangle + \dot\delta(t)\langle U_x, (\tilde K_\alpha)_x(\bar U_{xx} + U_{xx})\rangle.\end{aligned}$$

All of the terms not containing U_{xx} can be estimated by $C(C_*)(|U|^2_{H^1} + |\dot\delta(t)|^2)$ (having used once more (3.7)). For the remaining terms,

$$-\langle U_x, (\tilde K_\alpha)_x \tilde A U_{xx}\rangle + \dot\delta(t)\langle U_x, (\tilde K_\alpha)_x\, U_{xx}\rangle \le C\,\bar\zeta\,|U_{xx}|^2_{L^2} + C(C_*,\bar\zeta)\,|U_x|^2_{L^2}.$$

Hence,

$$\frac12\langle U_x, (\tilde K_\alpha)_x U_{xt}\rangle \le C\,\bar\zeta\,|U_{xx}|^2_{L^2} + C(C_*,\bar\zeta)(|U|^2_{H^1} + |\dot\delta(t)|^2).\tag{3.29}$$

Inserting (3.28), the first term on the right-hand side of (3.26) becomes

$$\begin{aligned}\langle U_{xx}, \tilde K_\alpha U_{xt}\rangle = &-\left\langle U_{xx}, \tilde K_\alpha \frac{d\tilde A}{d\tilde U}(\bar U_x + U_x)\, U_x\right\rangle - \langle U_{xx}, \tilde K_\alpha\, \tilde A U_{xx}\rangle\\ &+\left\langle U_{xx}, \tilde K_\alpha \frac{d\tilde Q}{d\tilde U}(\bar U_x + U_x)\, U_x\right\rangle + \langle U_{xx}, \tilde K_\alpha\, \tilde Q\, U_x\rangle\\ &+\left\langle U_{xx}, \tilde K_\alpha \frac{dM_1}{dU}\, U_x\, \bar U_x\right\rangle + \langle U_{xx}, \tilde K_\alpha\, M_1(U)\bar U_{xx}\rangle\\ &+\left\langle U_{xx}, \tilde K_\alpha\, (0, I_r)^t \frac{dM_2}{dU}\, U_x\right\rangle + \dot\delta(t)\langle U_{xx}, (\tilde K_\alpha)_x(\bar U_{xx} + U_{xx})\rangle.\end{aligned}$$

The terms containing at least one of $U_x, U, \dot\delta$ can be estimated (applying Young's inequality) by $C(\bar\zeta + \zeta)|U_{xx}|^2_{L^2} + C(\bar\zeta, C_*)(|U|^2_{H^1} + |\dot\delta(t)|^2)$; hence,

$$\begin{aligned}\langle U_{xx}, \tilde K_\alpha U_{xt}\rangle \le &-\langle U_{xx}, \tilde K_\alpha\, \tilde A U_{xx}\rangle + C(\bar\zeta + \zeta)|U_{xx}|^2_{L^2}\\ &+ C(\bar\zeta, C_*)(|U|^2_{H^1} + |\dot\delta(t)|^2).\end{aligned}\tag{3.30}$$

Finally, (3.26), (3.27), (3.29), and (3.30) give

$$\begin{aligned}\frac12\langle U_{xx}, \tilde K_\alpha U_x\rangle_t \le &-\langle U_{xx}, \tilde K_\alpha\, \tilde A U_{xx}\rangle + C(\bar\zeta + \zeta)|U_{xx}|^2_{L^2}\\ &+ C(C_*, \bar\zeta)(|U|^2_{H^1} + |\dot\delta(t)|^2).\end{aligned}\tag{3.31}$$

Adding (3.25) and (3.31), we obtain

$$\begin{aligned}&\frac12(\langle \tilde A^0_\alpha U_{xx}, U_{xx}\rangle_t + \langle U_{xx}, \tilde K_\alpha U_x\rangle)_t \le -\frac12\,\langle U_{xx}, (\tilde A^0)^{\frac12}\tilde O\,(\alpha_x \tilde D)\,\tilde O^t(\tilde A^0)^{\frac12} U_{xx}\rangle\\ &\quad + C\,\langle U_{xx}, |\bar U_x|\,\tilde A^0_\alpha U_{xx}\rangle + \langle U_{xx}, (-\tilde K_\alpha\, \tilde A + \tilde A^0_\alpha \tilde Q + (\tilde A^0)^{\frac12}\tilde O N \tilde O^t(\tilde A^0)^{\frac12})\, U_{xx}\rangle\\ &\quad + C(C_*)\,(\zeta + \bar\zeta)\,|U_{xx}|^2_{L^2} + C(C_*, \bar\zeta)\,(|U|^2_{H^1} + |\dot\delta(t)|^2).\end{aligned}\tag{3.32}$$

Recalling (3.8) and (3.9), we obtain, finally,

$$\begin{aligned}&(\langle \tilde A^0 U_{xx}, U_{xx}\rangle + \langle U_{xx}, \tilde K_\alpha U_x\rangle)_t \le -\,(\theta\, C_* - C)\,\langle U_{xx}, |\bar U_x|\,\tilde A^0_\alpha\, U_{xx}\rangle\\ &\quad + (-\theta(C_*) + C(C_*)(\zeta + \bar\zeta))|U_{xx}|^2_{L^2} + C(\bar\zeta, C_*)(|U|^2_{H^1} + |\dot\delta(t)|^2)\\ &\quad \le -\frac12\theta(C_*)|U_{xx}|^2_{L^2} + C(\bar\zeta, C_*)(|U|^2_{H^1} + |\dot\delta(t)|^2),\end{aligned}\tag{3.33}$$

provided that C_* is taken sufficiently large and $\bar\zeta$, ζ sufficiently small that $C \le \theta C_*$ and $-\theta(C_*) + C(C_*)(\zeta + \bar\zeta) \le -\theta(C_*)/2$.

Given $M > 0$, let us set

$$\mathcal E(U) := \langle \tilde A^0 U_{xx}, U_{xx}\rangle + \langle U_{xx}, \tilde K_\alpha U_x\rangle + M|U|^2_{L^2}.\tag{3.34}$$

Since, for $U \in H^2$, $|U_x|_{L^2}$ can be bounded by $C\left(|U|_{L^2} + |U_{xx}|_{L^2}\right)$ for some $C > 0$, then the functional defined in (3.34) is equivalent to $|U|^2_{H^2}$ if M is large enough. Moreover, from (3.1) it follows that

$$\frac{d|U|^2_{L^2}}{dt} \le C(|U|^2_{H^1} + |\dot\delta(t)|^2).$$

Therefore,

$$\frac{d\mathcal{E}}{dt} \leq -\frac{1}{2}\theta(C_*)\,|U_{xx}|^2_{L^2} + C(|U|^2_{H^1} + |\dot{\delta}(t)|^2)$$

for some $C > 0$. Passing through the Fourier transform, it is easy to see that for $U \in H^2$ there holds

$$|U_x|^2_{L^2} \leq C|U|^2_{L^2} + \frac{1}{C}|U_{xx}|^2_{L^2} \qquad \forall\, C > 0. \tag{3.35}$$

Hence, using $\mathcal{E} \geq C|U_{xx}|^2_{L^2}$ and choosing C big enough in (3.35), we get

$$\frac{d\mathcal{E}}{dt} \leq -\theta\,\mathcal{E} + C(|U|^2_{L^2} + |\dot{\delta}(t)|^2)$$

for some $C, \theta > 0$. Multiplying by $e^{\theta t}$ and integrating in time from 0 to t, we get (1.8), and the proof of Proposition 1.4 is complete.

Remark 3.2. The energy estimate (1.9) can be deduced from (3.33) as follows. Integrating (3.33) from 0 to t yields

$$\begin{aligned}(\langle \tilde{A}^0_\alpha U_{xx}, U_{xx}\rangle + \langle U_{xx}, \tilde{K}_\alpha U_x\rangle)|_0^t + \theta(C_*)\int_0^t |U_{xx}|^2_{L^2}(s)\,ds \\ \leq C(C_*, \bar{\zeta})\int_0^t (|U|^2_{H^1} + |\dot{\delta}|^2)(s)\,ds.\end{aligned}$$

Rearranging, using positive definiteness of $\tilde{A}^0_\alpha$, using Young's inequality to bound

$$\langle U_{xx}, K_\alpha U_x\rangle(t) \leq \bar{\zeta}|U_{xx}|^2_{L^2}(t) + C\bar{\zeta}^{-1}|U_x|^2_{L^2}(t) \leq \bar{\zeta}|U_{xx}|^2_{L^2}(t) + C\bar{\zeta}^{-1}\zeta^2,$$

and recalling, by assumption, that $(\langle \tilde{A}^0_\alpha U_x, U_x\rangle + \langle U_x, \tilde{K}_\alpha U\rangle)(0) \leq C\zeta^2$, we obtain (1.9) as claimed.

4. Nonlinear stability. We now establish Theorem 1.2 on nonlinear stability.

Linearized estimates. Linearizing (1.2) about the stationary solution $(\bar{u}, \bar{v})$, we obtain the linearized equations

$$U_t = \mathcal{L}U := -(AU)_x + QU, \tag{4.1}$$

where

$$A := \begin{pmatrix} df \\ dg \end{pmatrix}(\bar{u}, \bar{v}), \quad Q := \begin{pmatrix} 0 \\ dq(\bar{u}, \bar{v}) \end{pmatrix}, \quad \text{and} \quad U := \begin{pmatrix} u \\ v \end{pmatrix} \qquad (u \in \mathbb{R}^n,\ v \in \mathbb{R}^r).$$

Define the associated Green distribution $G(x, t; y)$ by

$$(\partial_t - \mathcal{L})G(x, t; y) = \delta_{(y,0)}(x, t).$$

We have the following bounds established in [24, Proposition 1.11] and [24, Lemmas 7.1–7.5]. (See also the "notes" below (7.22) of [24], which is used in the short-time estimate for $|e_y(\cdot, t)|_{L^p}$.)

Proposition 4.1 (see [24]). *Assuming* (A1)–(A2) *and* (H0)–(H5), *the Green distribution G may be decomposed as*

$$G = E + \tilde{G} + H, \tag{4.2}$$

where $E(x,t;y)=e(y,t)\,\bar U_x(x)$, *with*

$$
\begin{aligned}
|e_y(\cdot,t)|_{L^p} &\le Ct^{\frac12}(1+t)^{-\frac12(1-\frac1p)-\frac12}, \qquad |e_t(\cdot,t)|_{L^p}\le Ct^{-\frac12(1-\frac1p)},\\
|e_{ty}(\cdot,t)|_{L^p} &\le Ct^{-\frac12(1-\frac1p)-\frac12},\\
\left|\int_{\mathbb{R}}\tilde G(\cdot,t;y)\,f(y)\,dy\right|_{L^p} &\le C(1+t)^{-\frac12(1-\frac1r)}|f|_{L^q},\\
\left|\int_{\mathbb{R}}\tilde G(\cdot,t;y)\,(0,I_r)^t\,f(y)\,dy\right|_{L^p} &\le C(1+t)^{-\frac12(1-\frac1r)-\frac12}|f|_{L^q},\\
\left|\int_{\mathbb{R}}\tilde G_y(\cdot,t;y)\,f(y)\,dy\right|_{L^p} &\le C(1+t)^{-\frac12(1-\frac1r)-\frac12}|f|_{L^q}+Ce^{-\eta t}|f|_{L^p},
\end{aligned}
$$

and

$$
\left|\int_{\mathbb{R}}H(\cdot,t;y)\,f(y)\,dy\right|_{L^p}\le Ce^{-\eta t}|f|_{L^p}
$$

for all $t\ge 0$, *some* C, $\eta>0$, *for any* $1\le r\le p$ *and* $f\in L^q$ *(resp.,* L^p*), where* $1/r+1/q=1+1/p$.

Here, the "excited" component E accounts for contributions in the direction of the translational zero eigenfunction $\bar U_x$, while the "hyperbolic" component H accounts for propagation of signals along hyperbolic characteristics, its time-exponential damping a consequence of the genuine coupling condition (1.3). The reduced Green distribution $\tilde G$, accounting for long-time behavior in the far fields, is approximately a sum of Gaussian signals scattered by the shock layer. For further discussion, see [24].

Proof of Theorem 1.2. We first treat the case $p\ge 2$, afterward extending to $p\le 2$ by a bootstrap argument. ∎

stability, 2 ≤ ≤ ∞. Define the nonlinear perturbation

$$
U(x,t):=\begin{pmatrix}u\\ v\end{pmatrix}(x+\delta(t),t)-\begin{pmatrix}\bar u\\ \bar v\end{pmatrix}(x)=\tilde U(x+\delta(t),t)-\bar U(x),
$$

where $\delta(t)$ (estimating shock location) is to be determined later; for definiteness, fix $\delta(0)=0$. Then

$$
U_t-\mathcal{L}U=N_1(U)_x+(0,I_r)^tN_2(U)+\dot\delta(t)(\bar U_x+U_x),
$$

where

$$
N_j(U,U)=\mathcal{O}(|U|^2)\qquad\text{and}\qquad N_j(U,U)_x=\mathcal{O}(|U||U_x|)
$$

so long as $|U|$ remains bounded. By Duhamel's principle, and the fact that

$$
\int_{\mathbb{R}}G(x,t;y)\bar U_x(y)dy=e^{\mathcal{L}t}\bar U_x(x)=\bar U_x(x),
$$

we have

$$
\begin{aligned}
U(x,t)=\int_{\mathbb{R}}G(x,t;y)U_0(y)\,dy&-\int_0^t\!\!\int_{\mathbb{R}}G_y(x,t-s;y)(N_1(U)+\dot\delta U)(y,s)\,dy\,ds\\
&+\int_0^t\!\!\int_{\mathbb{R}}G(x,t-s;y)(0,I_r)^tN_2(U)(y,s)\,dy\,ds+\delta(t)\bar U_x.
\end{aligned}
$$

Defining the *instantaneous shock location*

$$\delta(t) = -\int_{\mathbb{R}} e(y,t)U_0(y)\,dy + \int_0^t\!\!\int_{\mathbb{R}} e_y(y,t-s)(N_1(U)+\dot{\delta}U)(y,s)\,dy\,ds, \tag{4.3}$$

where E, e are defined as in Proposition 4.1, and recalling decomposition (4.2), we thus obtain the *reduced equations*

$$\begin{aligned} U(x,t) = {} & \int_{\mathbb{R}} (H+\tilde{G})(x,t;y)U_0(y)\,dy \\ & + \int_0^t\!\!\int_{\mathbb{R}} H(x,t-s;y)(N_1(U)_x + (0,I_r)^t N_2(U) + \dot{\delta}U_x)(y,s)\,dy\,ds \\ & - \int_0^t\!\!\int_{\mathbb{R}} \tilde{G}_y(x,t-s;y)(N_1(U)+\dot{\delta}U)(y,s)\,dy\,ds \\ & + \int_0^t\!\!\int_{\mathbb{R}} \tilde{G}(x,t-s;y)(0,I_r)^t N_2(U)\,dy\,ds, \end{aligned} \tag{4.4}$$

and differentiating (4.3) with respect to t and using $|e_y(\cdot,s)|_{L^1} \to 0$ as $t \to 0$,

$$\dot{\delta}(t) = -\int_{\mathbb{R}} e_t(y,t)U_0(y)\,dy + \int_0^t\!\!\int_{\mathbb{R}} e_{yt}(y,t-s)(N_1(U)+\dot{\delta}U)(y,s)\,dy\,ds. \tag{4.5}$$

Define

$$\zeta(t) := \sup_{0\le s\le t,\, 2\le p\le\infty} \big[\, |U(\cdot,s)|_{L^p}(1+s)^{\frac{1}{2}(1-\frac{1}{p})} + |\dot{\delta}(s)|(1+s)^{\frac{1}{2}} + |\delta(s)|\big]. \tag{4.6}$$

We shall establish the following claim.

Claim. *For all $t \ge 0$ for which a solution exists with ζ uniformly bounded by some fixed, sufficiently small constant, there holds*

$$\zeta(t) \le C_2(|U_0|_{L^1\cap H^2} + \zeta(t)^2).$$

From this result, it follows by continuous induction that, provided $|U_0|_{L^1\cap H^2} < 1/4C_2^2$, there holds

$$\zeta(t) \le 2\,C_2|U_0|_{L^1\cap H^2} \tag{4.7}$$

for all $t \ge 0$ such that ζ remains small. By standard short-time theory/local well-posedness in H^2, and the standard principle of continuation, there exists a solution $U(\cdot,t) \in H^2$ on the open time-interval for which $|U|_{H^2}$ remains bounded, and on this interval ζ is well-defined and continuous. Now, let $[0,T)$ be the maximal interval on which $|U|_{H^2}$ remains strictly bounded by some fixed, sufficiently small constant $\delta > 0$. By Proposition 1.4 and the one-dimensional Sobolev bound $|U|_{W^{1,\infty}} \le C|U|_{H^2}$, we have

$$\begin{aligned} |U(t)|_{H^2}^2 & \le C|U(0)|_{H^2}^2 e^{-\theta t} + C\int_0^t e^{-\theta_2(t-\tau)}(|U|_{L^2}^2 + |\dot{\delta}|^2)(\tau)\,d\tau \\ & \le C_2\big(|U(0)|_{H^2}^2 + \zeta(t)^2\big)(1+t)^{-\frac{1}{2}}, \end{aligned} \tag{4.8}$$

and so the solution continues so long as ζ remains small, with bound (4.7), at once yielding existence and the claimed sharp $L^p \cap H^2$ bounds, $2 \le p \le \infty$.

Thus, it remains only to establish the claim above.

Proof of claim. We must show that each of the quantities $|U|_{L^p}(1+s)^{\frac{1}{2}(1-\frac{1}{p})}$, $|\dot\delta|(1+s)^{\frac{1}{2}}$, and $|\delta|$ is separately bounded by

$$C(|U_0|_{L^1\cap H^2}+\zeta(t)^2)$$

for some $C>0$, all $0\le s\le t$, so long as ζ remains sufficiently small. By (4.4)–(4.5), we have

$$\begin{aligned}
|U|_{L^p}(t) \le{}& \left|\int_{\mathbb{R}}(H+\tilde G)(x,t;y)U_0(y)\,dy\right|_{L^p}\\
&+\left|\int_0^t\!\!\int_{\mathbb{R}} H(x,t-s;y)\,N_1(U)_y(y,s)\,dy\,ds\right|_{L^p}\\
&+\left|\int_0^t\!\!\int_{\mathbb{R}} H(x,t-s;y)\,(0,I_r)^t N_2(U)(y,s)\,dy\,ds\right|_{L^p}\\
&+\left|\int_0^t\!\!\int_{\mathbb{R}} H(x,t-s;y)\,\dot\delta\,U_x(y,s)\,dy\,ds\right|_{L^p}\\
&+\left|\int_0^t\!\!\int_{\mathbb{R}} \tilde G_y(x,t-s;y)\,N_1(U)(y,s)\,dy\,ds\right|_{L^p}\\
&+\left|\int_0^t\!\!\int_{\mathbb{R}} \tilde G(x,t-s;y)\,(0,I_r)^t N_2(U)(y,s)\,dy\,ds\right|_{L^p}\\
&+\left|\int_0^t\!\!\int_{\mathbb{R}} \tilde G_y(x,t-s;y)\,\dot\delta\,U(y,s)dy\,ds\right|_{L^p}\\
=:{}& I_a+I_b+I_c+I_d+I_e+I_f+I_g,
\end{aligned}\tag{4.9}$$

(4.10)

$$|\dot\delta|(t)\le\left|\int_{\mathbb{R}} e_t(y,t)\,U_0(y)\,dy\right|+\left|\int_0^t\!\!\int_{\mathbb{R}} e_{yt}(y,t-s)\dot\delta U(y,s)\,dy\,ds\right|=:II_a+II_b,$$

and

(4.11)

$$|\delta|(t)\le\left|\int_{\mathbb{R}} e(y,t)U_0(y)dy\right|+\left|\int_0^t\!\!\int_{\mathbb{R}} e_y(y,t-s)\dot\delta\,U(y,s)\,dy\,ds\right|=:III_a+III_b.$$

We estimate each term in turn, following the approach of [24, 25, 27, 37]. Applying the bounds of Proposition 4.1, we find that the linear term I_a satisfies

$$\begin{aligned}
I_a &\le |U|_{L^p}(t)\left|\int_{\mathbb{R}} H\,U_0\,dy\right|_{L^p}+\left|\int_{\mathbb{R}}\tilde G\,U_0\ dy\right|_{L^p}\\
&\le Ce^{-\theta t}|U_0|_{L^p}+C(1+t)^{-\frac14}|U_0|_{L^1}\le C|U_0|_{L^1\cap H^2}(1+t)^{-\frac14}.
\end{aligned}\tag{4.12}$$

Likewise, applying the bounds of Proposition 4.1 together with definition (4.6) and energy estimate (4.8), we have

$$\begin{aligned} I_b &= \left| \int_0^t\!\!\int_{\mathbb{R}} H\, N_1(U)_y \, dy\, ds \right|_{L^p} \le C \int_0^t e^{-\eta(t-s)} |U|_{L^\infty} |U_x|_{L^p}(s)\, ds \\ &\le C \int_0^t e^{-\eta(t-s)} |U|_{L^\infty} |U|_{H^2}(s)\, ds \le C\zeta(t)^2 \int_0^t e^{-\eta(t-s)} (1+s)^{-\frac34}\, ds \\ &\le C\zeta(t)^2 (1+t)^{-\frac34}, \end{aligned} \tag{4.13}$$

$$\begin{aligned} I_c &= \left| \int_0^t\!\!\int_{\mathbb{R}} H\, (0, I_r)^t N_2(U)\, dy\, ds \right|_{L^p} \le C \int_0^t e^{-\eta(t-s)} |U|_{L^\infty} |U|_{L^p}(s)\, ds \\ &\le C\zeta(t)^2 \int_0^t e^{-\eta(t-s)} (1+s)^{-\frac34}\, ds \le C\zeta(t)^2 (1+t)^{-\frac34}, \end{aligned} \tag{4.14}$$

$$\begin{aligned} I_d &= \left| \int_0^t\!\!\int_{\mathbb{R}} H\, \dot\delta\, U_x \, dy\, ds \right|_{L^p} \le C \int_0^t e^{-\eta(t-s)} |\dot\delta| |U_x|_{L^p}(s)\, ds \\ &\le C\zeta(t)^2 \int_0^t e^{-\eta(t-s)} (1+s)^{-\frac34}\, ds \le C\zeta(t)^2 (1+t)^{-\frac34} \end{aligned} \tag{4.15}$$

(Proposition 4.1, $2 \le p = q \le \infty$, $r = 1$) and

$$\begin{aligned} I_e &= \left| \int_0^t\!\!\int_{\mathbb{R}} \tilde G_y\, N_1(U)\, dy\, ds \right|_{L^p} \le C \int_0^t \big(1+(t-s)\big)^{-\frac12(1-\frac1p)-\frac14} |U|_{L^\infty} |U|_{L^2}(s)\, ds \\ &\le C\zeta(t)^2 \int_0^t \big(1+(t-s)\big)^{-\frac12(1-\frac1p)-\frac14} (1+s)^{-\frac34}\, ds \le C\zeta(t)^2 (1+t)^{-\frac12(1-\frac1p)}, \end{aligned} \tag{4.16}$$

$$\begin{aligned} I_f &= \left| \int_0^t\!\!\int_{\mathbb{R}} \tilde G\, (0, I_r)^t N_2(U)\, dy\, ds \right|_{L^p} \le C \int_0^t \big(1+(t-s)\big)^{-\frac12(1-\frac1p)-\frac14} |U|_{L^\infty} |U|_{L^2}(s)\, ds \\ &\le C\zeta(t)^2 \int_0^t \big(1+(t-s)\big)^{-\frac12(1-\frac1p)-\frac14} (1+s)^{-\frac34}\, ds \le C\zeta(t)^2 (1+t)^{-\frac12(1-\frac1p)}, \end{aligned} \tag{4.17}$$

$$\begin{aligned} I_g &= \left| \int_0^t\!\!\int_{\mathbb{R}} \tilde G_y\, \dot\delta\, U\, dy\, ds \right|_{L^p} \le C \int_0^t \big(1+(t-s)\big)^{-\frac12(1-\frac1p)-\frac14} |\dot\delta| |U|_{L^2}(s)\, ds \\ &\le C\zeta(t)^2 \int_0^t \big(1+(t-s)\big)^{-\frac12(1-\frac1p)-\frac14} (1+s)^{-\frac34}\, ds \le C\zeta(t)^2 (1+t)^{-\frac12(1-\frac1p)} \end{aligned} \tag{4.18}$$

(Proposition 4.1, $2 \le p \le \infty$, $q = 2$). Summing bounds (4.12)–(4.18), we obtain the desired bound on $|U|_{L^p}$.

Similarly, applying the bounds of Proposition 4.1 together with definition (4.6), we find that

$$II_a = \left| \int_{\mathbb{R}} e_t\, U_0\, dy \right| \le |e_t(y,t)|_{L^\infty}(t) |U_0|_{L^1} \le C |U_0|_{L^1} (1+t)^{-\frac12} \tag{4.19}$$

and

(4.20)
$$II_b = \left|\int_0^t\!\!\int_{\mathbb{R}} e_{yt}\,\dot\delta U\,dy\,ds\right| \le \int_0^t |e_{yt}|_{L^2}(t-s)|\dot\delta||U|_{L^2}(s)\,ds$$
$$\le C\zeta(t)^2\int_0^t (t-s)^{-3/4}(1+s)^{-\frac{3}{4}}\,ds \le C\zeta(t)^2(1+t)^{-\frac{1}{2}},$$

while

$$III_a = \left|\int_{\mathbb{R}} e\,U_0\,dy\right| \le |e(y,t)|_{L^\infty}(t)|U_0|_{L^1} \le C|U_0|_{L^1} \tag{4.21}$$

and

$$\begin{aligned} III_b = \left|\int_0^t\!\!\int_{\mathbb{R}} e_y\,\dot\delta U\,dy\,ds\right| &\le \int_0^t |e_y|_{L^2}(t-s)|\dot\delta||U|_{L^2}(s)\,ds \\ &\le C\zeta(t)^2\int_0^t (t-s)^{-\frac{1}{4}}(1+s)^{-\frac{3}{4}}\,ds \le C\zeta(t)^2. \end{aligned} \tag{4.22}$$

Summing (4.19)–(4.20) and (4.21)–(4.22), we obtain the desired bounds on $\dot\delta$ and δ.

This completes the proof of the claim, giving the result for $2 \le p \le \infty$. □

stability, $1 \leq \quad \leq 2$. The source term $\dot\delta\, U_x$ appearing in the reduced equations is convenient for high norm estimates L^p, $p \ge 2$, but (since it would lead to a source term involving higher derivative factor $|U_x|_{L^p}$ not controlled by energy estimates) not for low norm estimates L^p, $1 \le p < 2$. To treat low norms, we redefine

$$U := \tilde U(x,t) - \bar U(x - \delta(t))$$

following [40], which has the effect of replacing $\dot\delta\, U_x$ in the reduced equations with "centering errors"

$$\begin{aligned} S_1(\delta,\dot\delta,U)_x + \begin{pmatrix} 0 \\ I_r \end{pmatrix} S_2(\delta) := &-((A(\bar U(x-\delta)) - A(\bar U(x)))U + \dot\delta(\bar U(x-\delta) - \bar U(x)))_x \\ &+ \delta(Q(\bar U(x-\delta)) - Q(\bar U(x)))U \end{aligned}$$

satisfying

$$\begin{aligned} |S_1(\delta,\dot\delta,U)(y,s)| &\le |\delta|(|U| + |\dot\delta|)e^{-\theta|y|}, \\ |S_1(\delta,\dot\delta,U)_x(y,s)| &\le |\delta|(|U| + |\dot\delta| + |U_x|)e^{-\theta|y|}, \\ |S_2(\delta)(y,s)| &\le |\delta||U|e^{-\theta|y|}, \end{aligned}$$

and therefore

$$\begin{aligned} |S_1|_{L^1} &\le (|U|_{L^\infty} + |\dot\delta|) \le C(1+t)^{-\frac{1}{2}}|U_0|_{L^1\cap H^2}, \\ |(S_1)_x|_{L^1} &\le (|U|_{L^\infty} + |U|_{H^1} + |\dot\delta|) \le C(1+t)^{-\frac{1}{4}}|U_0|_{L^1\cap H^2}, \end{aligned}$$

and

$$|S_2|_{L^1} \le |U|_{L^\infty}|\delta| \le C(1+t)^{-\frac{1}{2}}|U_0|_{L^1\cap H^2}$$

by the previously obtained L^∞ and H^1 estimates, which are unaffected by a spatial shift. Likewise,

$$|N_1(U)|_{W^{1,1}},\ |N_2(U)|_{L^1} \le |U|^2_{H^1} \le C(1+t)^{-\frac{1}{2}}|U_0|_{L^1\cap H^2}.$$

Thus, expressing U by Duhamel's formula similarly to (4.9)–(4.11) and estimating nonlinear terms using the bounds of Proposition 4.1 with $p=1$, $q=1$, we readily obtain the sharp L^1 decay estimate

$$|U|_{L^1} \le C|U_0|_{L^1\cap H^2}$$

and, by interpolation with the previously obtained L^2 bound, the sharp L^p estimate, $1\le p\le 2$ of $|U|_{L^p} \le C(1+t)^{\frac{1}{2}(1-\frac{1}{p})}|U_0|_{L^1\cap H^2}$, as claimed. We omit the details, which are entirely similar to those already carried out.

5. The characteristic case. Finally, we briefly discuss the uniformly characteristic case in which (H1)(i) is violated for a shock profile of a discrete kinetic model. This cannot occur for the simplest examples of the Broadwell or Jin–Xin models, for the reason that it would violate the subcharacteristic condition

$$a_j < a_j^{*\pm} < a_{j+r}, \tag{5.1}$$

which is in turn necessary for strict dissipativity, (A2); see, e.g., [24, 26, 33]. Indeed, this holds in general for models with the property that $r+1$ consecutive characteristics take on only two values ($r+1=2$ for Broadwell; $r+1=(n/2)+1\ge 2$ for Jin–Xin, but the total number of characteristic values is 2), for the subcharacteristic condition (5.1) then implies that $a_p^{*\pm}$ lie between the neighboring characteristic values a_p and a_{p+r}, whence, by the Lax condition $a_p^{*-}>0>a_p^{*+}$, the speed $s=0$ does as well.

This is clearly an accident of low dimension, however, and for general models there is no physical reason that (H1)(i) should be satisfied. Indeed, though it is evidently satisfied generically, there is ample reason to discard this hypothesis, for, discretizing the Boltzmann equations

$$f_t + \xi f_x = \mathcal{Q}(\xi, f), \quad \xi \in \mathbb{R}^1,$$

by velocity ξ, where ξ denotes velocity, $f(\xi,x)$ the probability distribution of speeds at spatial location x, and $\mathcal{Q}(\xi,f)$ a collision term (local in x but nonlocal in ξ), we find as the velocity mesh goes to zero that (H1)(i) is more and more poorly satisfied, so that uniformity of our estimates (or even the ball for which small-amplitude profiles are guaranteed to exist) is lost.

This is hardly the main difficulty in proceeding to the Boltzmann limit, which is rather the reverse problem of unboundedness of the multiplication operator $f\to\xi f$ (in our notation, blowup of the spectrum of A) and the associated lack of spectral gap between zero and the essential spectrum of the operator $A^{-1}Q$ appearing in the standing-wave and eigenvalue ODE; see, e.g., [6]. Nonetheless, it is an issue that arises and should be addressed.

Fortunately, there is a simple fix, at least for discrete kinetic models. Namely, in case characteristics $a_j,\dots,a_k$ coincide with shock speed $s=0$, we may substitute for (H1)(i) the more general hypotheses

$$\mathrm{Re}\ \begin{pmatrix} L_j \cdots L_k \end{pmatrix}\ d\mathcal{Q} \begin{pmatrix} R_j \\ \vdots \\ R_k \end{pmatrix} \le -\theta < 0 \tag{5.2}$$

for some fixed left and right zero eigenbases L_i and R_i, and

$$\ker A \cap \ker d\mathcal{Q} = \emptyset, \tag{5.3}$$

where without loss of generality A is taken to be diagonal. This holds necessarily at $x = \pm\infty$ for some choice of diagonalizing transformation, by strict dissipativity, (A2) (recall that $\operatorname{Re}\sigma(\text{ block-diag}\,\tilde{L}Q\tilde{R}_\pm) < 0$ follows from (A2) by Taylor expansion at infinity of the spectrum of the symbol $i\xi A_\pm + Q_\pm$; likewise, (A2) implies genuine coupling, (1.3), of which (5.3) is a weakened form), hence (H1)(i) is always satisfied in the small-amplitude case. Whether or not it holds globally for physically interesting examples we do not know.

Review of the argument of section 3 shows that auxiliary energy estimate (1.9) goes through under this hypothesis with constant weights $\alpha_j = \alpha_{j+1} = \cdots = \alpha_k \equiv 1$ in the zero-speed modes, since there are no error terms in these modes to be overcome and there is a uniformly good contribution by (5.2). Likewise, review of the arguments of [24] shows that the results obtained there carry through as well, with appropriate modification of the proofs. Namely, (5.2) and (5.3) together imply that

$$\begin{pmatrix} df \\ dQ_j \\ \vdots \\ dQ_k \\ dg_2 \end{pmatrix}, \qquad g =: \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}, \quad g_1 \in \mathbb{R}^{k-j+1},$$

is full rank under some choice of coordinate system, whence we can again rewrite both traveling-wave and eigenvalue/resolvent equations as nondegenerate first-order systems and proceed as before. We omit the details, as this diverges from our main purpose.

Note in particular that we obtain small-amplitude existence for fixed speed s by a minor adjustment of the argument of [24] (namely, fixing the speed, without loss of generality, $s = 0$, and letting endstates vary), without any assumption on the base state other than simplicity of the principal eigenvalue and strong dissipativity; this generalizes earlier results of [34, 24].

For $a_j, \dots, a_k$ close to $s = 0$, we may recover our previous results with uniform estimates by a singular perturbation version of the same argument. It would be interesting to extend this approach to more general relaxation models for which the speeds a_i are not constant, in particular the case for which they are sometimes but not always characteristic.

Acknowledgments. K.Z. thanks Istituto per le Applicazioni del Calcolo "M. Picone" (CNR) and European TMR project "Hyperbolic Systems of Conservation Laws" for their hospitality and for making possible the visit (April 10–May 10, 2000) during which this work was initiated.

REFERENCES

[1] G. I. Barenblatt, J. Garcia-Azorero, A. De Pablo, and J. L. Vazquez, *Mathematical model of the non-equilibrium water-oil displacement in porous strata*, Appl. Anal., 65 (1997), pp. 19–45.

[2] G. I. Barenblatt and A. P. Vinnichenko, *Nonequilibrium filtration of nonmixing fluids*, Adv. in Mech., 3 (1980), pp. 35–50 (in Russian).

[3] A. A. Barmin and S. A. Egorushkin, *Stability of shock waves*, Adv. Mech., 15 (1992), pp. 3–37.
[4] S. Benzoni-Gavage, D. Serre, and K. Zumbrun, *Alternate Evans functions and viscous shock waves*, SIAM J. Math. Anal., 32 (2001), pp. 929–962.
[5] S. Bianchini, B. Hanouzet, and R. Natalini, *Asymptotic behavior of smooth solutions for weakly dissipative hyperbolic systems with a convex entropy in several space dimensions*, in preparation.
[6] R. Caflisch and B. Nickolaenko, *Shock profile solutions of the Boltzmann equation*, Comm. Math. Phys., 86 (1982), pp. 161–194.
[7] H. Freistühler and K. Zumbrun, *Examples of Unstable Viscous Shock Waves*, unpublished note, Institut für Mathematik, RWTH Aachen, 1998.
[8] K. O. Friedrichs, *Symmetric hyperbolic linear differential equations*, Comm. Pure Appl. Math., 7 (1954), pp. 345–392.
[9] K. O. Friedrichs, *On the laws of relativistic electro-magneto-fluid dynamics*, Comm. Pure Appl. Math., 27 (1974), pp. 749–808.
[10] R. Gardner and K. Zumbrun, *The gap lemma and geometric criteria for instability of viscous shock profiles*, Comm. Pure Appl. Math., 51 (1998), pp. 797–855.
[11] P. Godillon, *Linear stability of shock profiles for systems of conservation laws with semi-linear relaxation*, Phys. D, 148 (2001), pp. 289–316.
[12] J. Goodman, *Remarks on the stability of viscous shock waves*, in Viscous Profiles and Numerical Methods for Shock Waves, Raleigh, NC, 1990, SIAM, Philadelphia, PA, 1991, pp. 66–72.
[13] P. Howard and K. Zumbrun, *Pointwise estimates for dispersive-diffusive shock waves*, Arch. Ration. Mech. Anal., 155 (2000), pp. 85–169.
[14] P. Howard and K. Zumbrun, *Stability of Undercompressive Shock Profiles*, preprint, 2004.
[15] J. Humpherys, *Viscous-Dispersive Traveling Waves* I: *Admissibility and Existence*, preprint, 2003.
[16] S. Jin and Z. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., 48 (1995), pp. 235–276.
[17] S. Kawashima, *Systems of a Hyperbolic-Parabolic Composite Type, with Applications to the Equations of Magnetohydrodynamics*, thesis, Kyoto University, 1983.
[18] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, Heidelberg, 1985.
[19] P. D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CBMS-NSF Reg. Conf. Ser. in Appl. Math. 11, SIAM, Philadelphia, PA, 1973.
[20] T.-P. Liu, *Nonlinear stability of shock waves for viscous conservation laws*, Mem. Amer. Math. Soc., 56 (1985).
[21] A. Majda, *The stability of multi-dimensional shock fronts—a new problem for linear hyperbolic equations*, Mem. Amer. Math. Soc., 275 (1983).
[22] A. Majda, *The existence of multi-dimensional shock fronts*, Mem. Amer. Math. Soc., 281 (1983).
[23] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Springer-Verlag, New York, 1984.
[24] C. Mascia and K. Zumbrun, *Pointwise Green's function bounds and stability of relaxation shocks*, Indiana Univ. Math. J., 51 (2002), pp. 773–904.
[25] C. Mascia and K. Zumbrun, *Stability of small-amplitude shock profiles of symmetric hyperbolic–parabolic systems*, Comm. Pure Appl. Math., 57 (2004), pp. 841–876.
[26] C. Mascia and K. Zumbrun, *Pointwise Green function bounds for shock profiles of systems with real viscosity*, Arch. Ration. Mech. Anal., 169 (2003), pp. 177–263.
[27] C. Mascia and K. Zumbrun, *Stability of large-amplitude viscous shock profiles of hyperbolic–parabolic systems*, Arch. Ration. Mech. Anal., 172 (2004), pp. 93–131.
[28] R. Plaza and K. Zumbrun, *An Evans function approach to spectral stability of small-amplitude viscous shock profiles*, Discrete Contin. Dyn. Syst., 10 (2004), pp. 885–924.
[29] M.-R. Raoofi, *L^1 Asymptotic Behavior of Viscous Shock Profiles of Systems with Real Viscosity*, preprint, 2004.
[30] Y. Shizuta and S. Kawashima, *Systems of equations of hyperbolic–parabolic type with applications to the discrete Boltzmann equation*, Hokkaido Math. J., 14 (1985), pp. 249–275.
[31] J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Chapman and Hall, New York, 1989.
[32] G. B. Whitham, *Linear and Nonlinear Waves*, Pure Appl. Math., Wiley-Interscience, New York, London, Sidney, 1974.
[33] W.-A. Yong, *Basic Properties of Hyperbolic Relaxation Systems*, Progr. Nonlinear Differential Equations Appl., Birkhäuser Boston, Boston, 2001.

[34] W.-A. Yong and K. Zumbrun, *Existence of relaxation shock profiles for hyperbolic conservation laws*, SIAM J. Appl. Math., 60 (2000), pp. 1565–1575.

[35] Y. Zeng, *Gas dynamics in thermal nonequilibrium and general hyperbolic systems with relaxation*, Arch. Ration. Mech. Anal., 150 (1999), pp. 225–279.

[36] K. Zumbrun, *Stability of Viscous Shock Waves*, lecture notes, Indiana University, 1998.

[37] K. Zumbrun, *Refined wave-tracking and nonlinear stability of viscous Lax shocks*, Methods. Appl. Anal., 7 (2000), pp. 747–768.

[38] K. Zumbrun, *Multidimensional stability of planar viscous shock waves*, in Advances in the Theory of Shock Waves, Progr. Nonlinear Differential Equations Appl. 47, Birkhäuser Boston, Boston, MA, 2001, pp. 307–516.

[39] K. Zumbrun, *Planar Stability Conditions for Large-Amplitude Viscous Shock Profiles*, lecture notes, CIME, Cetraro, Italy, 2003.

[40] K. Zumbrun, *Stability of large-amplitude viscous shock profiles of the equations of fluid dynamics*, in Handbook of Fluids, Elsevier, to appear.

[41] K. Zumbrun and P.Howard, *Pointwise semigroup methods and stability of viscous shock waves*, Indiana Univ. Math. J., 47 (1998), pp. 741–871.

[42] K. Zumbrun and D. Serre, *Viscous and inviscid stability of multidimensional planar shock fronts*, Indiana Univ. Math. J., 48 (1999), pp. 937–992.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 914–940

ANALYSIS OF SOLUTIONS TO THE LAWRENCE–DONIACH SYSTEM FOR LAYERED SUPERCONDUCTORS*

PATRICIA BAUMAN† AND YANGSUK KO‡

Abstract. We consider a coupled Ginzburg–Landau system, the so-called Lawrence–Doniach system, which models layered superconductors as a stack of nonlinearly coupled, parallel two-dimensional superconducting layers, separated by an insulating material or vacuum, in an applied magnetic field. We prove that weak solutions (e.g., energy minimizers) in an appropriate divergence-free gauge are uniformly bounded and continuous and satisfy a priori estimates based on elliptic theory and single layer potentials. Moreover, we show the existence of an upper critical field $\bar{h}$ such that when the modulus of a constant applied magnetic field $\vec{H} = h\vec{v}$ in a direction $\vec{v}$ nontangential to the layers (where $|\vec{v}| = 1$) is greater than $\bar{h}$, the normal (nonsuperconducting) state is the only solution to the Lawrence–Doniach system. It follows from these results and methods developed earlier by Chapman, Du, and Gunzburger [*SIAM J. Appl. Math.,* 55 (1995), pp. 156–174] that under certain assumptions on the relative values of parameters in the model, minimizers of the Lawrence–Doniach energy converge, as the interlayer spacing tends to zero, to minimizers of an appropriate anisotropic Ginzburg–Landau energy in three dimensions. Finally, we derive that $\bar{h} \le C\kappa/\mu$ for all κ sufficiently large and all unit vectors $\vec{v}$ satisfying $\vec{v} \cdot \vec{e_3} \ge \mu > 0$ for the Lawrence–Doniach system, where κ is the Ginzburg–Landau constant for the superconducting material and C is independent of κ and μ.

Key words. superconductivity, Lawrence–Doniach equations, Ginzburg–Landau equations, anisotropic Ginzburg–Landau equations, layered superconductors, regularity, upper critical fields, homogenization

AMS subject classifications. 35J60, 35J65, 35Q40

DOI. 10.1137/S0036141004444597

1. Introduction and statement of main results.

1.1. Background on the Lawrence–Doniach model. The standard Ginzburg–Landau model (cf. [7]) has been well accepted as a macroscopic model for isotropic (and homogeneous) superconductors for temperatures near the critical temperature T_c of the material. However, this model cannot account for the anisotropy of layered or high temperature superconductors. (See [3], [18].) Therefore, alternative models have been developed. One of these is the Lawrence–Doniach model proposed by Lawrence and Doniach [15], whose solutions are analyzed here. Their model describes a layered superconductor as a finite number of infinitely thin parallel superconducting layers, each pair of which is separated by an insulating material or vacuum, occupying a bounded domain in $\mathbb{R}^3$. The model includes Josephson coupling in adjacent superconducting layers and is described by an energy which includes two-dimensional (isotropic) Ginzburg–Landau-type integrals in terms of an order parameter and magnetic potential on each layer, and a three-dimensional integral which includes the global effect of the applied magnetic field on the layered superconductor. The class of layered superconductors includes low-temperature layered superconductors and copper oxide high temperature superconductors (which are composed of layered Perovskite structures).

*Received by the editors July 17, 2004; accepted for publication (in revised form) February 22, 2005; published electronically December 6, 2005.
http://www.siam.org/journals/sima/37-3/44459.html
†Department of Mathematics, Purdue University, W. Lafayette, IN 47907-1395 (bauman@math.purdue.edu). This author was supported in part by NSF grant 9971974-DMS.
‡Department of Mathematics, California State University, Bakersfield, CA 93311 (yko@csub.edu).

The scaling parameters most important for characterizing superconducting materials are the coherence length ξ (representing the length scale for spatial variation of the superconducting order parameter $\psi(x)$, whose modulus is the density of superconducting electrons at position x in the reference configuration) and the penetration depth λ (which sets the length scale for electromagnetic response). In layered superconductors with uniaxial symmetry, these quantities are isotropic within the plane of the layers (which we assume is oriented so that it is parallel to the horizontal (x_1x_2) plane), and they take different values along the perpendicular (x_3 axis) direction (namely $\xi_\|, \lambda_\|, \xi_\perp, \lambda_\perp$). The degree of anisotropy, referred to in the physics literature as "quasi-two-dimensionality" varies over a wide range for different layered superconducting materials, depending on the strength of coupling between the adjacent two-dimensional superconducting layers. (A table of these values for various types of layered and high temperature superconducting materials can be found in [11].)

The Lawrence–Doniach model is widely accepted as a model for layered superconductors at temperatures close to the critical temperature T_C and is considered qualitatively correct even for lower temperatures. It is considered a more complete theory than the competing three-dimensional anisotropic Ginzburg–Landau model (described in section 6 of this paper). The latter model occurs as a limit of the Lawrence–Doniach model for layered superconductors in the case in which the coherence length perpendicular to the layers, $\xi_\perp$, is much larger than the layer separation, d, by letting $d/\xi_\perp \to 0$. This limiting process course-grains the layered structure into a continuum in which the anisotropy is incorporated into an effective mass tensor in the three-dimensional reference configuration of the material. The resulting three-dimensional anisotropic Ginzburg–Landau model is physically relevant for layered superconductors in which the coherence length perpendicular to the layers is much less than the interlayer spacing, and it is not expected to be even qualitatively correct at lower temperatures, where this coherence length can become less than or comparable to the layer separation. (See [13].) For layered superconductors in which $\xi_\perp$ is much less than d (as is the case in many high temperature superconductors), it is necessary to take the discrete nature of the layered structure into account. An example of a high temperature superconducting material in which this is the case is $BiSr_2Ca_2Cu_2O_{8+y}$. Thus, the Lawrence–Doniach model is more appropriate for describing $BiSr_2Ca_2Cu_2O_{8+y}$ than the three-dimensional anisotropic Ginzburg–Landau model. (See [11].)

A detailed study comparing the phenomenological Lawrence–Doniach model for layered superconductors to a full-fledged microscopic treatment was done by Klemm, Luther, and Beasley in [13]. An extensive description of experimental results on highly anisotropic layered high temperature superconductors described by the Lawrence–Doniach model can be found in [11]. The experimental observations and physical predictions in these papers motivate a careful study of solutions to the Lawrence–Doniach system (described below) in regimes in which the applied magnetic field $\vec{H} = h\vec{v}$ is allowed to take various directions. To understand why this is important, we note that in standard (isotropic) Ginzburg–Landau models in three dimensions, it is expected that vortices (i.e., superconducting defects) occur in filaments whose cross sections form an Abrikosov triangular lattice, and the qualitative behavior of solutions is similar as we vary the direction of the applied magnetic field. However, in layered superconductors of the type in which the discrete layers must be taken into account, the situation is different: experiments suggest that when the applied magnetic field $\vec{H} = h\vec{v}$ is perpendicular to the horizontal (x_1x_2) plane of the layers, the lateral distance between two vortices of a single flux line in adjacent layers increases on the order of $\cot\theta$, where $\theta = \arcsin(\vec{v}\cdot\vec{e_3})$. As long as θ is not too small, the vortices of single flux lines in adjacent layers have sufficient overlap to form a three-dimensional supercurrent flow pattern that is not very different from the usual Abrikosov vortex. As θ is reduced, there occurs a crossover from Abrikosov-like to staircase-like flux lines at a critical angle. As θ is further reduced, there may be a second critical angle below which the whole length of a flux line is locked into the region between the superconducting layers. Moreover, when θ is zero (so that the applied magnetic field is parallel to the layers), shielding current flows only within individual layers and not between them. (See [11].) The phenomena described above may be related to an interesting conjecture by Beasley, Klemm, and Luther, which we describe later in the introduction, asserting that when $\vec{H} = h\vec{v}$ is parallel to the $x_x x_2$ plane of the layers and the temperature T of the superconductor is sufficiently low, the upper critical modulus $H_{c_3}(v)$ for the Lawrence–Doniach model should be infinite. (See [13].) Physically, this means that no matter how strong the parallel applied magnetic field is, the superconducting material never becomes normal; i.e., it never reaches a state in which superconductivity is completely destroyed and the applied magnetic field has completely penetrated the superconducting layers. It is interesting to note that this kind of behavior is impossible for the three-dimensional anisotropic Ginzburg–Landau model, since methods developed in [9] for the standard isotropic three-dimensional Ginzburg–Landau model can be used to show easily that upper critical fields are uniformly bounded above independent of the direction $\vec{v}$, since the anisotropy governed by the effective mass tensor is not so different qualitatively from the isotropic case in which this tensor is the identity.

Before stating our results, let us introduce the Lawrence–Doniach model and the Euler–Lagrange equations and gauge invariance associated with the model. In the (nondimensionalized) Lawrence–Doniach model, a layered superconductor occupies a bounded domain $\mathcal{D} = \Omega \times (0, S)$ in $\mathbb{R}^3$, where Ω is a bounded Lipschitz domain in $\mathbb{R}^2$ and $S > 0$. The superconducting layers are perpendicular to the x_3 axis and occupy $N+1$ equally spaced planar regions, $\Omega_n = \Omega \times \{ns\}$, in $\mathcal{D}$ for $n = 0, 1, \ldots, N$, where $S = Ns$ and N is a positive integer. Thus, the distance s between adjacent superconducting planes is equal to $\frac{S}{N}$.

We assume throughout this paper that $\vec{H} = (h_1, h_2, h_3)$ is a given constant applied magnetic field, and we write $\vec{H} = h\vec{v}$, where $\vec{v}$ is a unit vector in $\mathbb{R}^3$ and $h \equiv |\vec{H}| \geq 0$. We say that $\vec{H}$ is nontangential if it is nontangential to the layers, i.e., $h > 0$ and $\vec{v}\cdot\vec{e_3} \neq 0$. For each unit vector $\vec{v}$ in $\mathbb{R}^3$, we assume that a fixed smooth divergence-free vector field $\vec{a} : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ has been chosen such that $\nabla \times \vec{a} = \vec{v}$ in $\mathbb{R}^3$. For example, if $\vec{v} = (v_1, v_2, v_3)$, one can choose $\vec{a(x)} = (v_2 x_3, v_3 x_1 - v_1 x_3, 0)$, and any other choice differs from this by the gradient of a harmonic function on $\mathbb{R}^3$. Thus, $\nabla \times h\vec{a} = h\vec{v} = \vec{H}$ in $\mathbb{R}^3$.

The Lawrence–Doniach energy functional for a layered superconductor in an applied magnetic field $\vec{H}$ in nondimensionalized form (see [4], [6], [13], and [15]) is given by

$$
\begin{aligned}
\mathcal{G}^s_{LD}\left(\{\psi_n\}_{n=0}^N, \vec{A}\right) = {} & s\sum_{n=0}^{N}\int_\Omega \left(\frac{1}{2}\left(|\psi_n|^2 - 1\right)^2 + \left|\left(\frac{\imath}{\kappa}\mathbf{grad} + \mathbf{A}_n\right)\psi_n\right|^2\right) d\mathbf{x} \\
& + s\sum_{n=0}^{N-1}\int_\Omega \sigma\left|\psi_{n+1}\exp\left(-\imath\kappa\int_{ns}^{(n+1)s} A^3 dx_3\right) - \psi_n\right|^2 d\mathbf{x} \\
& + \int_{\mathbb{R}^3}\left|\left(\nabla\times\vec{A}\right) - \vec{H}\right|^2 d\vec{x},
\end{aligned}
\tag{1.1}
$$

where $\mathbf{x} = (x_1, x_2)$ so that $\vec{x} = (x_1, x_2, x_3) = (\mathbf{x}, x_3)$, **grad** denotes the gradient operator with respect to the x_1 and x_2 coordinates, $\imath$ is the imaginary unit, and σ and κ are positive constants. Here, ψ_n is a complex-valued function defined on Ω, called the order parameter for the nth superconducting layer occupying $\Omega_n \equiv \Omega \times \{ns\}$; $\vec{A} = \vec{A}(\vec{x}) = (A^1, A^2, A^3) = (\mathbf{A}, A^3)$ is a vector field defined on $\mathbb{R}^3$, called the magnetic potential; and $\mathbf{A}_n = \mathbf{A}_n(\mathbf{x}) = (A_n^1(\mathbf{x}), A_n^2(\mathbf{x}))$ is a function from $\mathbb{R}^2$ to $\mathbb{R}^2$ defined by $\mathbf{A}_n(\mathbf{x}) = \mathbf{A}(\mathbf{x}, ns) = (A^1(x_1, x_2, ns), A^2(x_1, x_2, ns))$. More precisely, $\mathbf{A}_n(\mathbf{x})$ and $A_n^i(\mathbf{x})$ for $i = 1, 2$ are the traces of $\mathbf{A}$ and A^i, respectively, at $(\mathbf{x}, ns)$ on the plane $\{x_3 = ns\}$ for $n = 0, 1, \ldots, N$. Note that ψ_n is a function whose domain is Ω, even though it represents a physical quantity (the nth order parameter) corresponding to Ω_n. Similarly, $\mathbf{A}_n$ is a function whose domain is $\mathbb{R}^2$, though it corresponds to the trace of $\mathbf{A}$ on the plane $\{x_3 = ns\}$.

Given $\vec{H}$, the Lawrence–Doniach energy is defined on pairs $\big(\{\psi_n\}_{n=0}^N, \vec{A}\big)$ such that

$$
(1.2)\qquad \begin{cases} \{\psi_n\}_{n=0}^N \in \big[H^1(\Omega; \mathbb{C})\big]^{N+1} = \big[\mathcal{H}^1(\Omega)\big]^{N+1} \quad \text{and} \\ \vec{A} \in E \equiv \Big\{ \vec{C} \in H^1_{loc}(\mathbb{R}^3; \mathbb{R}^3) : (\nabla \times \vec{C}) - \vec{H} \in L^2(\mathbb{R}^3; \mathbb{R}^3) \Big\}. \end{cases}
$$

Note that for these pairs it follows from the trace theorem that $\mathbf{A}_n \in H^{\frac{1}{2}}_{loc}\big(\mathbb{R}^2; \mathbb{R}^2\big) \hookrightarrow L^4_{loc}\big(\mathbb{R}^2; \mathbb{R}^2\big)$, so that $\mathcal{G}^s_{LD}\big(\{\psi_n\}_{n=0}^N, \vec{A}\big)$ is well defined and finite. According to the Lawrence–Doniach model, a minimizer or stable equilibrium $\big(\{\psi_n\}_{n=0}^N, \vec{A}\big)$ of the Lawrence–Doniach energy, $\mathcal{G}^s_{LD}$, in $[\mathcal{H}^1(\Omega)]^{N+1} \times E$ corresponds to a physically realistic state of the layered superconductor: $\big|\psi_n(\mathbf{x})\big|^2$ is the density of superconducting electron pairs at position $(\mathbf{x}, ns)$ in the nth superconducting layer Ω_n; $\big(\nabla \times \vec{A}\big)(\vec{x})$ is the induced magnetic field at position $\vec{x}$ in $\mathbb{R}^3$; and $\big(\nabla \times (\nabla \times \vec{A})\big)(\vec{x})$ is the induced current at position $\vec{x}$ in $\mathbb{R}^3$. We note that in the above nondimensionalized model, several dimensional constants have been "nondimensionalized" or scaled from their original values. Thus, $\kappa = \lambda/\xi$, the ratio of the penetration depth and coherence length in the superconducting layers. Also, the nondimensionalized interlayer spacing, s, equals d/λ, where d is the original (dimensionalized) interlayer spacing. (See [4] for more details on the nondimensionalized formulation.) The constant σ is a parameter related to the strength of the interlayer coupling (Josephson coupling) of the layers in the superconducting material.

The Euler–Lagrange equations and natural boundary conditions associated with minimizers or equilibria of the Lawrence–Doniach energy $\mathcal{G}^s_{LD}$, called the Lawrence–Doniach system, are given by

$$
(1.3)\qquad \begin{cases} \text{(a)}\ \Big(\dfrac{\imath}{\kappa}\mathbf{grad} + \mathbf{A}_n\Big)^2 \psi_n + \big(|\psi_n|^2 - 1\big)\psi_n + \mathrm{P}_n = 0 & \text{on } \Omega, \\ \text{(b)}\ \nabla \times \big(\nabla \times \vec{A}\big) = (j_1, j_2, j_3) & \text{in } \mathbb{R}^3, \\ \text{(c)}\ \Big(\dfrac{\imath}{\kappa}\mathbf{grad}\ \psi_n + \mathbf{A}_n \psi_n\Big) \cdot \mathbf{n} = 0 & \text{on } \partial\Omega, \\ \text{(d)}\quad (\nabla \times \vec{A}) - \vec{H} \in L^2(\mathbb{R}^3; \mathbb{R}^3) & \end{cases}
$$

for all $n = 0, 1, \dots, N$, where

$$\mathrm{P}_n = \begin{cases} \sigma\big(\psi_0 - \psi_1 e^{-\imath\phi_0^1}\big) & \text{if } n = 0, \\ \sigma\big(2\psi_n - \psi_{n-1} e^{\imath\phi_{n-1}^n} - \psi_{n+1} e^{-\imath\phi_n^{n+1}}\big) & \text{if } 0 < n < N, \\ \sigma\big(\psi_N - \psi_{N-1} e^{\imath\phi_{N-1}^N}\big) & \text{if } n = N, \end{cases}$$

$$\phi_n^{n+1} = \kappa \int_{ns}^{(n+1)s} A^3\, dx_3 \quad \text{for } n = 0, 1, \dots, N-1,$$

$$j_i = s \sum_{n=0}^{N} \left[\frac{\imath}{2\kappa} \left(\psi_n \frac{\partial \psi_n^*}{\partial x_i} - c.c. \right) - |\psi_n|^2 A_n^i \right] \chi_\Omega(x_1, x_2) dx_1\, dx_2\, \delta_{ns}(x_3) \quad \text{for } i = 1, 2,$$

$$j_3 = \frac{1}{2} s\sigma\kappa\imath \sum_{n=0}^{N-1} \left[\psi_n \psi_{n+1}^* e^{\imath\phi_n^{n+1}} - c.c. \right] \chi_\Omega(x_1, x_2) \chi_{[ns,(n+1)s]}(x_3).$$

Here $\mathbf{n}$ is the outward normal to $\partial\Omega$, * means the complex conjugate, *c.c.* denotes the complex conjugate of the previous term, χ_Ω is the characteristic function for Ω, and $\delta_{ns} \in \mathcal{D}'(\mathbb{R})$ is the delta distribution supported at the point ns. The Lawrence–Doniach system (1.3) is to be interpreted in the weak sense defined by (2.3). Note that j_1 and j_2 are real measures in $\mathbb{R}^3$ supported in $\bigcup_{n=0}^N \overline{\Omega}_n$.

The Lawrence–Doniach energy, $\mathcal{G}_{LD}^s$, in (1.1) and the family of weak solutions of (1.2) and (1.3) are invariant in $[H^1(\Omega)]^{N+1} \times E$ under the gauge transformation

$$\Big(\{\psi_n\}_{n=0}^N, \vec{A}\Big) \longrightarrow \Big(\{\xi_n\}_{n=0}^N, \vec{Q}\Big),$$

where

$$\text{(1.4)} \qquad \begin{cases} \xi_n(\mathbf{x}) = \psi_n(\mathbf{x}) e^{\imath\kappa g(\mathbf{x}, ns)} & \text{in } \Omega, \\ \vec{Q} = \vec{A} + \nabla g & \text{in } R^3 \end{cases}$$

for all $n = 0, 1, \dots, N$ and any $g \in H^2_{loc}(\mathbb{R}^3)$. If $\big(\{\psi_n\}_{n=0}^N, \vec{A}\big) \in [H^1(\Omega)]^{N+1} \times E$ and (1.4) holds for a function $g \in H^2_{loc}(\mathbb{R}^3)$, we say that $\big(\{\psi_n\}_{n=0}^N, \vec{A}\big)$ and $\big(\{\xi_n\}_{n=0}^N, \vec{Q}\big)$ are gauge-equivalent. Note that the Lawrence–Doniach energy and the physical properties of a solution are invariant under this transformation: the density of superconducting electron pairs is $|\psi_n|^2 = |\xi_n|^2$, the induced magnetic field is $\nabla\times\vec{A} = \nabla\times\vec{Q}$, and the current is $\nabla\times\big(\nabla\times\vec{A}\big) = \nabla\times\big(\nabla\times\vec{Q}\big)$.

We prove in section 2 that every pair in $[H^1(\Omega)]^{N+1} \times E$ is gauge-equivalent to a pair $\big(\{\psi_n\}_{n=0}^N, \vec{A}\big)$, satisfying

$$\text{(1.5)} \qquad \begin{aligned} &\Big(\{\psi_n\}_{n=0}^N, \vec{A}\Big) \in [H^1(\Omega)]^{N+1} \times K, \\ &\text{where } K = \Big\{ \vec{C} \in E : \nabla\cdot\vec{C} = 0 \text{ and } \vec{C} - h\vec{a} \in \check{H}^1(\mathbb{R}^3) \cap L^6(R^3; R^3) \Big\}, \end{aligned}$$

which is unique up to uniform rotations: $\psi_n \to \psi_n e^{i\kappa c}$ for $n = 0, \dots, N$, where $c \in R$. (See Lemma 2.1.) Here $\check{H}^1(\mathbb{R}^3)$ denotes the completion of $C_0^\infty(\mathbb{R}^3; \mathbb{R}^3)$ with respect to the norm

$$\|\vec{C}\|_{\check{H}^1(\mathbb{R}^3)} = \left(\int_{\mathbb{R}^3} \big|\nabla\vec{C}\big|^2 d\vec{x} \right)^{\frac{1}{2}},$$

where

$$\nabla\vec{C} = \left(\frac{\partial C_j}{\partial x_i}\right) \text{ and } \left|\nabla\vec{C}\right|^2 = \operatorname{tr}\left[(\nabla\vec{C})^t(\nabla\vec{C})\right].$$

It follows easily that minimizers of $\mathcal{G}^s_{LD}$ (and hence weak solutions) of (1.3) in $[H^1(\Omega)]^{N+1} \times E$ exist. Moreover, for any weak solution of (1.3) in $[H^1(\Omega)]^{N+1} \times E$, g can be chosen so as to obtain a new gauge-equivalent weak solution of (1.3) in $[H^1(\Omega)]^{N+1} \times K$. (This process is called "choosing a gauge.")

The layered superconductor is said to be in a perfect superconducting state if $|\psi_n|^2 = 1$ in Ω for all n, and in the normal (nonsuperconducting) state if $|\psi_n|^2 = 0$ in Ω for all n and $\nabla \times \vec{A} \equiv \vec{H}$ in $\mathbb{R}^3$. In general, it is expected that $|\psi_n| \leq 1$ in Ω for each n; this was proved for a weak solution of (1.2) and (1.3) by Chapman, Du, and Gunzburger in [4] under the assumption that Ω is a smooth simply connected domain, and that the solution is gauge-equivalent to some $\left(\{\psi_n\}_{n=0}^N, \vec{A}\right)$ whose restriction to $\overline{\Omega}^{N+1} \times \overline{\mathcal{D}}$ is continuous, and satisfies $\left(\{\psi_n\}_{n=0}^N, \vec{A}\right) \in [H^1(\Omega)]^{N+1} \times \check{H}^1(\mathbb{R}^3)$,

$$\begin{aligned} &\nabla\cdot\vec{A} = 0 \text{ in } R^3, \\ &\vec{A}\cdot\vec{n} = 0 \text{ on } \partial D \text{ and } \nabla\psi\cdot\vec{n} = 0 \text{ on } \partial\Omega, \end{aligned} \tag{1.6}$$

where $\vec{n}$ is the outer unit normal on ∂D. (See [4, sections 1 and 4].) In addition, they proved that if γ and S are fixed positive numbers, with $\sigma\kappa^2 s^2 = \gamma^{-1}$ and $S = Ns$, and if minimizers of (1.1) satisfy $|\psi_n| \leq 1$ for all $n = 0, 1, \ldots, N$, then there is an extension of $\left(\{\psi_n\}_{n=0}^N, \vec{A}\right)$, denoted by $(\psi^s, \vec{A}^s)$, to $H^1(D) \times E$, such that $(\psi^s, \vec{A}^s)$ forms a minimizing sequence, as $s = s_j \equiv \frac{S}{N_j} \to 0$, for an appropriate anisotropic Ginzburg–Landau energy (which involves the parameters γ and κ) defined on $H^1(D) \times E$. (See (6.1) of section 6 for the definition of this anisotropic Ginzburg–Landau energy.) The proof of this result involved an intricate analysis in which the minimum Lawrence–Doniach energy was shown to converge (as $s \to 0$) to the minimum energy for the anisotropic Ginzburg–Landau model, assuming the above relationship between σ, κ, s, and γ. This result explains why, under certain assumptions on the relative sizes of the parameters σ, κ, and s, the Lawrence–Doniach system "homogenizes" to an anisotropic version of the classical three-dimensional Ginzburg–Landau model as the interlayer spacing s tends to 0.

We remark that the "gauge" (1.6) used in [4] is overdetermined. This is easily adjusted, however, by choosing the gauge (1.5), for example, instead of (1.6), and when this is done, it follows from the work of Chapman, Du, and Gunzburger in [4] that if weak solutions in the gauge (1.5) satisfy $|\psi_n| \leq 1$ in Ω for each n, the convergence result described above as $s \to 0$ holds in the gauge (1.5).

The question of whether minimizers of (1.1) or weak solutions to the Lawrence–Doniach system are gauge-equivalent to a pair whose restriction to $\overline{\Omega}^{N+1} \times \overline{\mathcal{D}}$ is continuous was left open in [4]. This question is nontrivial, since the magnetic potential $\vec{A} = (A^1, A^2, A^3)$ has the property that $\nabla \times \nabla \times \vec{A}$ is given by a measure supported in the layers, for $i = 1, 2$, and thus analysis of the regularity of the trace of A^i on Ω_n, and hence of ψ_n, is nontrivial. This question is of independent interest, since defects in superconductivity (known as vortices) in the nth layer are described by points $\mathbf{x}$ in Ω satisfying $\psi_n(\mathbf{x}) = 0$ (since $|\psi_n(\mathbf{x})|^2$ is the density of superconducting electron pairs in the nth layer at position $(\mathbf{x}, ns)$). This definition of defects makes sense only if higher regularity, such as continuity, of the order parameter ψ_n is proved. In addition, jump discontinuities in the induced magnetic field, $(\nabla \times \vec{A})(\vec{x})$, across the layers (corresponding to bending of the induced magnetic field across the layers) are expected for solutions of the Lawrence–Doniach system. (See [1] and [2].) A rigorous analysis of such properties cannot be done without results on the regularity of $(\nabla \times \vec{A})(\vec{x})$ that identifies how and where jumps will occur across each layer for solutions of the Lawrence–Doniach system.

1.2. Main results and outline of the paper. In this paper, assuming that Ω is a bounded Lipschitz domain in $\mathbb{R}^2$, we prove that $|\psi_n| \leq 1$ almost everywhere in Ω for $n = 0, 1, \ldots, N$ for all weak solutions, without assuming continuity of solutions of the Lawrence–Doniach system. In addition, we prove a priori estimates, which imply that all weak solutions (and in particular, minimizers of the Lawrence–Doniach energy) are gauge-equivalent to a pair satisfying (1.5), whose restriction to $\overline{\Omega}^{N+1} \times \overline{\mathcal{D}}$ is continuous, with $\vec{A}$ piecewise C^1 in $\Omega \times \mathbb{R}$, such that all discontinuities of $\nabla\vec{A}$ in $\Omega \times \mathbb{R}$ occur as jump discontinuities in $\frac{\partial A^i}{\partial x_3}$ for $i = 1, 2$ from above and below at points in Ω_n. Our methods of proof involve elliptic regularity and a priori estimates of single layer potentials. The latter arise because A^1 and A^2 in this gauge have Laplacians given by measures supported in the layers, and hence can be represented as single layer potentials in $\mathbb{R}^3$. A consequence of these results and the methods of Chapman, Du, and Gunzburger is that minimizers of the Lawrence–Doniach energy converge, as the nondimensionalized interlayer spacing s tends to zero, to minimizers of an appropriate anisotropic Ginzburg–Landau energy in three dimensions (provided that $\sigma\kappa^2 s^2 = \gamma^{-1}$ for some $\gamma > 0$). (See section 6.) Another consequence is the existence of a smallest nonnegative number, $\bar{h} = \bar{h}(\vec{v}, \kappa, s, \sigma, S, \Omega) = \frac{1}{\mu}O(\kappa)$, called the upper critical modulus, such that if $h > \bar{h}$ and $\vec{H} = h\vec{v}$, where $\vec{v}$ is a nontangential unit vector in $\mathbb{R}^3$ and $0 < \mu \leq |\vec{v} \cdot \vec{e_3}|$, the only weak solution to the Lawrence–Doniach system is the normal (nonsuperconducting) state, defined by $\psi_n \equiv 0$ for $n = 0, 1, \ldots, N$ and $\nabla \times \vec{A} \equiv \vec{H}$ in $\mathbb{R}^3$. (See (5.1) for a formal definition of $\bar{h}$.)

The infimum of the values $h'' > 0$ such that all normal states are stable for $h > h''$ in an applied magnetic field $\vec{H} = h\vec{v}$ is denoted by $H_{c_3} = H_{c_3}(\vec{v}, \kappa, s, \sigma, S, \Omega)$. (If there are no such values h'', one defines $H_{c_3} = \infty$.) Here, a weak solution is called stable if the Lawrence–Doniach energy has nonnegative second variation at the solution. From our result on the finiteness of $\bar{h} < \infty$ and the fact that the Lawrence–Doniach energy has a minimizer in $[H^1(\Omega)]^{N+1} \times E$, it follows that $H_{c_3} \leq \bar{h} = \frac{1}{\mu}O(\kappa)$ (for each fixed $\vec{v}, \kappa, s, \sigma, S$, and Ω as above), for nontangential applied magnetic fields $\vec{H} = h\vec{v}$. This proves a result conjectured by physicists (see [13] and the references cited there). Note that since μ is the sine of the angle between $\vec{H} = h\vec{v}$ and the x_1x_2 plane when $\vec{v} \cdot \vec{e_3} = \mu$, our upper bound on H_{c_3} blows up as $\mu \to 0$. Moreover, an examination of the proof of this estimate (based on the inequality (5.6) of Lemma 5.4, which involves integral estimates on the two-dimensional layers) shows that our proof of this result does not allow a finite upper bound on H_{c_3} as $\vec{v}$ approaches a direction parallel to the layers.

We remark that Klemm, Luther, and Beasley have predicted that $H_{c_3} = \infty$ at a sufficiently low temperature $T < T_c$ for applied magnetic fields $\vec{H} = h\vec{v}$ in directions $\vec{v}$ that are parallel to the layers, based on a theory that the normal cores in these materials can effectively fit between the layers (so that $|\psi_n| > 0$ in Ω_n for each n) in this case. The behavior of solutions in this regime has been studied recently by Alama, Berlinsky, and Bronsard in [1] for layered superconductors occupying a finite number of parallel infinite strips, under some assumptions on the spatial dependence of the order parameter in each layer and on the magnetic potential $\vec{A}$ in $\mathbb{R}^3$. Since the constants in the Lawrence–Doniach model are temperature dependent, an interesting open problem (that will require a different approach from the one we develop here in Lemma 5.4 and Theorem 5.5) is to compute $H_{c_3} = H_{c_3}(T, \vec{v}, \kappa, s, \sigma, S, \Omega)$ when $\vec{v}$ is parallel to the x_1x_2 plane. We conjecture that $H_{c_3}(v)$ is finite in certain regimes depending on the constants in the Lawrence–Doniach model in parallel fields $\vec{v}$ (since it converges to the three-dimensional Ginzburg–Landau anisotropic model under certain assumptions on these constants), but it may be infinite for parallel fields at certain temperatures (i.e., in other regimes depending on the constants in the Lawrence–Doniach model).

As far as we know, our results on the existence and bounds of the upper critical moduli H_{c_3} and $\bar{h}$ for nontangential applied magnetic fields, $\vec{H} = h\vec{v}$, are the first rigorous results on upper critical moduli for the Lawrence–Doniach system, even in the case of a perpendicular applied magnetic field, $\vec{H} = h\vec{e_3}$.

A detailed summary of our results and the organization of our paper is as follows. In section 2 we prove the maximum principle, $|\psi_n|^2 \leq 1$ on Ω, for all solutions of (1.2) and (1.3), without assuming any continuity of solutions. (See Theorem 2.4.) In addition, we observe that for a solution $\big(\{\psi_n\}_{n=0}^N, \vec{A}\big)$ in the "Coulomb gauge" (1.5), it follows from elliptic regularity theory (since $\nabla\times\big(\nabla\times\vec{A}\big) = -\Delta\vec{A}$ for divergence-free fields in $\mathbb{R}^3$) that A^3 is in $W^{2,q}_{loc}(\mathbb{R}^3)$ for all $q \in (1,\infty)$, and thus A^3 is in $C^{1,\beta}_{loc}(R^3)$ for all $\beta \in (0,1)$. (See Theorem 2.6.)

Unlike the case of A^3, the Laplacians of A^1 and A^2 are singular measures rather than L^p functions. In addition, ψ_n satisfies a nonlinear elliptic equation in Ω involving $\mathbf{A}_n$ for each n, where $\mathbf{A}_n(\cdot)$ is the trace of (A^1, A^2) on $\{x_3 = ns\}$. Therefore, techniques for obtaining higher regularity for the standard Ginzburg–Landau system do not apply. In section 3 we find explicit integral representations of A^1, A^2, A^1_n, and A^2_n involving single layer potentials, and using results from the theory of single layer potentials, we improve the regularity of A^1_n and A^2_n from $H^{\frac{1}{2}}_{loc}(\mathbb{R}^2)$ to $H^1_{loc}(\mathbb{R}^2)$.

In section 4, we use the above result and a bootstrapping method (combining results from the theory of single layer potentials and elliptic theory) to obtain a regularity result, Theorem 4.6, which states that there exist constants $p > 2$ and α in (0,1) such that in the gauge (1.5), all weak solutions of the Lawrence–Doniach system satisfy

$$
\begin{aligned}
&\psi_n \in W^{1,p}\big(\Omega\big) \cap C^{\alpha}\big(\overline{\Omega}\big) \cap C^{2,\alpha}\big(\Omega\big),\\
&A^1_n, A^2_n \in W^{1,p}_{loc}\big(\mathbb{R}^2\big) \cap C^{\alpha}_{loc}\big(\mathbb{R}^2\big) \cap C^1\big(\Omega\big) \cap C^{\infty}\Big(\mathbb{R}^2 \setminus \overline{\Omega}\Big) \text{ for all } n = 0, 1, \ldots, N,\\
&A^1, A^2 \in W^{1,p}_{loc}\big(\mathbb{R}^3\big) \cap C^{\alpha}_{loc}\big(\mathbb{R}^3\big) \cap C^{\infty}\left(\mathbb{R}^3 \Big\backslash \bigcup_{i=o}^{N} \overline{\Omega}_n\right),\\
&\frac{\partial A^i}{\partial x_1} \text{ and } \frac{\partial A^i}{\partial x_2} \in C\big(\Omega \times \mathbb{R}\big) \text{ for } i = 1, 2,\\
&A^3 \in W^{2,q}_{loc}\big(\mathbb{R}^3\big) \cap C^{1,\beta}_{loc}\big(\mathbb{R}^3\big) \text{ for all } q \in (1,\infty) \text{ and } \beta \in (0,1).
\end{aligned}
$$

In addition, for $i = 1, 2$ we prove that A^i is piecewise C^1 in $\Omega \times \mathbb{R}$, with all discontinuities of ∇A^i in $\Omega \times \mathbb{R}$ occurring as jump discontinuities in $\frac{\partial A^i}{\partial x_3}$ from above and below at points in Ω_n where the ith component of the current density in $R^2 \times \{ns\}$, namely, $g^i_n(\mathbf{x}, ns)$ (defined by (3.4)), is nonzero. If Ω is a $C^{1,1}$ domain, we prove that the above results hold for all $p \in (1,\infty)$ and all $\alpha \in (0,1)$, with additional regularity up to the boundary of Ω. (See Theorems 4.4 and 4.6.) Thus ψ_n is continuous in Ω for each n, and the induced magnetic field $\nabla \times \vec{A}$ is continuous in $\Omega \times \mathbb{R}^3$ except at points in Ω_n, where it has jump discontinuities.

In section 5, we prove the existence of $\bar{h} = \bar{h}(\vec{v}, \kappa, s, \sigma, S, \Omega) < \infty$ such that if $h > \bar{h}$, $\vec{v}$ is a nontangential unit vector, and $\vec{H} = h\vec{v}$, any weak solution to the Lawrence–Doniach system is normal (nonsuperconducting). We also prove that $H_{c_3} \leq \bar{h} \leq C\kappa/\mu$ for all κ sufficiently large if $|\vec{v} \cdot \vec{e_3}| \geq \mu > 0$, where C is a positive constant depending only on S and Ω. (See Theorem 5.5 and Corollary 5.6.) The above result generalizes a theorem of Giorgi and Phillips in [9] for the standard Ginzburg–Landau model with $\vec{v} = \vec{e_3}$ as $\kappa \to \infty$ in two and three dimensions.

Finally, in section 6, we describe the three-dimensional anisotropic Ginzburg–Landau model, and conclude as a consequence of Theorem 2.4 and the results of Chapman, Du, and Gunzburger in [4], that if $\sigma\kappa^2 s^2 = \gamma^{-1}$, where $\gamma > 0$ is fixed, then minimizers of the Lawrence–Doniach energy in the gauge (1.5) have an extension to $H^1(D) \times E$ in a divergence-free gauge in $\mathbb{R}^3$ and form a minimizing sequence as $s \to 0$ of the anisotropic Ginzburg–Landau energy. (See Theorem 6.1.)

2. Preliminary results. Throughout this paper, we assume that Ω is a given bounded Lipschitz domain in $\mathbb{R}^2$ and $\vec{H} = h\vec{v}$ is a constant applied magnetic field on $\mathbb{R}^3$ with h, $\vec{v}$, and $\vec{a} = \vec{a}(\vec{x})$ as described in the introduction.

In this section we prove preliminary results on weak solutions of (1.3) and minimizers of the Lawrence–Doniach energy in the gauge $[\mathcal{H}^1(\Omega)]^{N+1} \times K$ defined by (1.5). Recall that $\check{H}^1(\mathbb{R}^3)$ is the completion of $C_0^\infty(\mathbb{R}^3; \mathbb{R}^3)$ with respect to the norm

$$\|\vec{C}\|_{\check{H}^1(\mathbb{R}^3)} = \left(\int_{\mathbb{R}^3} \left|\nabla \vec{C}\right|^2 d\vec{x} \right)^{\frac{1}{2}}.$$

It follows from the Sobolev inequality that each $\vec{C} \in \check{H}^1(\mathbb{R}^3)$ has a representative in $L^6(R^3; R^3)$ such that

$$\|\vec{C}\|_{L^6(\mathbb{R}^3, \mathbb{R}^3)} \leq 2\|\vec{C}\|_{\check{H}^1(\mathbb{R}^3)}. \tag{2.1}$$

Moreover,

$$\|\vec{C}\|^2_{\check{H}^1(\mathbb{R}^3)} = \int_{\mathbb{R}^3} \left(\left|\nabla \cdot \vec{C}\right|^2 + \left|\nabla \times \vec{C}\right|^2 \right) d\vec{x}. \tag{2.2}$$

(See [14].) We need the following lemma to prove that the "Coulomb gauge" (1.5) is an appropriate gauge for weak solutions of (1.3) in $[\mathcal{H}^1(\Omega)]^{N+1} \times E$.

LEMMA 2.1. *Let* $\left(\{\xi_n\}_{n=0}^N, \vec{Q}\right) \in [\mathcal{H}^1(\Omega)]^{N+1} \times E$. *Then there exists a gauge-equivalent pair* $\left(\{\psi_n\}_{n=0}^N, \vec{A}\right)$ *satisfying* (1.5). *Moreover, if* $\left(\{\tilde{\psi}_n\}_{n=0}^N, \tilde{A}\right)$ *is another such pair, then* $\tilde{A} = \vec{A}$ *and there is a constant* $c \in \mathbb{R}$ *such that* $\tilde{\psi}_n = \psi_n e^{\imath\kappa c}$ *for all* $n = 0, 1, \ldots, N$.

Proof. Let $\left(\{\xi_n\}_{n=0}^N, \vec{Q}\right) \in [\mathcal{H}^1(\Omega)]^{N+1} \times E$ be given. Then $\vec{Q} \in H^1_{loc}(\mathbb{R}^3; \mathbb{R}^3)$ and $\nabla \times \vec{Q} - \vec{H} = \nabla \times \left(\vec{Q} - h\vec{a}\right) \in L^2(\mathbb{R}^3; \mathbb{R}^3)$. Set $\vec{u} = \nabla \times \left(\vec{Q} - h\vec{a}\right)$. Since $\vec{u} \in L^2(\mathbb{R}^3; \mathbb{R}^3)$ and $\nabla \cdot \vec{u} = 0$ in $\mathcal{D}'(\mathbb{R}^3)$, Lemma 3.1 of [9] states that there is a unique $\vec{C} \in \check{H}^1(\mathbb{R}^3) \cap L^6(R^3; R^3)$ such that $\nabla \times \vec{C} = \vec{u}$ and $\nabla \cdot \vec{C} = 0$. Set $\vec{A} = \vec{C} + h\vec{a}$. Then $\vec{A} \in H^1_{loc}(\mathbb{R}^3; \mathbb{R}^3)$, $\vec{A} - h\vec{a} = \vec{C} \in \check{H}^1(\mathbb{R}^3) \cap L^6(R^3; R^3)$, and $\nabla \cdot \vec{A} = \nabla \cdot \vec{C} + h\nabla \cdot \vec{a} = 0$

in $\mathbb{R}^3$. Since

$$\begin{aligned}\nabla\times\left(\vec{A}-\vec{Q}\right) &= \nabla\times\left(\vec{A}-h\vec{a}\right)-\nabla\times\left(\vec{Q}-h\vec{a}\right)\\ &= \nabla\times\vec{C}-\vec{u}=\vec{0}\end{aligned}$$

and $\vec{A}-\vec{Q}\in H^1_{loc}(\mathbb{R}^3;\mathbb{R}^3)$, there exists a function g in $H^2_{loc}(\mathbb{R}^3)$ such that $\vec{A}=\vec{Q}+\nabla g$. Set $\psi_n(\mathbf{x})=\xi_n(\mathbf{x})e^{\imath\kappa g(\mathbf{x},ns)}$ for all $\mathbf{x}$ in Ω and for all $n=0,1,\dots,N$. Then the pair $\left(\{\psi_n\}_{n=0}^N,\vec{A}\right)$ is gauge-equivalent to $\left(\{\xi_n\}_{n=0}^N,\vec{Q}\right)$ and $\left(\{\psi_n\}_{n=0}^N,\vec{A}\right)\in[\mathcal{H}^1(\Omega)]^{N+1}\times K$.

Now assume that there is another pair $(\{\tilde{\psi}_n\}_{n=0}^N,\tilde{A})$ in $[\mathcal{H}^1(\Omega)]^{N+1}\times K$, which is gauge-equivalent to $\left(\{\xi_n\}_{n=0}^N,\vec{Q}\right)$; i.e., there is a function $\tilde{g}$ in $H^2_{loc}(\mathbb{R}^3)$ such that $\tilde{A}=\vec{Q}+\nabla\tilde{g}$ and $\tilde{\psi}_n(\mathbf{x})=\xi_n(\mathbf{x})e^{\imath\kappa\tilde{g}(\mathbf{x},ns)}$ for each n. To show that $\vec{A}=\tilde{A}$ and that there is a constant $c\in\mathbb{R}$ such that $\tilde{\psi}_n=\psi_n e^{\imath\kappa c}$ for each n, it is sufficient to prove that $\tilde{g}-g=c$ for some constant $c\in\mathbb{R}$. Set $\tilde{C}=\tilde{A}-h\vec{a}$. Then $\vec{C},\ \tilde{C}\in\check{H}^1(\mathbb{R}^3)\cap L^6(R^3;R^3)$, $\nabla\times\vec{C}=\nabla\times\tilde{C}=\nabla\times\left(\vec{Q}-h\vec{a}\right)\in L^2(\mathbb{R}^3;\mathbb{R}^3)$, and $\nabla\cdot\vec{C}=\nabla\cdot\tilde{C}=0$. By Lemma 3.1 in [9], such a vector field is unique, and hence $\vec{C}=\tilde{C}$. Thus, $\vec{A}=\tilde{A}$ and we have $\nabla g=\nabla\tilde{g}$, so that $\tilde{g}=g+c$ for some constant $c\in\mathbb{R}$. ∎

By Lemma 2.1 and the invariance of the Lawrence–Doniach energy under gauge transformations, the existence of minimizers in $[\mathcal{H}^1(\Omega)]^{N+1}\times E$ is equivalent to the existence of minimizers in $[\mathcal{H}^1(\Omega)]^{N+1}\times K$. Thus we have the following theorem.

THEOREM 2.2. *There exists* $\left(\{\psi_n\}_{n=0}^N,\vec{A}\right)\in[\mathcal{H}^1(\Omega)]^{N+1}\times K$, *which minimizes* $\mathcal{G}^s_{LD}$ *in* $[\mathcal{H}^1(\Omega)]^{N+1}\times E$.

Proof. Let $\{\left(\{\psi^j_n\}_{n=0}^N,\vec{A}^j\right)\}\subset[\mathcal{H}^1(\Omega)]^{N+1}\times E$ be a minimizing sequence for $\mathcal{G}^s_{LD}$. By Lemma 2.1, we may assume that $\vec{A}^j\in K$ for all j without loss of generality. Using (2.1), (2.2), and $\nabla\cdot\vec{A}^j=0$ for all j, it follows that $\left(\{\psi^j_n\}_{n=0}^N,\vec{A}^j-h\vec{a}\right)$ is bounded in $[\mathcal{H}^1(\Omega)]^{N+1}\times\ [\check{H}^1(\mathbb{R}^3)\cap L^6(R^3;R^3)]$ so that there is a weakly convergent subsequence. Let $(\{\psi_n\}_{n=0}^N,\vec{A}-h\vec{a})$ be the weak limit of the subsequence. Since the integrands in (1.1) are each weakly lower semicontinuous in $[\mathcal{H}^1(\Omega)]^{N+1}\times K$, $\mathcal{G}^s_{LD}$ is weakly lower semicontinuous, and thus $\left(\{\psi_n\}_{n=0}^N,\vec{A}\right)$ is a minimizer in the gauge $[\mathcal{H}^1(\Omega)]^{N+1}\times K$. ∎

The weak formulation of (1.3) is

(2.3)

$$\begin{aligned}
&\text{(a)}\quad s\sum_{n=0}^{N}\int_\Omega\left[\left(|\psi_n|^2-1\right)\psi_n\varphi_n^*+\left(\frac{\imath}{\kappa}\mathbf{grad}\ \psi_n+\mathbf{A}_n\psi_n\right)\cdot\left(\frac{\imath}{\kappa}\mathbf{grad}\ \varphi_n+\mathbf{A}_n\varphi_n\right)^*\right]d\mathbf{x}\\
&\qquad+s\sigma\sum_{n=0}^{N-1}\int_\Omega\left[\left(\psi_{n+1}-\psi_n e^{\imath\phi_n^{n+1}}\right)\varphi_{n+1}^*+\left(\psi_n-\psi_{n+1}e^{-\imath\phi_n^{n+1}}\right)\varphi_n^*\right]d\mathbf{x}=0,\\
&\text{(b)}\quad 2\int_{\mathbb{R}^3}\nabla\times\left(\vec{A}-h\vec{a}\right)\cdot\nabla\times\vec{B}\,d\vec{x}+s\sigma\int_{\mathbb{R}^3}\mathrm{I}\cdot B^3\,d\vec{x}\\
&\qquad+s\sum_{n=0}^{N}\int_{\mathbb{R}^2}\chi_{[\Omega\times\{ns\}]}(\mathbf{x},x_3)\left[\frac{\imath}{\kappa}\left(\psi_n^*\,\mathbf{grad}\ \psi_n-c.c.\right)+2\mathbf{A}_n\left|\psi_n\right|^2\right]\cdot\mathbf{B}_n\,d\mathbf{x}=0
\end{aligned}$$

for any $\{\varphi_n\}_{n=0}^N\in[\mathcal{H}^1(\Omega)]^{N+1}$ and $\vec{B}=(B^1,B^2,B^3)\in H^1(\mathbb{R}^3;\mathbb{R}^3)$ with bounded support. Here,

$$
\begin{aligned}
\mathrm{I} &= \imath\kappa \sum_{n=0}^{N-1} \chi_\Omega(\mathbf{x})\chi_{[ns,(n+1)s]}(x_3)\Big(\psi_{n+1}\psi_n^* e^{-\imath\phi_n^{n+1}} - c.c.\Big) \\
&= 2\kappa \sum_{n=0}^{N-1} \chi_\Omega(\mathbf{x})\chi_{[ns,(n+1)s]}(x_3)\Im\Big(\psi_{n+1}^*\psi_n e^{\imath\phi_n^{n+1}}\Big),
\end{aligned} \tag{2.4}
$$

where $\Im$ denotes the imaginary part of the argument and *c.c.* denotes the complex conjugate of the previous term. By a weak solution of (1.3) we mean a pair, $\big(\{\psi_n\}_{n=0}^N, \vec{A}\big) \in [\mathcal{H}^1(\Omega)]^{N+1} \times E$, satisfying (2.3). From Theorem 2.2, we have the following.

COROLLARY 2.3. *There exists* $\big(\{\psi_n\}_{n=0}^N, \vec{A}\big) \in [\mathcal{H}^1(\Omega)]^{N+1} \times K$, *which is a weak solution to the Lawrence–Doniach system* (1.3).

Next we prove that the magnitude of the order parameters of weak solutions to (1.3) is uniformly bounded by 1, which is the density of the perfectly superconducting state. This result generalizes Proposition 4.5 in [4], in which the maximum principle was established under the assumption that ψ_n and $\vec{A}$ are continuous on $\overline{\Omega}$ and $\overline{\mathcal{D}}$ for all $n = 0, 1, \ldots, N$.

THEOREM 2.4. *If* $\big(\{\psi_n\}_{n=0}^N, \vec{A}\big)$ *is a weak solution to* (1.3), *then* $|\psi_n| \le 1$ *almost everywhere in* Ω *for all* $n = 0, 1, \ldots, N$.

Proof. Set $\varphi_n = \big(|\psi_n| - 1\big)_+ d_n$, where $d_n = \frac{\psi_n}{|\psi_n|}$ and where $q_+ = q$ if $q \ge 0$ and $q_+ = 0$ if $q < 0$. Since $\psi_n \in H^1(\Omega;\mathbb{C})$, we have $|\psi_n|$ and $(|\psi_n| - 1)_+$ in $H^1(\Omega)$ (see [10, Chapter 7]), and it follows that $\varphi_n \in H^1(\Omega, \mathbb{C})$ for each n. Let $O_n = \{x \in \Omega : |\psi_n(x)| > 1\}$. Then almost everywhere in O_n, we have

$$
\frac{\imath}{\kappa}\,\mathbf{grad}\,\psi_n + \mathbf{A}_n\psi_n = \frac{\imath}{\kappa} d_n\Big(\mathbf{grad}\ |\psi_n|\Big) + |\psi_n|\left(\frac{\imath}{\kappa}\mathbf{grad}\, d_n + \mathbf{A}_n d_n\right),
$$

$$
\left(\frac{\imath}{\kappa}\,\mathbf{grad}\,\varphi_n + \mathbf{A}_n\varphi_n\right)^* = -\frac{\imath}{\kappa} d_n^*\Big(\mathbf{grad}\ |\psi_n|\Big) + \big(|\psi_n| - 1\big)\left(-\frac{\imath}{\kappa}\mathbf{grad}\, d_n^* + \mathbf{A}_n d_n^*\right),
$$

and

$$
\big(|\psi_n|^2 - 1\big)\psi_n\varphi_n^* = |\psi_n|\,\big(|\psi_n| + 1\big)\big(|\psi_n| - 1\big)^2
$$

so that

$$
\begin{aligned}
&\Re\left\{\left(\frac{\imath}{\kappa}\mathbf{grad}\,\psi_n + \mathbf{A}_n\psi_n\right)\left(\frac{\imath}{\kappa}\mathbf{grad}\,\varphi_n + \mathbf{A}_n\varphi_n\right)^*\right\} \\
&\quad = \frac{1}{\kappa^2}\big|\mathbf{grad}\ |\psi_n|\big|^2 + |\psi_n|\,\big(|\psi_n| - 1\big)\left|\frac{\imath}{\kappa}\mathbf{grad}\, d_n + \mathbf{A}_n d_n\right|^2.
\end{aligned}
$$

Taking the real part of (2.3a), we obtain

$$
\begin{aligned}
& s\sum_{n=0}^{N}\int_{O_n}\left(\frac{1}{\kappa^2}\big|\mathbf{grad}\ |\psi_n|\big|^2 + |\psi_n|\,\big(|\psi_n| - 1\big)\left|\frac{\imath}{\kappa}\mathbf{grad}\, d_n + \mathbf{A}_n d_n\right|^2\right)d\mathbf{x} \\
&+ s\sum_{n=0}^{N}\int_{O_n}|\psi_n|\,\big(|\psi_n| + 1\big)\big(|\psi_n| - 1\big)^2\,d\mathbf{x} \\
&+ s\sigma\sum_{n=0}^{N-1}\Re\left\{\int_\Omega\left(\left(\psi_n - \psi_{n+1}e^{-i\phi_n^{n+1}}\right)\varphi_n^* + \left(\psi_{n+1} - \psi_n e^{i\phi_n^{n+1}}\right)\varphi_{n+1}^*\right)d\mathbf{x}\right\} = 0.
\end{aligned}
$$

We write this as $\mathrm{I}+\mathrm{II}+\mathrm{III}=0$. Note that $\mathrm{I}\geq 0$ and $\mathrm{II}\geq 0$. Rewriting III, we have

$$\mathrm{III} = s\sigma\sum_{n=0}^{N-1}\left[\int_\Omega \Re\left\{\left(\psi_n-\psi_{n+1}e^{-i\phi_n^{n+1}}\right)\psi_n^*\frac{|\psi_n|-1}{|\psi_n|}\chi_{O_n}\right\}\right. \\ \left.+\Re\left\{\left(\psi_{n+1}-\psi_n e^{i\phi_n^{n+1}}\right)\psi_{n+1}^*\frac{|\psi_{n+1}|-1}{|\psi_{n+1}|}\chi_{O_{n+1}}\right\}d\mathbf{x}\right]$$

so that

$$\mathrm{III} = s\sigma\sum_{n=0}^{N-1}\left[\int_{O_n}\left(|\psi_n|^2-\Re\left\{\psi_n^*\psi_{n+1}e^{-i\phi_n^{n+1}}\right\}\right)\frac{|\psi_n|-1}{|\psi_n|}\,d\mathbf{x}\right. \\ \left.+\int_{O_{n+1}}\left(|\psi_{n+1}|^2-\Re\left\{\psi_{n+1}^*\psi_n e^{i\phi_n^{n+1}}\right\}\right)\frac{|\psi_{n+1}|-1}{|\psi_{n+1}|}\,d\mathbf{x}\right].$$

Thus,

$$\mathrm{III} \geq s\sigma\sum_{n=0}^{N-1}\left[\int_{O_n}\big(|\psi_n|-|\psi_{n+1}|\big)\big(|\psi_n|-1\big)\,d\mathbf{x}\right. \\ \left.+\int_{O_{n+1}}\big(|\psi_{n+1}|-|\psi_n|\big)\big(|\psi_{n+1}|-1\big)\,d\mathbf{x}\right].$$

Splitting the domains of integration appropriately, we have

$$\mathrm{III} \geq s\sigma\sum_{n=0}^{N-1}\left[\left(\int_{O_n\cap O_{n+1}}+\int_{O_n\setminus O_{n+1}}\right)\big(|\psi_n|-|\psi_{n+1}|\big)\big(|\psi_n|-1\big)\,d\mathbf{x}\right. \\ \left.+\left(\int_{O_{n+1}\cap O_n}+\int_{O_{n+1}\setminus O_n}\right)\big(|\psi_{n+1}|-|\psi_n|\big)\big(|\psi_{n+1}|-1\big)\,d\mathbf{x}\right].$$

Rearranging the terms in the integrand of the above, we get

$$\mathrm{III} \geq s\sigma\sum_{n=0}^{N-1}\left[\int_{O_n\cap O_{n+1}}\big(|\psi_n|^2-2|\psi_n||\psi_{n+1}|+|\psi_{n+1}|^2\big)\,d\mathbf{x}\right. \\ +\int_{O_n\setminus O_{n+1}}\big(|\psi_n|-|\psi_{n+1}|\big)\big(|\psi_n|-1\big)\,d\mathbf{x} \\ \left.+\int_{O_{n+1}\setminus O_n}\big(|\psi_{n+1}|-|\psi_n|\big)\big(|\psi_{n+1}|-1\big)\,d\mathbf{x}\right].$$

Since $|\psi_n|-|\psi_{n+1}|\geq|\psi_n|-1\geq 0$ on $O_n\setminus O_{n+1}$ and $|\psi_{n+1}|-|\psi_n|\geq|\psi_{n+1}|-1\geq 0$ on $O_{n+1}\setminus O_n$, we have

$$\mathrm{III} \geq s\sigma\sum_{n=0}^{N-1}\left[\int_{O_n\cap O_{n+1}}\big(|\psi_n|-|\psi_{n+1}|\big)^2\right. \\ \left.+\int_{O_n\setminus O_{n+1}}\big(|\psi_n|-1\big)^2+\int_{O_{n+1}\setminus O_n}\big(|\psi_{n+1}|-1\big)^2\right]\geq 0. \tag{2.5}$$

Using (2.5) and the fact that $\mathrm{I} \geq 0$, $\mathrm{II} \geq 0$, and $\mathrm{I} + \mathrm{II} + \mathrm{III} = 0$, we obtain $\mathrm{I} = \mathrm{II} = \mathrm{III} = 0$ so that

$$\sum_{n=0}^{N} \int_{O_n} |\psi_n| \big(|\psi_n| + 1\big)\big(|\psi_n| - 1\big)^2 \, d\mathbf{x} = 0.$$

Therefore, the measure of O_n is zero, i.e., $|\psi_n| \leq 1$ almost everywhere in Ω for all $n = 0, 1, \ldots, N$. $\square$

The above result, namely, $|\psi_n| \leq 1$ for all $n = 0, 1, \ldots, N$, was assumed as a hypothesis in a result of Chapman, Du, and Gunzburger in [4], in which case it was shown that if the parameters in the Lawrence–Doniach system are related by the equation $\gamma\sigma s^2 = \kappa^{-2}$, then minimizers of the Lawrence–Doniach energy (with S fixed and $s = S/N \to 0$), after extending $\{\psi_n\}$ to be defined on $D = \Omega \times (0, S)$ by linear interpolation in x_3 between consecutive layers, form a minimizing sequence for an appropriate anisotropic Ginzburg–Landau energy. (See Lemma 5.5, Theorem 5.1, and Corollary 5.6 of [4].) Thus, our Theorem 2.4 and the results in [4] can be combined to obtain this limiting result without the added hypothesis $|\psi_n| \leq 1$ in Ω for all $n = 0, 1, \ldots, N$. For completeness, we describe the anisotropic Ginzburg–Landau energy and state this consequence in section 6 of this paper.

We need the following lemma, which follows by approximation with smooth functions.

Lemma 2.5. *Let $\vec{C} = (C^1, C^2, C^3) \in H^1_{loc}(\mathbb{R}^3; \mathbb{R}^3)$ such that $\nabla \cdot \vec{C} = 0$ and let $\vec{B} = (B^1, B^2, B^3) \in H^1(\mathbb{R}^3; \mathbb{R}^3)$ with bounded support. Then*

$$\int_{\mathbb{R}^3} \left(\nabla \times \vec{C}\right) \cdot \left(\nabla \times \vec{B}\right) d\vec{x} = \int_{\mathbb{R}^3} \left(\nabla \vec{C}\right) \cdot \left(\nabla \vec{B}\right) d\vec{x}, \tag{2.6}$$

where

$$\left(\nabla \vec{C}\right) \cdot \left(\nabla \vec{B}\right) = \sum_{i=1}^{3} \nabla C^i \cdot \nabla B^i.$$

We can now improve the regularity of A^3, the third component of $\vec{A}$, for weak solutions in the gauge (1.5).

Theorem 2.6. *Let $\left(\{\psi_n\}_{n=0}^N, \vec{A}\right)$ be a weak solution of (1.3) in $[\mathcal{H}^1(\Omega)]^{N+1} \times K$. Then $A^3 \in W^{2,q}_{loc}(\mathbb{R}^3)$ for all $1 < q < \infty$, and thus $A^3 \in C^{1,\beta}_{loc}(\mathbb{R}^3)$ for all $0 < \beta < 1$.*

Proof. Let $\vec{B} = (0, 0, B^3)$ in (2.3b), where $B^3 \in H^1(\mathbb{R}^3)$ with bounded support. Using Lemma 2.5, we have

$$\Delta(A^3 - ha^3) = \Delta A^3 = \frac{s\sigma}{2}\,\mathrm{I} \text{ in } \mathbb{R}^3,$$

where I is the function defined in (2.4). By Theorem 2.4, we have $\frac{s\sigma}{2}\,\mathrm{I} \in L^\infty(\mathbb{R}^3)$. Since $\mathrm{I} = 0$ in $\mathbb{R}^3 \backslash \overline{D}$ and $A \in K$, it follows from this and (1.3) that $A^3 - ha^3$ is the Newtonian potential of $\frac{s\sigma}{2}\,\mathrm{I}$ in $\mathbb{R}^3$. The proof now follows from standard elliptic regularity results and the Sobolev imbedding theorem; see, e.g., [10]. $\square$

3. Integral representations of ^{1} ^{2} ^{1}, and ^{2}. Throughout this section, let $\left(\{\psi_n\}_{n=0}^N, \vec{A}\right)$ denote a weak solution of (1.3) in the gauge (1.5) so that

$$\left(\{\psi_n\}_{n=0}^N, \vec{A}\right) \in [\mathcal{H}^1(\Omega)]^{N+1} \times K.$$

Thus,

$$\vec{A} - h\vec{a} \in \check{H}^1(\mathbb{R}^3), \quad \nabla\cdot(\vec{A} - h\vec{a}) = 0 \quad \text{in } \mathbb{R}^3. \tag{3.1}$$

By choosing an appropriate $\vec{B} = (B^1, B^2, B^3)$ in (2.3b), we obtain

$$\Delta(A^i - ha^i) = j_i \quad \text{in } \mathcal{D}'(\mathbb{R}^3) \tag{3.2}$$

for $i = 1, 2$. Here j_i is the measure defined in (1.3), i.e.,

$$j_i(\varphi) = \sum_{n=0}^{N} \int_{\mathbb{R}^2} g_n^i(\mathbf{x}, ns)\varphi(\mathbf{x}, ns)\, d\mathbf{x} \tag{3.3}$$

for all $\varphi \in C_0^\infty(\mathbb{R}^3)$ and $i = 1, 2$, where

$$g_n^i(\mathbf{x}, ns) \equiv h_n^i(\mathbf{x}) = s\chi_\Omega(\mathbf{x}) \left[\frac{\imath}{2\kappa} \left(\psi_n^* \frac{\partial \psi_n}{\partial x_i} - c.c. \right) + A_n^i \left|\psi_n\right|^2 \right](\mathbf{x}) \tag{3.4}$$

for all $\mathbf{x} \in \mathbb{R}^2$, $n = 0, 1, \ldots, N$, and $i = 1, 2$. Note that $A_n^i \in H_{loc}^{\frac{1}{2}}(\mathbb{R}^2)$ by the trace theorem and that h_n^i is real valued. By Theorem 2.4, $h_n^i \in L^2(\mathbb{R}^2)$ (with support in $\overline{\Omega}$), and hence $g_n^i \in L^2(\mathbb{R}^2 \times \{ns\})$ (with support in $\overline{\Omega}_n$) for all $n = 0, 1, \ldots, N$ and $i = 1, 2$. Thus $j_i \in \mathcal{E}'(\mathbb{R}^3)$ for $i = 1, 2$, where $\mathcal{E}'(\mathbb{R}^3)$ is the class of tempered distributions in $\mathbb{R}^3$.

One can extend j_i to be defined in $C_c(\mathbb{R}^3)$ by formula (3.3), where $C_c(\mathbb{R}^3)$ is the set of continuous functions defined in $\mathbb{R}^3$ with compact support. Let us denote the extension again by j_i. Thus, j_i is a measure in $\mathbb{R}^3$ with support in $\bigcup_{n=0}^N \overline{\Omega}_n$.

For $\vec{x}$ in $\mathbb{R}^3$, let $\Gamma_3(\vec{x}) = \frac{c}{|\vec{x}|}$ be the fundamental solution of Δ in $\mathbb{R}^3$ (so that $c = -\frac{1}{4\pi}$). Define

$$\begin{aligned} [S(g_n^i)](\vec{x}) &= \int_{\mathbb{R}^2 \times \{ns\}} \frac{c}{|\vec{x} - \vec{Q}|} g_n^i(\vec{Q})\, d\sigma(\vec{Q}) \\ &= \int_{\mathbb{R}^2} \frac{c}{\left|\vec{x} - (\mathbf{y}, ns)\right|} h_n^i(\mathbf{y})\, d\mathbf{y}, \end{aligned} \tag{3.5}$$

where $d\sigma$ denotes surface measure on $\mathbb{R}^2 \times \{ns\}$. Note that $S(g_n^i)$ is the single layer potential of g_n^i on $\mathbb{R}^2 \times \{ns\}$. By the L^2 theory of single layer potentials on smooth domains (see [8]), $S(g_n^i) \in W_{loc}^{1,2}(\mathbb{R}^3) \cap C^\infty(\mathbb{R}^3 \backslash \overline{\Omega_n})$.

Our analysis of the regularity of $A^i - ha^i$ for $i = 1, 2$ is based on the following observations.

LEMMA 3.1. *$A^i - ha^i = \Gamma_3 * j_i = \sum_{n=0}^N S(g_n^i)$ in $L_{loc}^2(\mathbb{R}^3)$.*

Proof. Let $\hat{A} = \vec{A} - h\vec{a}$. By (2.3), we have $\Delta \hat{A}^i = j_i$ in $\mathcal{D}'(\mathbb{R}^3)$ for $i = 1, 2$. Thus for all $\varphi \in C_c^\infty(\mathbb{R}^3)$,

$$\begin{aligned} \langle \hat{A}^i, \varphi \rangle &= \langle \Delta\Gamma_3 * \hat{A}^i, \varphi \rangle \\ &= \langle \Gamma_3 * \Delta \hat{A}^i, \varphi \rangle = \langle \Gamma_3 * j_i, \varphi \rangle, \end{aligned}$$

and hence $\hat{A}^i = \Gamma_3 * j_i$ in $\mathcal{D}'(\mathbb{R}^3)$. On the other hand,

$$\begin{aligned} \langle S(g_n^i), \varphi \rangle &= \int_{\mathbb{R}^3} S(g_n^i)(\vec{x})\varphi(\vec{x})\, d\vec{x} \\ &= \int_{\mathbb{R}^3} \varphi(\vec{x}) \int_{\mathbb{R}^2} \frac{c}{\left|\vec{x} - (\mathbf{y}, ns)\right|} h_n^i(\mathbf{y})\, d\mathbf{y}\, d\vec{x}. \end{aligned}$$

Thus (by Fubini's theorem)

$$\begin{aligned}\langle \Gamma_3 * j_i, \varphi\rangle &= \langle j_i * \Gamma_3, \varphi\rangle \\ &= \sum_{n=0}^{N} \int_{\mathbb{R}^2} h_n^i(\mathbf{y}) \int_{\mathbb{R}^3} \frac{c}{|\vec{x}-(\mathbf{y},ns)|}\varphi(\vec{x})\, d\vec{x}\, d\mathbf{y} \\ &= \sum_{n=0}^{N} \int_{\mathbb{R}^3} \varphi(\vec{x}) \int_{\mathbb{R}^2} \frac{c}{|\vec{x}-(\mathbf{y},ns)|} h_n^i(\mathbf{y})\, d\mathbf{y}\, d\vec{x} \\ &= \sum_{n=0}^{N} \int_{\mathbb{R}^3} \varphi(\vec{x})[S(g_n^i)](\vec{x})\, d\vec{x},\end{aligned}$$

and we conclude that $\hat{A}^i = \Gamma_3 * j_i = \sum_{n=0}^N S(g_n^i)$ in $\mathcal{D}'(\mathbb{R}^3)$ for $i = 1,2$. Since $\hat{A}^i$ and $S(g_n^i)$ are in $L^2_{loc}(\mathbb{R}^3)$ for all $n = 0,1,\dots,N$ and $i = 1,2$, it follows that $\hat{A}^i = \sum_{n=0}^N S(g_n^i)$ in $L^2_{loc}(\mathbb{R}^3)$. $\square$

Define $t_n^i \in L^2_{loc}(\mathbb{R}^2 \times \{ns\})$ for all $n = 0,1,\dots,N$ and $i = 1,2$ by

$$\begin{aligned} t_n^i(\mathbf{x}, ns) &= \int_{\mathbb{R}^2\times\{ns\}} \frac{c}{|(\mathbf{x},ns)-\vec{Q}|} g_n^i(\vec{Q})\, d\sigma(\vec{Q}) \\ &= \int_{\mathbb{R}^2} \frac{c}{|\mathbf{x}-\mathbf{y}|} h_n^i(\mathbf{y})\, d\mathbf{y}. \end{aligned} \tag{3.6}$$

We shall need the following notation concerning nontangential limits and nontangential maximal functions on the plane $\{x_3 = ns\}$ for $n = 0,1,\dots,N$ in order to state our next result. Fix any $R > 0$ and assume that $0 < \theta < \frac{\pi}{2}$. Let

$$\Gamma \equiv \Gamma_{R,\theta} = \{\vec{x} \in \mathbb{R}^3 : |\vec{x}| < R \text{ and } |\vec{x}\cdot\vec{e_3}| > |\vec{x}|\cdot\cos\theta\}.$$

Let $\Gamma^+ = \{\vec{x} \in \Gamma : x_3 > 0\}$ and $\Gamma^- = \{\vec{x} \in \Gamma : x_3 < 0\}$. For each $\mathbf{x} \in \mathbb{R}^2$ and $n \in \{0,1,\dots,N\}$, let

$$\Gamma(\mathbf{x}, ns) = \{\vec{y} \in \mathbb{R}^3 : \vec{y} - (\mathbf{x}, ns) \in \Gamma\}.$$

Similarly, let $\Gamma^+(\mathbf{x}, ns) = \{\vec{y} \in \mathbb{R}^3 : \vec{y} - (\mathbf{x}, ns) \in \Gamma^+\}$ and $\Gamma^-(\mathbf{x}, ns) = \{\vec{y} \in \mathbb{R}^3 : \vec{y} - (\mathbf{x}, ns) \in \Gamma^-\}$. Note that $\Gamma(\mathbf{x}, ns)$ is a cone which is nontangential to the plane $\mathbb{R}^2 \times \{ns\}$ at $(\mathbf{x}, ns)$, since $\overline{\Gamma(\mathbf{x}, ns)} \bigcap [\mathbb{R}^2 \times \{ns\}] = \{(\mathbf{x}, ns)\}$.

For a function u defined at all points $\vec{y}$ in $\Gamma(\mathbf{x}, ns)$ with $0 < |y_3 - ns|$ sufficiently small, we define the nontangential limit (n.t. lim) of $u(\vec{y})$ as $\vec{y} \to (\mathbf{x}, ns)$ by

$$\operatorname*{n.t.lim}_{\vec{y}\to(\mathbf{x},ns)} u(\vec{y}) = \lim_{\vec{y}\to(\mathbf{x},ns)} \{u(\vec{y}) : \vec{y} \in \Gamma(\mathbf{x}, ns)\},$$

provided that the limit exists for each $\Gamma = \Gamma_{R,\theta}$ with $0 < \theta < \frac{\pi}{2}$. Similarly, we define

$$\operatorname*{n.t.lim}_{\vec{y}\to(\mathbf{x},ns^+)} u(\vec{y}) \text{ and } \operatorname*{n.t.lim}_{\vec{y}\to(\mathbf{x},ns^-)} u(\vec{y})$$

by replacing $\Gamma(\mathbf{x}, ns)$ in the above definition with $\Gamma^+(\mathbf{x}, ns)$ and $\Gamma^-(\mathbf{x}, ns)$, respectively.

Finally, we define the nontangential maximal function of u at $(\mathbf{x}, ns)$, denoted $u^*(\mathbf{x}, ns) = u^*_{R,\theta}(\mathbf{x}, ns)$, by

$$u^*(\mathbf{x}, ns) = \sup\{|u(\vec{y})| : \vec{y} \in \Gamma(\mathbf{x}, ns)\}$$

for a fixed $R > 0$ and $0 < \theta < \frac{\pi}{2}$.

By standard results on single layer potentials (see [8] and [19]), we have the following.

THEOREM 3.2. *For all $n = 0, 1, \ldots, N$ and $i = 1, 2$, $S(g_n^i) \in W_{loc}^{1,2}(\mathbb{R}^3) \cap C^\infty(\mathbb{R}^3 \backslash \overline{\Omega}_n)$, $t_n^i \in W_{loc}^{1,2}(\mathbb{R}^2 \times \{ns\})$, and t_n^i is the trace of $S(g_n^i)$ on $\mathbb{R}^2 \times \{ns\}$. The nontangential maximal functions of $S(g_n^i)$ and $\nabla S(g_n^i)$ from above and below $\mathbb{R}^2 \times \{ns\}$ are in $L_{loc}^2(\mathbb{R}^2 \times \{ns\})$. In addition, t_n^i and $\mathbf{grad}\, t_n^i$ are the nontangential limits of $S(g_n^i)$ and $\mathbf{grad} S(g_n^i)$, respectively, pointwise almost everywhere in $\mathbb{R}^2 \times \{ns\}$, and in $L_{loc}^2(\mathbb{R}^2 \times \{ns\})$ via the translation $(\mathbf{x}, ns + \epsilon) \to (\mathbf{x}, ns)$ as $\epsilon \to 0$. The gradients of $S(g_n^i)$ and t_n^i satisfy*

$$\begin{aligned}
\nabla S\big(g_n^i\big)(\vec{x}) &= \int_{\mathbb{R}^2 \times \{ns\}} \frac{-c\big(\vec{x} - \vec{Q}\big)}{\big|\vec{x} - \vec{Q}\big|^3} g_n^i\big(\vec{Q}\big)\, d\sigma\big(\vec{Q}\big) \\
&= \int_{\mathbb{R}^2} \frac{-c\big(\vec{x} - (\mathbf{y}, ns)\big)}{\big|\vec{x} - (\mathbf{y}, ns)\big|^3} h_n^i(\mathbf{y})\, d\mathbf{y} \;\; \text{a.e. in } \mathbb{R}^3, \\
(\mathbf{grad}\, t_n^i)(\mathbf{x}, ns) &= P.V. \int_{\mathbb{R}^2 \times \{ns\}} \frac{-c\big((\mathbf{x}, ns) - \vec{Q}\big)}{\big|(\mathbf{x}, ns) - \vec{Q}\big|^3} g_n^i\big(\vec{Q}\big)\, d\sigma\big(\vec{Q}\big) \\
&= P.V. \int_{\mathbb{R}^2} \frac{-c(\mathbf{x} - \mathbf{y})}{|\mathbf{x} - \mathbf{y}|^3} h_n^i(\mathbf{y})\, d\mathbf{y} \;\; \text{a.e. in } \mathbb{R}^2 \times \{ns\},
\end{aligned}$$

and

$$\begin{aligned}
&\underset{\vec{y} \to (\mathbf{x}, ns)}{\text{n.t.lim}}\; \mathbf{grad}\, S\big(g_n^i\big)(\vec{y}) = \mathbf{grad}\, t_n^i(\mathbf{x}, ns), \\
&\underset{\vec{y} \to (\mathbf{x}, ns^+)}{\text{n.t.lim}} \left(\frac{\partial S(g_n^i)}{\partial x_3}\right)(\vec{y}) = \frac{1}{2} g_n^i(\mathbf{x}, ns) = \frac{1}{2} h_n^i(\mathbf{x}), \\
&\underset{\vec{y} \to (\mathbf{x}, ns^-)}{\text{n.t.lim}} \left(\frac{\partial S(g_n^i)}{\partial x_3}\right)(\vec{y}) = -\frac{1}{2} g_n^i(\mathbf{x}, ns) = -\frac{1}{2} h_n^i(\mathbf{x})
\end{aligned}$$

pointwise a.e. in $\mathbb{R}^2 \times \{ns\}$, and in $L_{loc}^2(\mathbb{R}^2 \times \{ns\})$ via the translation $(\mathbf{x}, ns + \epsilon) \to (\mathbf{x}, ns)$, as $\epsilon \to 0$ (or $\epsilon \to 0^+$ or $\epsilon \to 0^-$, respectively). Here P.V. denotes the principal-valued integral.

Define $\mathbf{a}_n(\mathbf{x}) = (a_n^1(\mathbf{x}), a_n^2(\mathbf{x})) = (a^1(\mathbf{x}, ns), a^2(\mathbf{x}, ns))$. A consequence of Lemma 3.1, Theorem 3.2, and the fact that $S(g_n^i) \in C^\infty(\mathbb{R}^3 \backslash \bigcup_{n=0}^N \overline{\Omega}_n)$ is Corollary 3.3.

COROLLARY 3.3. *For all $n = 0, 1, \ldots, N$ and $i = 1, 2$, we have*

$$\begin{aligned}
A_n^i(\mathbf{x}) - h a_n^i(\mathbf{x}) &= t_n^i(\mathbf{x}, ns) + \sum_{\substack{k=0 \\ k \neq n}}^N S(g_k^i)(\mathbf{x}, ns) \\
&= \sum_{k=0}^N \int_{\mathbb{R}^2} \frac{c}{\big|(\mathbf{x}, ns) - (\mathbf{y}, ks)\big|} h_k^i(\mathbf{y})\, d\mathbf{y}
\end{aligned}$$

for almost every $\mathbf{x}$ in $\mathbb{R}^2$. Thus, $A_n^i \in W_{loc}^{1,2}(\mathbb{R}^2)$ for all n and i as above.

Proof. Fix i and n. The formula for $A_n^i - h a_n^i$ follows from Lemma 3.1 and Theorem 3.2. Since a_n^i is smooth, we have $A_n^i \in W_{loc}^{1,2}(\mathbb{R}^2)$. $\square$

Remark 3.4. We remark that if $g_n^i \in L^p(\Omega_n)$ for $2 < p < \infty$, the L^p theory of single layer potentials on C^1 domains gives the results of Theorem 3.2 and Corollary 3.3 with L_{loc}^2 and $W_{loc}^{1,2}$ replaced by L_{loc}^p and $W_{loc}^{1,p}$ throughout. (See [8].)

4. Higher regularity of solutions to the Lawrence–Doniach system. Throughout this section, let $\left(\{\psi_n\}_{n=0}^N, \vec{A}\right)$ denote a weak solution of (1.3) in the Coulomb gauge (1.5). In Corollary 3.3, we improved the regularity of A_n^i for $i = 1, 2$ from $H_{loc}^{\frac{1}{2}}(\mathbb{R}^2)$ to $W_{loc}^{1,2}(\mathbb{R}^2)$. In this section, we use this improved regularity together with the L^p theory of single layer potentials and elliptic estimates to prove higher regularity of ψ_n, $\mathbf{A}_n$, and $\vec{A}$. In particular, we prove that ψ_n and $\mathbf{A}_n$ are C^1 in Ω for each n, A^1 and A^2 are continuous in $\mathbb{R}^3$ and piecewise C^1 in $\Omega \times \mathbb{R}$, and all discontinuities in the gradient of A^i for $i = 1, 2$ occur as jump discontinuities in $\frac{\partial A^i}{\partial x_3}$ at those points of Ω_n at which $g_n^i \neq 0$.

Although our only assumption on Ω throughout this paper is that it is a bounded Lipschitz domain, we also state the regularity results, which hold when Ω is a bounded $C^{1,1}$ domain, and which are of independent interest, since most of the literature on Ginzburg–Landau equations has been written for the case of bounded smooth domains.

Lemma 4.1. *$\psi_n \in W_{loc}^{2,2}(\Omega) \cap W^{1,p}(\Omega) \cap C^\alpha(\overline{\Omega})$ for some $2 < p < \infty$, $0 < \alpha < 1$, and for all $n = 0, 1, \ldots, N$. Moreover, if Ω is a $C^{1,1}$ domain, then $\psi_n \in W^{2,2}(\Omega) \cap W^{1,q}(\Omega) \cap C^\beta(\overline{\Omega})$ for all $1 < q < \infty$ and $0 < \beta < 1$.*

Proof. Let **div** denote the divergence operator with respect to the x_1 and x_2 coordinates. By (1.5), Theorem 2.6, and Corollary 3.3, $\mathbf{div}\mathbf{A}_n = \frac{\partial A_n^1}{\partial x_1} + \frac{\partial A_n^2}{\partial x_2}$ is well defined and is in $C_{loc}^\beta(\mathbb{R}^2)$ for all $0 < \beta < 1$. Using this result, we can rewrite (1.3a) and (1.3c) as follows:

$$\begin{cases} \dfrac{1}{\kappa^2}\Delta\psi_n = F_n & \text{in } \Omega, \\ \dfrac{\partial\psi_n}{\partial\mathbf{n}} = G_n & \text{in } \partial\Omega \end{cases} \tag{4.1}$$

for all $n = 0, 1, \ldots, N$, where

$$\begin{cases} F_n = \left(|\psi_n|^2 - 1\right)\psi_n + \dfrac{2i}{\kappa}\mathbf{A}_n \cdot \mathbf{grad}\,\psi_n + \dfrac{\imath}{\kappa}\psi_n\mathbf{div}\mathbf{A}_n + |\mathbf{A}_n|^2\psi_n + P_n, \\ G_n = i\kappa(\psi_n\mathbf{A}_n)\cdot\mathbf{n} \end{cases} \tag{4.2}$$

for all $n = 0, 1, \ldots, N$, and where P_n is the bounded function defined in (1.3). By Theorem 2.4 and Corollary 3.3, $F_n \in L^{2-\varepsilon}(\Omega)$ for all $0 < \epsilon < 1$, and thus $\psi_n \in W_{loc}^{2,2-\epsilon}(\Omega)$ by interior elliptic estimates. It follows that $F_n \in L_{loc}^2(\Omega)$ and $\psi_n \in W_{loc}^{2,2}(\Omega) \cap W^{1,2}(\Omega) \cap L^\infty(\Omega)$ for all $n = 0, 1, \ldots, N$.

Set $w = \Gamma_2 * \chi_\Omega F_n$, where Γ_2 is the fundamental solution for the Laplacian in $\mathbb{R}^2$. By interior elliptic estimates, $w \in W_{loc}^{2,2-\varepsilon}(\mathbb{R}^2) \bigcap W_{loc}^{2,2}(\Omega)$ for all $0 < \epsilon < 1$. Hence $\frac{\partial w}{\partial x_j} \in W_{loc}^{1,2-\epsilon}(\mathbb{R}^2) \cap L_{loc}^q(\mathbb{R}^2)$ for all $0 < \epsilon < 1$ and $1 < q < \infty$, and $w \in C_{loc}^\beta(\mathbb{R}^2)$ for all $0 < \beta < 1$ by Sobolev imbedding. Since $\nabla[(\frac{\partial w}{\partial x_j})^q] = q(\frac{\partial w}{\partial x_j})^{q-1} \cdot \nabla(\frac{\partial w}{\partial x_j}) \in L_{loc}^{2-\epsilon}(\mathbb{R}^2)$ by Holder's inequality, we have $(\frac{\partial w}{\partial x_j})^q \in W_{loc}^{1,2-\epsilon}(\mathbb{R}^2)$ for all $1 < q < \infty$ and $0 < \epsilon < 1$. Similarly, since $A_n^i \in W_{loc}^{1,2}(\mathbb{R}^2) \subset L_{loc}^q(\mathbb{R}^2)$ for all $1 < q < \infty$, $0 < \epsilon < 1$, and $i = 1, 2$, $\nabla[(A_n^i)^q] = q(A_n^i)^{q-1} \cdot \nabla A_n^i \in L_{loc}^{2-\epsilon}(\mathbb{R}^2)$, and hence $(A_n^i)^q \in W_{loc}^{1,2-\epsilon}(\mathbb{R}^2)$ for all $1 < q < \infty$ and $0 < \epsilon < 1$. Thus the traces of ∇w and A_n^i are in $L^q(\partial\Omega)$ for all $1 < q < \infty$ and $G_n \in L^q(\partial\Omega)$ for all $1 < q < \infty$.

Now set $U = \psi_n - w$. Then $U \in W^{1,2}(\Omega) \cap W_{loc}^{2,2}(\Omega)$, $\Delta U = 0$ in Ω, and $\nabla U \cdot \mathbf{n} = G_n - \nabla w \cdot \mathbf{n} \equiv \tilde{G}_n \in L^q(\partial\Omega)$ for all $1 < q < \infty$. By existence-uniqueness results (up to an additive constant) on the Neumann problem in $W^{1,r}(\Omega)$ with compatible L^r

Neumann data in bounded Lipschitz domains for $1 < r < 2+\epsilon_0$, where ϵ_0 is a positive constant depending only on the domain (see [5], [12], [16], and [19]), we conclude that $U \in W^{1,p}(\Omega)$ for some $p \in (2, 2+\epsilon_0)$, and thus $\psi_n \in W^{2,2}_{loc}(\Omega) \cap W^{1,p}(\Omega) \cap C^{\alpha}(\overline{\Omega})$ for $\alpha = 1 - \frac{2}{p}$ by Sobolev imbedding.

If Ω is a $C^{1,1}$ domain, since $\tilde{G}_n$ is in $L^q(\partial\Omega)$, existence-uniqueness results for the Neumann problem in C^1 domains with L^q Neumann data imply that $U \in W^{1,q}(\Omega)$ for all $1 < q < \infty$. (See [8].) Thus $\psi_n \in W^{1,q}(\Omega)$ for all $1 < q < \infty$, and it follows from this and Corollary 3.3 that $G_n \in H^{\frac{1}{2}}(\partial\Omega)$ and $F_n \in L^2(\Omega)$. Thus, $\psi_n \in W^{2,2}(\Omega)$ by the elliptic regularity theory for $C^{1,1}$ domains. The desired result in this case now follows by Sobolev imbedding. $\square$

For the remainder of this section, let $p \in (2,\infty)$ and $\alpha \in (0,1)$ denote the constants of Lemma 4.1.

LEMMA 4.2. *For all $n = 0, 1, \ldots, N$ and $i = 1, 2$, $A^i_n \in W^{1,p}_{loc}(\mathbb{R}^2) \cap C^{\alpha}_{loc}(\mathbb{R}^2)$. Moreover, if Ω is a $C^{1,1}$ domain, then $A^i_n \in W^{1,q}_{loc}(\mathbb{R}^2) \cap C^{\beta}_{loc}(\mathbb{R}^2)$ for all $1 < q < \infty$ and $0 < \beta < 1$.*

Proof. Fix n and i as above. By (3.4), Theorem 2.4, Corollary 3.3, Lemma 4.1, and the Sobolev imbedding theorem, we have $g^i_n \in L^p(\mathbb{R}^2 \times \{ns\})$ with suppt $g^i_n \in \overline{\Omega}_n$. Since t^i_n is the trace of $S(g^i_n)$ on $\mathbb{R}^2 \times \{ns\}$, it follows from Remark 3.4 that $t^i_n \in W^{1,p}_{loc}(\mathbb{R}^2 \times \{ns\})$. By Corollary 3.3, we have $A^i_n \in W^{1,p}_{loc}(\mathbb{R}^2)$ and thus $A^i_n \in C^{\alpha}_{loc}(\mathbb{R}^2)$.

If Ω is a $C^{1,1}$ domain, then Lemma 4.1 implies that $g^i_n \in L^q(\mathbb{R}^2 \times \{ns\})$ for all $1 < q < \infty$, and thus $A^i_n \in W^{1,q}_{loc}(\mathbb{R}^2) \cap C^{\beta}_{loc}(\mathbb{R}^2)$ in this case. $\square$

The additional regularity obtained for A^i_n and ψ_n in Lemma 4.2 yields additional regularity on A^i for $i = 1, 2$.

LEMMA 4.3. *For $i = 1, 2$, $A^i \in W^{1,p}_{loc}(\mathbb{R}^3) \cap C^{\alpha}_{loc}(\mathbb{R}^3)$. Moreover, if Ω is a $C^{1,1}$ domain, then $A^i \in W^{1,q}_{loc}(\mathbb{R}^3) \cap C^{\beta}_{loc}(\mathbb{R}^3)$ for all $1 < q < \infty$ and $0 < \beta < 1$.*

Proof. Fix i as above. Since $A^i \in W^{1,2}_{loc}(\mathbb{R}^3)$ and ha^i is harmonic in $\mathbb{R}^3$ for $i = 1, 2$, it follows from (1.3) and (1.5) that A^i is harmonic in $\mathbb{R}^3 \setminus \bigcup_{n=0}^{N} \overline{\Omega}_n$ and the trace of A^i on $\mathbb{R}^2 \times \{ns\}$ is in $W^{1,p}_{loc}(\mathbb{R}^2)$ by Lemma 4.2. Thus elliptic regularity in C^1 domains implies that $A^i \in W^{1,p}_{loc}(\mathbb{R}^3) \cap C^{\alpha}_{loc}(\mathbb{R}^3)$. (See [8].)

If Ω is a $C^{1,1}$ domain, the above holds with p and α replaced by q and β for any $1 < q < \infty$ and $0 < \beta < 1$. $\square$

THEOREM 4.4. *$\psi_n \in C^{2,\alpha}(\Omega)$ and, if Ω is a $C^{1,1}$ domain, $\psi_n \in W^{2,q}(\Omega) \cap C^{1,\beta}(\overline{\Omega}) \cap C^{2,\beta}(\Omega)$ for all $1 < q < \infty$ and $0 < \beta < 1$. Moreover, if Ω is a $C^{2,\delta}$ domain with $0 < \delta < 1$, $\psi_n \in C^{2,\delta}(\overline{\Omega})$.*

Proof. By Lemma 4.1, $\psi_n \in W^{1,p}(\Omega) \cap C^{\alpha}(\overline{\Omega})$. By Lemma 4.2 and the Sobolev imbedding theorem, $F_n \in L^p(\Omega)$. It follows from interior elliptic estimates that $\psi_n \in W^{2,p}_{loc}(\Omega)$, and thus $\psi_n \in C^{1,\alpha}(\Omega)$. By this and Theorem 2.6, we have $F_n \in C^{\alpha}(\Omega)$ and thus $\psi_n \in C^{2,\alpha}(\Omega)$.

If Ω is a $C^{1,1}$ domain, then, arguing as above, we have $F_n \in L^q(\Omega)$. By the trace theorem, Lemma 4.1, and Lemma 4.2, we have $G_n \in W^{1-\frac{1}{q},q}(\partial\Omega) \cap C^{\beta}(\partial\Omega)$ for all $1 < q < \infty$, $0 < \beta < 1$, and $n = 0, \ldots, N$. Thus $\psi_n \in W^{2,q}(\Omega) \cap C^{1,\beta}(\overline{\Omega})$ for all $1 < q < \infty$ and $0 < \beta < 1$ by the elliptic regularity theory. It follows that $F_n \in C^{\beta}(\overline{\Omega})$ for all $0 < \beta < 1$ and $\psi_n \in C^{2,\beta}(\Omega)$. Moreover, if $\partial\Omega$ is $C^{2,\delta}$, then $\psi_n \in C^{2,\delta}(\overline{\Omega})$. $\square$

COROLLARY 4.5. *$A^i_n \in C^1(\Omega)$ for all $n = 0, 1, \ldots, n$ and $i = 1, 2$.*

Proof. Fix n and i as above. By Corollary 3.3, it suffices to show that $t^i_n(\cdot, ns) \in C^1(\Omega)$. By (3.4), Lemmas 4.1 and 4.2, and Theorem 4.4, $h^i_n \in L^p(\mathbb{R}^2) \cap C^{\alpha}(\Omega)$. Thus, by Theorem 3.2 and Remark 3.4, $t^i_n(\cdot, ns) \in W^{1,p}_{loc}(\mathbb{R}^2) \subset C^{\alpha}_{loc}(\mathbb{R}^2)$. In addition, for each open set Ω' such that $\overline{\Omega'} \subset\subset \Omega$, there is a constant $C = C(\Omega', \Omega)$ such that

$$|h_n^i(\mathbf{y}) - h_n^i(\mathbf{x})| \leq C|\mathbf{x} - \mathbf{y}|^\alpha \text{ for all } \mathbf{x}, \mathbf{y} \in \Omega'.$$

Recall that

$$t_n^i(\mathbf{x}, ns) = \int_{\mathbb{R}^2} \frac{c}{|\mathbf{x} - \mathbf{y}|} h_n^i(\mathbf{y})\, d\mathbf{y} = \int_\Omega \frac{c}{|\mathbf{x} - \mathbf{y}|} h_n^i(\mathbf{y})\, d\mathbf{y}.$$

To show that $t_n^i(\cdot, ns)$ is C^1 in Ω, we use an argument similar to that found in Lemma 4.2 of [10]. Let $\zeta(\mathbf{x})$ denote $\frac{c}{|\mathbf{x}|}$ and $\eta(t)$ denote a continuously differentiable function such that $0 \leq \eta \leq 1$, $0 \leq \eta' \leq 2$, $\eta(t) = 0$ if $t \leq 1$, and $\eta(t) = 1$ if $t \geq 2$. For $\epsilon > 0$, define $\eta_\epsilon(t)$ to be $\eta\left(\frac{t}{\epsilon}\right)$. Let

$$t_{n,\epsilon}^i(\mathbf{x}, ns) = \int_\Omega \zeta(\mathbf{x} - \mathbf{y}) \cdot \eta_\epsilon(|\mathbf{x} - \mathbf{y}|) \cdot h_n^i(\mathbf{y})\, d\mathbf{y}.$$

Note that $t_{n,\epsilon}^i \in C^1(\mathbb{R}^2)$, and $t_{n,\epsilon}^i(\cdot, ns)$ converges uniformly in compact subsets of Ω to $t_n^i(\cdot, ns)$. For $j = 1, 2$, $\mathbf{x} \in \Omega$, and $\epsilon < \frac{1}{2} dist(\mathbf{x}, \partial\Omega)$, we have

$$\begin{aligned}
\frac{\partial}{\partial x_j}(t_{n,\epsilon}^i)(\mathbf{x}, ns) &= \int_\Omega \frac{\partial}{\partial x_j}(\zeta\eta_\epsilon) h_n^i(\mathbf{y})\, d\mathbf{y} \\
&= \int_\Omega \frac{\partial}{\partial x_j}(\zeta\eta_\epsilon)[h_n^i(\mathbf{y}) - h_n^i(\mathbf{x})]\, d\mathbf{y} + h_n^i(\mathbf{x}) \int_\Omega \frac{\partial}{\partial x_j}(\zeta\eta_\epsilon)\, d\mathbf{y} \\
&= \int_\Omega \frac{\partial}{\partial x_j}(\zeta\eta_\epsilon)[h_n^i(\mathbf{y}) - h_n^i(\mathbf{x})]\, d\mathbf{y} - h_n^i(\mathbf{x}) \int_{\partial\Omega} \zeta(\mathbf{x} - \mathbf{y}) n_j\, dl(\mathbf{y}).
\end{aligned}$$

Here, $\zeta = \zeta(\mathbf{x} - \mathbf{y})$, $\eta_\epsilon = \eta_\epsilon(\mathbf{x} - \mathbf{y})$, n_j is the jth component of the outward unit normal to $\partial\Omega$, and $dl(\mathbf{y})$ is the arclength measure on $\partial\Omega$. Define

$$u_n^{ij}(\mathbf{x}) = \int_\Omega \frac{\partial}{\partial x_j}\big(\zeta(\mathbf{x} - \mathbf{y})\big)\big(h_n^i(\mathbf{y}) - h_n^i(\mathbf{x})\big)\, d\mathbf{y} - h_n^i(\mathbf{x}) \int_{\partial\Omega} \zeta(\mathbf{x} - \mathbf{y}) n_j\, dl(\mathbf{y}).$$

Since $h_n^i \in C^\alpha(\Omega)$, it follows that u_n^{ij} is well defined in Ω, and for each subdomain Ω' as above,

$$\sup \left\{ \left| \frac{\partial}{\partial x_j}(t_{n,\epsilon}^i)(\mathbf{x}, ns) - u_n^{ij}(\mathbf{x}) \right| : \mathbf{x} \in \overline{\Omega'} \right\} \to 0$$

as $\epsilon \to 0$ and thus $u_n^{ij} \in C(\Omega)$ for all i and j in $\{1, 2\}$. We have shown that $t_{n,\epsilon}^i \to t_n^i$ and $\frac{\partial t_{n,\epsilon}^i}{\partial x_j}(\cdot, ns) \to u_n^{ij}$ uniformly on compact subsets of Ω, $u_n^{ij} \in C(\Omega)$, and $t_n^i(\cdot, ns) \in W_{loc}^{1,p}(\mathbb{R}^2) \subset C_{loc}^\alpha(\mathbb{R}^2)$. It follows that $\frac{\partial t_n^i}{\partial x_j}(\cdot, ns) = u_n^{ij}$ in the sense of distributions in Ω, and thus $t_n^i(\cdot, ns) \in C^1(\Omega)$. $\square$

Since A^i and ∇A^i are harmonic in $\mathbb{R}^3 \backslash \bigcup_{n=0}^N \overline{\Omega}_n$ for $i = 1, 2$, the continuity of $\mathbf{grad}\, A_n^i$ in Ω and g_n^i in Ω_n implies that $\frac{\partial A^i}{\partial x_j}$ is continuous up to the boundary in the subdomains of $\Omega' \times \mathbb{R}$ located between two adjacent layers Ω_n, below Ω_0, or above Ω_n, where $\overline{\Omega'} \subset\subset \Omega$. From this and the regularity results proved in sections 3 and 4, we obtain the following.

THEOREM 4.6. *Let $\big(\{\psi_n\}_{n=0}^N, \vec{A}\big)$ be a weak solution of* (1.3) *satisfying* (1.5). *There exist constants $p > 2$ and $0 < \alpha < 1$ such that for all $n = 0, \dots, N$,*

$$\begin{aligned}
&\psi_n \in W^{1,p}(\Omega) \cap C^{2,\alpha}(\Omega) \cap C^{\alpha}(\overline{\Omega}),\\
&A_n^1, A_n^2 \in W^{1,p}_{loc}(\mathbb{R}^2) \cap C^{\alpha}_{loc}(\mathbb{R}^2) \cap C^1(\Omega) \cap C^{\infty}\Big(\mathbb{R}^2 \setminus \overline{\Omega}\Big),\\
&A^1, A^2 \in W^{1,p}_{loc}(\mathbb{R}^3) \cap C^{\alpha}_{loc}(\mathbb{R}^3) \cap C^{\infty}\left(\mathbb{R}^3 \Big\backslash \bigcup_{i=o}^{N} \overline{\Omega}_n\right),\\
&\frac{\partial A^i}{\partial x_1} \text{ and } \frac{\partial A^i}{\partial x_2} \in C(\Omega \times \mathbb{R}) \text{ for } i = 1, 2,\\
&A^3 \in W^{2,q}_{loc}(\mathbb{R}^3) \cap C^{1,\beta}_{loc}(\mathbb{R}^3) \text{ for all } q \in (1, \infty) \text{ and } \beta \in (0, 1).
\end{aligned}$$

In addition, for $i = 1, 2$, A^i is piecewise C^1 in $\overline{\Omega'} \times \mathbb{R}$ for all domains $\Omega' \subset \overline{\Omega'} \subset\subset \Omega$, with all discontinuities of ∇A^i in $\Omega \times \mathbb{R}$ occurring as jump discontinuities in $\frac{\partial A^i}{\partial x_3}$ from above and below at points in Ω_n at which $g_n^i(\mathbf{x}, ns)$ is nonzero. Moreover, if Ω is a $C^{1,1}$ domain, the above results hold for all $1 < p < \infty$ and $0 < \alpha < 1$, and we have, in addition,

$$\begin{aligned}
&\psi_n \in W^{2,q}(\Omega) \cap C^{2,\beta}(\Omega) \cap C^{1,\beta}(\overline{\Omega}),\\
&A_n^1, A_n^2 \in W^{1,q}_{loc}(\mathbb{R}^2) \cap C^{\beta}_{loc}(\mathbb{R}^2),\\
&A^1, A^2 \in W^{1,q}_{loc}(\mathbb{R}^3) \cap C^{\beta}_{loc}(\mathbb{R}^3)
\end{aligned}$$

for all $1 < q < \infty$ and $0 < \beta < 1$.

Proof. The proof follows immediately from elliptic regularity and Theorem 3.2, Lemmas 4.1–4.3, Theorem 4.4, and Corollary 4.5, since $\Delta A^i = \vec{0}$ in $\mathbb{R}^3 \backslash \bigcup_{n=0}^N \overline{\Omega}_n$ for $i = 1, 2$. $\square$

5. Breakdown of superconductivity due to strong magnetic fields. In this section we prove various a priori estimates for solutions of the Lawrence–Doniach system in the gauge (1.5), which generalize results of Giorgi and Phillips (cf. [9]) for the standard (isotropic) Ginzburg–Landau model predicted by physicists (cf. [17]). Using our a priori estimates, we show that the upper critical modulus, $\bar{h}(\vec{v}, \kappa, s, \sigma, S, \Omega)$, is finite when $\vec{v}$ is nontangential to the layers.

Recall that we have assumed $\vec{H}$ is a constant applied magnetic field in $\mathbb{R}^3$ and $\vec{H} = h\vec{v}$, where $|\vec{v}| = 1$ and $h \geq 0$. The vector $\vec{v}$ is called nontangential if $\vec{v} \cdot \vec{e_3} \neq 0$.

We define $\bar{h} = \bar{h}(\vec{v}, \kappa, s, \sigma, S, \Omega)$ by

$$\begin{aligned}
\bar{h} = \inf\{h' > 0 :& \text{ normal states are the only solutions of (1.3)}\\
&\text{for all } h > h' \text{ with } \vec{H} = h\vec{v}\},
\end{aligned} \tag{5.1}$$

with $\bar{h} = \infty$ if there are no such h'.

We begin with the following observation: Let $\big(\{\psi_n\}_{n=0}^N, \vec{A}\big)$ be a weak solution of (1.3). Computing real and imaginary parts of $(\frac{\imath}{\kappa}\mathbf{grad}\ \psi_n + \mathbf{A}_n\psi_n)\psi_n^*$, we have

$$\begin{aligned}
\Re\left[\Big(\frac{\imath}{\kappa}\mathbf{grad}\ \psi_n + \mathbf{A}_n\psi_n\Big)\psi_n^*\right] &= \frac{\imath}{2\kappa}\Big(\psi_n^*\mathbf{grad}\ \psi_n - c.c.\Big) + \mathbf{A}_n|\psi_n|^2,\\
\Im\left[\Big(\frac{\imath}{\kappa}\mathbf{grad}\ \psi_n + \mathbf{A}_n\psi_n\Big)\psi_n^*\right] &= \frac{1}{2\kappa}\mathbf{grad}\ \Big(|\psi_n|^2\Big) = \frac{|\psi_n|}{\kappa}\mathbf{grad}\ (|\psi_n|).
\end{aligned}$$

Thus,

$$\begin{aligned}&\frac{\imath}{\kappa}\textbf{grad }\psi_n + \mathbf{A}_n\psi_n\\ &= \left\{\left[\frac{\imath}{2\kappa}\Big(\psi_n^*\textbf{grad }\psi_n - c.c.\Big) + \mathbf{A}_n\,|\psi_n|^2\right]|\psi_n|^{-1} + \Im\left[\frac{1}{\kappa}\textbf{grad }|\psi_n|\right]\right\}\frac{\psi_n}{|\psi_n|}\end{aligned} \tag{5.2}$$

for almost all $\mathbf{x}$ in $\{\mathbf{x} \in \Omega : \psi_n(\mathbf{x}) \neq 0\}$, which we denote by $\{\psi_n \neq 0\}$. Since $\psi_n \in H^1(\Omega)$, $\textbf{grad }\psi_n = 0$ almost everywhere in $\{\psi_n = 0\}$. (See [10, Lemma 7.7].) Therefore, (5.2) holds almost everywhere in Ω. Now, letting $\varphi_n = \psi_n$ in (2.3a), using (5.2), Theorem 2.4, and the fact that $|\psi_{n+1}|^2 + |\psi_n|^2 - \big(\psi_n\psi_{n+1}^* e^{\imath\phi_n^{n+1}} + c.c.\big) = |\psi_{n+1}e^{-\imath\phi_n^{n+1}} - \psi_n|^2$, we get the following.

Lemma 5.1. *If* $\big(\{\psi_n\}_{n=0}^N, \vec{A}\big)$ *is a weak solution of* (1.3)*, then*

$$\begin{aligned}&\sum_{n=0}^N\int_\Omega\left(\left|\frac{1}{\kappa}\textbf{grad }|\psi_n|\right|^2 + \left|\left[\frac{\imath}{2\kappa}\Big(\psi_n^*\textbf{grad }\psi_n - c.c.\Big) + \mathbf{A}_n\,|\psi_n|^2\right]|\psi_n|^{-1}\right|^2\right)d\mathbf{x}\\ &= \sum_{n=0}^N\int_\Omega\left|\frac{\imath}{\kappa}\textbf{grad }\psi_n + \mathbf{A}_n\psi_n\right|^2 d\mathbf{x}\\ &= \sum_{n=0}^N\int_\Omega\Big(1 - |\psi_n|^2\Big)|\psi_n|^2 - \sigma\sum_{n=0}^{N-1}\int_\Omega\left|\psi_{n+1}e^{-\imath\phi_n^{n+1}} - \psi_n\right|^2 d\mathbf{x}\\ &\leq \sum_{n=0}^N\int_\Omega\Big(1 - |\psi_n|^2\Big)|\psi_n|^2\, d\mathbf{x}\\ &\leq \sum_{n=0}^N\int_\Omega|\psi_n|^2\, d\mathbf{x}.\end{aligned}$$

Recall that $\mathbf{a}_n(\mathbf{x}) = \big(a_n^1(\mathbf{x}), a_n^2(\mathbf{x})\big)$ denotes the trace of the first two components of $\vec{a}(\mathbf{x}, ns)$ on $\mathbb{R}^2$. We need the following lemma.

Lemma 5.2. *Let* $\big(\{\psi_n\}_{n=0}^N, \vec{A}\big)$ *be a weak solution of* (1.3) *satisfying* (1.5)*. There exists a constant M depending only on Ω and S (and hence depending only on $\mathcal{D}$) such that*

$$\sum_{n=0}^N\int_\Omega|\mathbf{A}_n - h\mathbf{a}_n|^2\, d\mathbf{x} \leq M\sum_{n=0}^N\int_\Omega|\psi_n|^2\, d\mathbf{x}.$$

Proof. By Corollary 3.3,

$$A_n^i(\mathbf{x}) - ha_n^i(\mathbf{x}) = \sum_{k=0}^N\int_{\mathbb{R}^2}\frac{c}{\big|(\mathbf{x}, ns) - (\mathbf{y}, ks)\big|}h_k^i(\mathbf{y})\, d\mathbf{y}$$

for $i = 1, 2$. Thus

$$\begin{aligned}\int_\Omega\left|A_n^i - ha_n^i\right|^2 d\mathbf{x} &\leq \tilde{C}N\sum_{k=0}^N\int_\Omega\left|h_k^i\right|^2 d\mathbf{x}\\ &= \tilde{C}N\sum_{k=0}^N\int_{\Omega\times\{ks\}}\left|g_k^i(\mathbf{x}, ks)\right|^2 d\mathbf{x},\end{aligned} \tag{5.3}$$

where $\tilde{C}$ is a constant depending only on Ω. Since $|\psi_n| \leq 1$, it follows from formula (3.4) and Lemma 5.1 that

$$\begin{aligned}\sum_{k=0}^{N} \int_{\Omega\times\{ks\}} \left|g_k^i(\mathbf{x}, ks)\right|^2 \, d\mathbf{x} &\leq s^2 \sum_{k=0}^{N} \int_{\Omega\times\{ks\}} |\psi_k|^2 \, d\mathbf{x} \\ &= S^2 N^{-2} \sum_{k=0}^{N} \int_{\Omega\times\{ks\}} |\psi_k|^2 \, d\mathbf{x}.\end{aligned} \tag{5.4}$$

Summing over $i = 1, 2$ and $n = 0, 1, \ldots, N$ in (5.3) and using (5.4), we obtain

$$\sum_{n=0}^{N} \int_{\Omega} |\mathbf{A}_n - h\mathbf{a}_n|^2 \, d\mathbf{x} \leq \tilde{C} S^2 \sum_{n=0}^{N} \int_{\Omega} |\psi_n|^2 \, d\mathbf{x}.$$

The theorem follows with $M = \tilde{C}S^2$. ◻

LEMMA 5.3. *Let* $\left(\{\psi_n\}_{n=0}^N, \vec{A}\right)$ *be a weak solution of* (1.3) *satisfying* (1.5). *Then*

$$\sum_{n=0}^{N} \int_{\Omega} \left|(\imath \mathbf{grad} + \kappa h \mathbf{a}_n)\psi_n\right|^2 d\mathbf{x} \leq C_1 \kappa^2 \sum_{n=0}^{N} \int_{\Omega} |\psi_n|^2 \, d\mathbf{x},$$

where C_1 is a positive constant depending only on $\mathcal{D}$.

Proof. We write

$$\left(\frac{\imath}{\kappa}\mathbf{grad} + h\mathbf{a}_n\right)\psi_n = \left(\frac{\imath}{\kappa}\mathbf{grad} + \mathbf{A}_n\right)\psi_n - (\mathbf{A}_n - h\mathbf{a}_n)\psi_n. \tag{5.5}$$

Thus, we have

$$\begin{aligned}&\sum_{n=0}^{N} \int_{\Omega} \left|\left(\frac{\imath}{\kappa}\mathbf{grad} + h\mathbf{a}_n\right)\psi_n\right|^2 d\mathbf{x} \\ &\leq 2 \sum_{n=0}^{N} \left[\int_{\Omega} \left|\left(\frac{\imath}{\kappa}\mathbf{grad} + \mathbf{A}_n\right)\psi_n\right|^2 d\mathbf{x} + \int_{\Omega} \left|\mathbf{A}_n - h\mathbf{a}_n\right|^2 |\psi_n|^2 \, d\mathbf{x}\right].\end{aligned}$$

Using this inequality, Theorem 2.4, Lemma 5.1, and Lemma 5.2, we obtain

$$\sum_{n=0}^{N} \int_{\Omega} \left|\left(\frac{\imath}{\kappa}\mathbf{grad} + h\mathbf{a}_n\right)\psi_n\right|^2 d\mathbf{x} \leq 2(1+M) \sum_{n=0}^{N} \int_{\Omega} |\psi_n|^2 \, d\mathbf{x}.$$

The lemma follows with $C_1 = 2(1 + M) = C_1(\mathcal{D})$. ◻

We shall need the following result, which is an extension of Lemma 2.8 of [9]. Note that when $\theta = 0$ in the statement of this result, no positive lower bound on the right-hand side of (5.6) is obtained. This is the reason that our estimate from above on $h_{c_3}(\vec{v})$ (proved on Theorem 5.5) applies only to the case in which $\vec{H} = h\vec{v}$ is nontangential to the layers $\Omega \times \{ns\}$.

LEMMA 5.4. *Given $m > 0$, there is a constant $C_2 = C_2(m, \Omega)$ satisfying $0 < C_2 \leq 1$ such that if $\theta \in \mathbb{R}$, w is a nonzero real number, $\vec{b} = (b^1, b^2, b^3) = (\mathbf{b}, b^3)$ is a vector field in $H^1(\Omega, \mathbb{R}^3)$ satisfying $\left(\nabla \times \vec{b}\right) \cdot \vec{e_3} \equiv \left(\frac{\partial b_2}{\partial x_1} - \frac{\partial b_1}{\partial x_2}\right) \equiv \theta$ in Ω, and $w^2 |\theta| \geq m$, then*

$$C_2 w^2 |\theta| \int_{\Omega} |\zeta|^2 \, d\mathbf{x} \leq \int_{\Omega} \left|(\imath \mathbf{grad} + w^2 \mathbf{b})\zeta\right|^2 d\mathbf{x} \tag{5.6}$$

for all $\zeta \in \mathcal{H}^1(\Omega;\mathbb{C})$.

Proof. If $\theta = 0$, the inequality is trivial. Assume that $\theta \neq 0$.

It was shown by Giorgi and Phillips (see [9, Lemma 2.8]) that, given $m > 0$, there exists $C_2 = C_2(m,\Omega))$, $0 < C_2 \leq 1$, such that if $\lambda^2 \geq m$, then

$$C_2\lambda^2 \int_\Omega |\zeta|^2 \, d\mathbf{x} \leq \int_\Omega \left|(\imath\mathbf{grad} + \lambda^2\mathbf{c})\zeta\right|^2 d\mathbf{x}$$

for all $\zeta \in H^1(\Omega;\mathbb{C})$ and $\mathbf{c} = (c_1, c_2) \in H^1(\Omega;\mathbb{R}^2)$ such that $\frac{\partial c_2}{\partial x_1} - \frac{\partial c_1}{\partial x_2} \equiv 1$ in Ω. Since $\frac{1}{\theta}(\frac{\partial b_2}{\partial x_1} - \frac{\partial b_1}{\partial x_2}) = 1$ in Ω, if $\theta > 0$, we can apply the result of Giorgi and Phillips with $\mathbf{c} = \frac{1}{\theta}\mathbf{b} = \frac{1}{\theta}(b_1, b_2)$ and $\lambda^2 = \theta w^2$ to obtain (5.6). If $\theta < 0$, the result follows from (5.6) for the case just proved, applied to $\theta' = |\theta| = -\theta$, $\vec{b}' = -\vec{b}$, and $\zeta' = -(\mathrm{Re}\zeta) + i(\mathrm{Im}\zeta)$ in $\mathcal{H}^1(\Omega;\mathbb{C})$. $\square$

We can now prove the existence of a finite upper critical modulus, assuming that $\vec{H} = h\vec{v}$, where $\vec{v}$ is a unit vector in $\mathbb{R}^3$ which is nontangential to the superconducting layers, $\Omega \times \{ns\}$.

THEOREM 5.5. *Given* $m > 0$ *and* $0 < \mu < 1$*, there exists a positive constant* $\phi = \phi(m,\mathcal{D})$ *so that if* $\vec{H} = h\vec{v}$ *with* $|\vec{v}| \equiv 1$ *and* $|\vec{v}\cdot\vec{e}_3| \geq \mu$ *and if* $h > \frac{1}{\mu}\max\left(\frac{m}{\kappa}, \phi\kappa\right)$*, then any weak solution to* (1.3) *is a normal state. Thus*

$$\begin{aligned} H_{c_3}(\vec{v},\kappa,s,\sigma,S,\Omega) &\leq \bar{h}(\vec{v},\kappa,s,\sigma,S,\Omega) \\ &\leq \frac{1}{\mu}\max\left(\frac{m}{\kappa}, \phi\kappa\right). \end{aligned}$$

Moreover, the constant ϕ *can be chosen to satisfy* $\phi = \frac{C_1}{C_2}$*, where* $C_1 = C_1(\mathcal{D})$ *and* $C_2 = C_2(m,\Omega)$ *are the constants of Lemmas* 5.3 *and* 5.4*, respectively.*

Proof. Let $\left(\{\psi_n\}_{n=0}^N, \vec{A}\right)$ be a weak solution of (1.3) and assume that $h > \frac{1}{\mu}\max\left(\frac{m}{\kappa}, \phi\kappa\right)$. Note that a weak solution is a normal state if and only if its entire gauge-equivalence class consists of normal states. Thus we may assume without loss of generality that $\left(\{\psi_n\}_{n=0}^N, \vec{A}\right)$ satisfies (1.5). Set $w^2 = h\kappa$. Then, by hypothesis, we have $w^2|\vec{v}\cdot\vec{e}_3| \geq w^2\mu \geq m$. Since $[\nabla\times(\vec{a}(x_1,x_2,ns))]\cdot\vec{e}_3 = \vec{v}\cdot\vec{e}_3$ is constant in Ω for each n, we may apply (5.6) of Lemma 5.4 with $\zeta = \psi_n$, $\vec{b}(x_1,x_2) = \vec{a}(x_1,x_2,ns)$, $\mathbf{b} = \mathbf{a}_n$, and $\theta = \vec{v}\cdot\vec{e}_3$, to get

$$C_2 h\kappa\mu \sum_{n=0}^N \int_\Omega |\psi_n|^2 \, d\mathbf{x} \leq \sum_{n=0}^N \int_\Omega \left|(i\mathbf{grad} + h\kappa\mathbf{a}_n)\psi_n\right|^2 d\mathbf{x}.$$

Using this inequality, Lemma 5.3, and setting $\phi = \frac{C_1}{C_2}$, we have

$$h\sum_{n=0}^N \int_\Omega |\psi_n|^2 \, d\mathbf{x} \leq \frac{\phi\kappa}{\mu}\sum_{n=0}^N \int_\Omega |\psi_n|^2 \, d\mathbf{x}.$$

By hypothesis, $h > \frac{\phi\kappa}{\mu}$. Therefore we have $\int_\Omega |\psi_n|^2 \, d\mathbf{x} = 0$ for all $n = 0, 1, \ldots, N$, and $\psi_n = 0$ almost everywhere in Ω for all n. Using this result and (1.3), we have $\Delta(\vec{A} - h\vec{a}) = -\nabla\times\nabla\times(\vec{A} - h\vec{a}) = 0$. Since $\vec{A} - h\vec{a} \in \check{H}^1(\mathbb{R}^3)$, we conclude that $\vec{A} = h\vec{a}$, and thus $\nabla\times\vec{A} = h\vec{v} = \vec{H}$. $\square$

COROLLARY 5.6. *Assume that* $|\vec{v}| = 1$ *and* $|\vec{v}\cdot\vec{e}_3| \geq \mu > 0$*. There exist positive constants* κ_0 *and* C_0 *depending only on* $\mathcal{D}$ *such that for all* $\kappa \geq \kappa_0$*, we have*

$$H_{c_3}(\vec{v},\kappa,s,\sigma,S,\Omega) \leq \bar{h}(\vec{v},\kappa,s,\sigma,S,\Omega) \leq C_0\cdot\frac{\kappa}{\mu}.$$

Proof. Set $m = 1$ in Theorem 5.5 and let $C_0 = \phi(1, \mathcal{D}) = C_1(\mathcal{D})/C_2(1, \mathcal{D})$. The result follows from Theorem 5.5 if $\kappa \geq \kappa_0 = \phi(1, \mathcal{D})^{-\frac{1}{2}}$. $\square$

6. Homogenization of the Lawrence–Doniach model. For completeness, we conclude this paper by stating a result on the homogenization of solutions of the Lawrence–Doniach system (in the gauge (1.5)) to solutions of a three-dimensional anisotropic Ginzburg–Landau system, which follows from the maximum principle (Theorem 2.4) and results proved by Chapman, Du, and Gunzburger in [4] under the assumption that solutions of the Lawrence–Doniach system satisfy $|\psi_n| \leq 1$ for all n.

The three-dimensional anisotropic Ginzburg–Landau energy (in nondimensionalized form) is given by

$$
\begin{aligned}
\mathcal{G}_{AGL}(\psi, \vec{A}) = &\int_{\mathcal{D}} \frac{1}{2}\left(|\psi|^2 - 1\right)^2 d\vec{x} + \int_{\mathbb{R}^3} \left|\nabla \times (\vec{A} - h\vec{a})\right|^2 d\vec{x} \\
&+ \int_{\mathcal{D}} \left(\left| \left(\frac{\imath}{\kappa} \mathbf{grad} + \mathbf{A} \right) \psi \right|^2 + \frac{1}{\gamma} \left| \left(\frac{\imath}{\kappa} \frac{\partial}{\partial x_3} + A^3 \right) \psi \right|^2 \right) d\vec{x},
\end{aligned}
\tag{6.1}
$$

where κ is the Ginzburg–Landau constant and γ is an anisotropy constant. Here, the order parameter ψ is defined in $\mathcal{D} = \Omega \times (0, S)$ and $(\psi, \vec{A}) \in H^1(\mathcal{D}) \times E$, where E is defined as in (1.2). We remark that the anisotropic Ginzburg–Landau energy $\mathcal{G}_{AGL}$ is invariant under the gauge transformation

$$
(\psi, \vec{A}) \longrightarrow (\xi, \vec{Q}),
$$

where

$$
\begin{cases}
\xi(\vec{x}) = \psi(\vec{x}) e^{\imath\kappa g(\vec{x})}, \\
\vec{Q} = \vec{A} + \nabla g,
\end{cases}
$$

which maps $H^1(\mathcal{D}) \times E$ to itself whenever $g \in H^2_{loc}(\mathbb{R}^3)$. By choosing g appropriately, one obtains $(\xi, \vec{Q}) \in H^1(\mathcal{D}) \times K$, where K is defined as in (1.5), and it follows that minimizers of $\mathcal{G}_{AGL}$ in $H^1(\mathcal{D}) \times E$ exist and are gauge-equivalent to some $(\psi, \vec{A})$ in $H^1(\mathcal{D}) \times K$. Moreover, minimizers are weak solutions of the Euler–Lagrange equations (called the anisotropic equations)

$$
\begin{cases}
\left(\frac{\imath}{\kappa} \mathbf{grad} + \mathbf{A} \right)^2 \psi + \frac{1}{\gamma} \left(\frac{\imath}{\kappa} \frac{\partial}{\partial x_3} + A^3 \right)^2 \psi + \left(|\psi|^2 - 1\right)\psi = 0 \quad \text{in } \mathcal{D}, \\
\nabla \times \left(\nabla \times \vec{A} \right) = (J_1, J_2, J_3) \quad \text{in } \mathbb{R}^3, \\
\left(\frac{\imath}{\kappa} \nabla \psi + \vec{A}\psi \right) \cdot \hat{n} = 0 \quad \text{on } \partial\mathcal{D}, \\
\nabla \times \left(\vec{A} - h\vec{a} \right) \in L^2(\mathbb{R}^3; \mathbb{R}^3),
\end{cases}
\tag{6.2}
$$

where

$$
J_i = \frac{1}{\gamma_i} \left[\frac{\imath}{2\kappa} \left(\psi \frac{\partial \psi^*}{\partial x_i} - c.c. \right) - |\psi|^2 A^i \right] \chi_{\mathcal{D}}
$$

for $i = 1, 2, 3$, and where $\gamma_1 = \gamma_2 = 1$ and $\gamma_3 = \gamma$. Here, $\hat{n}$ is the outward normal to $\partial\mathcal{D}$. More precisely, a minimizer of $\mathcal{G}_{AGL}$ in $H^1(\mathcal{D}) \times K$ satisfies

(6.3)

$$\begin{aligned}
&\int_{\mathcal{D}} \Bigg[(|\psi|^2 - 1)\psi\varphi^* + \left(\frac{\imath}{\kappa}\mathbf{grad}\ \psi + \mathbf{A}\psi\right) \cdot \left(\frac{\imath}{\kappa}\mathbf{grad}\ \varphi + \mathbf{A}\varphi\right)^* \\
&\qquad + \frac{1}{\gamma}\left(\frac{\imath}{\kappa}\frac{\partial\psi}{\partial x_3} + A^3\psi\right)\left(\frac{\imath}{\kappa}\frac{\partial\varphi}{\partial x_3} + A^3\varphi\right)^*\Bigg] d\vec{x} = 0, \\
&2\int_{\mathbb{R}^3} \nabla\times\left(\vec{A} - h\vec{a}\right) \cdot \nabla\times\vec{B}\, d\vec{x} + \int_{\mathbb{R}^3} \chi_{\mathcal{D}}\left[\frac{\imath}{\kappa}(\psi^*\,\mathbf{grad}\ \psi - c.c.) + 2\mathbf{A}\,|\psi|^2\right] \cdot \mathbf{B}\, d\mathbf{x} \\
&\qquad + \int_{\mathbb{R}^3} \chi_{\mathcal{D}}\,\frac{1}{\gamma}\left[\frac{\imath}{\kappa}\left(\psi^*\frac{\partial\psi}{\partial x_3} - c.c.\right) + 2A^3\,|\psi|^2\right] B^3\, d\vec{x} = 0
\end{aligned}$$

for any $\varphi \in \mathcal{H}^1(\mathcal{D})$ and $\vec{B} \in H^1(\mathbb{R}^3;\mathbb{R}^3)$ with bounded support.

As in the case of the Lawrence–Doniach system, one can prove that weak solutions satisfy $|\psi| \le 1$ almost everywhere in $\mathcal{D}$ by choosing an appropriate test function in (6.3).

Assume that the parameters σ, κ, s, and γ are related by

$$\frac{1}{\gamma} = \sigma\kappa^2 s^2. \tag{6.4}$$

(See [4] and [13] for details on the physical meaning of these relationships.) Briefly, the above equation allows one to define $m_\perp$, $\lambda_\perp$, and $\xi_\perp$, the mass of superconducting charge carriers, penetration depth, and coherence length in directions perpendicular to the layers in the dimensionalized Lawrence–Doniach model by

$$\gamma = \frac{m_\perp}{m_\|} = \left(\frac{\lambda_\perp}{\lambda_\|}\right)^2 = \left(\frac{\xi_\perp}{\xi_\|}\right)^2$$

(where $m_\|, \lambda_\|$, and $\xi_\|$ are defined to be their values within the layers for the Lawrence–Doniach model) in such a way that

$$\sigma = \left(\frac{\xi_\perp}{d}\right)^2,$$

and it can be shown that $s \to 0$ in the nondimensionalized Lawrence–Doniach model corresponds to $d/\xi_\perp \to 0$ in the dimensionalized Lawrence–Doniach model. (See [4].) The homogenization of the Lawrence–Doniach model is achieved by fixing the domain $\mathcal{D} = \Omega \times [0, S]$ and letting N go to ∞. Since $sN = S$, there is a set $\mathcal{A}$ of the available values for s, i.e., $\mathcal{A} = \left\{\frac{S}{N} : N = 1, 2, \dots\right\}$. Thus, letting $s \to 0^+$ means that $s \in \mathcal{A}$ and $s \to 0$ by letting $N \to \infty$.

Now let ξ^s denote $\{\xi_n\}_{n=0}^N$ if $s \in \mathcal{A}$ and $\{\xi_n\}_{n=0}^N \in \left[\mathcal{H}^1(\Omega)\right]^{N+1}$. Define $I_s\xi^s$ to be the linear interpolant of ξ^s in the x_3 direction in $\mathcal{D}$ with respect to the superconducting layers $\{\Omega_n\}_{n=0}^N$, i.e.,

$$\left(I_s\xi^s\right)(\mathbf{x}, x_3) = \left(1 - \frac{x_3 - ns}{s}\right)\xi_n(\mathbf{x}) + \frac{x_3 - ns}{s}\,\xi_{n+1}(\mathbf{x}) \tag{6.5}$$

if $(\mathbf{x}, x_3) \in \Omega \times [ns, (n+1)s]$ for all $n = 0, 1, \ldots, N-1$. Thus, we have

$$\begin{aligned} \nabla\big(I_s\xi^s\big) &= \left(\mathbf{grad}\,\big(I_s\xi^s\big), \frac{\partial I_s\xi^s}{\partial x_3}\right) \\ &= \left(\left(1 - \frac{x_3 - ns}{s}\right)\mathbf{grad}\,\xi_n + \frac{x_3 - ns}{s}\mathbf{grad}\,\xi_{n+1}\,, \frac{\xi_{n+1} - \xi_n}{s}\right). \end{aligned} \tag{6.6}$$

It is easy to show that if $\xi^s \in \big[\mathcal{H}^1(\Omega)\big]^{N+1}$, then $I_s\xi^s \in \mathcal{H}^1(\mathcal{D})$.

From Theorem 2.4 of this paper and Theorem 5.1, Lemma 5.5, and Corollary 5.6 of [4] (using the Coulomb gauge (1.5)), we have the following.

THEOREM 6.1. *Let* $\big(\psi^s, \vec{A}^s\big) \in \big[\mathcal{H}^1(\Omega)\big]^{N+1} \times K$ *denote a minimizer of* $\mathcal{G}^s_{LD}$ *in* $\big[\mathcal{H}^1(\Omega)\big]^{N+1} \times E$ *for each* $s \in \mathcal{A}$. *If* (6.4) *is satisfied, then as* $s \to 0^+$, $\big\{\big(I_s\psi^s, \vec{A}^s\big)\big\}_{s\in\mathcal{A}}$ *forms a minimizing sequence of* $\mathcal{G}_{AGL}$ *in* $H^1(\mathcal{D}) \times E$. *Moreover, if* l *and* l_s *denote the minimum values of* $\mathcal{G}_{AGL}$ *and* $\mathcal{G}^s_{LD}$, *respectively, i.e.,*

$$\begin{aligned} l &= \min\big\{\mathcal{G}_{AGL}\big(\xi, \vec{Q}\big) : \big(\xi, \vec{Q}\big) \in \mathcal{H}^1(\mathcal{D}) \times E\big\}, \\ l_s &= \min\big\{\mathcal{G}^s_{LD}\big(\xi^s, \vec{Q}\big) : \big(\xi^s, \vec{Q}\big) \in [\mathcal{H}^1(\Omega)]^{N+1} \times E\big\} \end{aligned}$$

for $s \in \mathcal{A}$, *we have* $l = \lim_{s\to 0^+} l_s = \lim_{s\to 0^+} \mathcal{G}_{AGL}\big(I_s\psi^s, \vec{A}^s\big)$.

REFERENCES

[1] S. Alama, A. J. Berlinsky, and L. Bronsard, *Minimizers of the Lawrence–Doniach energy in the small-coupling limit: Finite width samples in a parallel field*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 19 (2002), pp. 281–312.

[2] S. Alama, L. Bronsard, and A. J. Berlinsky, *Periodic vortex lattices for the Lawrence–Doniach model of layered superconductors in a parallel field*, Commun. Contemp. Math., 3 (2001), pp. 457–494.

[3] G. Burns, *High-Temperature Superconductivity, An Introduction*, Academic Press, New York, 1992.

[4] S. J. Chapman, Q. Du, and M. D. Gunzburger, *On the Lawrence–Doniach and anisotropic Ginzburg–Landau models for layered superconductors*, SIAM J. Appl. Math., 55 (1995), pp. 156–174.

[5] B. E. J. Dahlberg and C. E. Kenig, *Hardy spaces and the Neumann problem in L^p for Laplace's equation in Lipschitz domains*, Ann. of Math. (2), 125 (1987), pp. 437–465.

[6] Q. Du and P. Gray, *Numerical algorithms of the Lawrence–Doniach model for layered superconductors and their parallel implementation*, SIAM J. Sci. Comput., 20 (1999), pp. 2122–2139.

[7] Q. Du, M. D. Gunzburger, and J. S. Peterson, *Analysis and approximation of the Ginzburg–Landau model of superconductivity*, SIAM Rev., 34 (1992), pp. 54–81.

[8] E. B. Fabes, M. Jodeit, Jr., and N. M. Riviere, *Potential techniques for boundary value problems on C^1 domains*, Acta Math., 141 (1978), pp. 165–186.

[9] T. Giorgi and D. Phillips, *The breakdown of superconductivity due to strong fields for the Ginzburg–Landau model*, SIAM J. Math. Anal., 30 (1999), pp. 341–359.

[10] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, New York, 1983.

[11] Y. Iye, *How anisotropic are the cuprate high T C superconductors?*, Comments Cond. Mat. Phys., 16 (1992), pp. 89–111.

[12] C. Kenig, *Harmonic Analysis Techniques for Second Order Elliptic Boundary Value Problems*, AMS Reg. Conf. Ser. Math. 83, AMS, Providence, RI, 1994.

[13] R. Klemm, A. Luther, and M. Beasley, *Theory of the upper critical field in layered superconductors*, Phys. Rev. B, 12 (1975), pp. 877–891.

[14] O. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1969.

[15] W. Lawrence and S. Doniach, *Theory of layer structure superconductors*, in Proceedings of the 12th International Conference on Low Temperature Physics, Academic Press of Japan, Kyoto, 1971, pp. 361–362.

[16] D. Mitrea, *Sharp L^p-Hodge decompositions for Lipschitz domains in R^2*, Adv. Differential Equations, 7 (2002), pp. 343–364.

[17] D. Saint-James and P. G. de Gennes, *Onset of superconductivity in decreasing fields*, Phys. Lett., 7 (1963), pp. 306–308.

[18] M. Tinkham, *Introduction to Superconductivity*, 2nd ed., McGraw–Hill, New York, 1996.

[19] G. Verchota, *Layer potentials and regularity for the Dirichlet problem for Laplace's equation in Lipschitz domains*, J. Funct. Anal., 59 (1984), pp. 572–611.

[20] Y. Yang, *Existence, regularity, and asymptotic behavior of the solutions to the Ginzburg–Landau equations on* $\mathbb{R}^3$, Comm. Math. Phys., 123 (1989), pp. 147–161.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 941–981

CURRENT COUPLING OF DRIFT-DIFFUSION MODELS AND SCHRÖDINGER–POISSON SYSTEMS: DISSIPATIVE HYBRID MODELS*

M. BARO†, H. NEIDHARDT†, AND J. REHBERG†

Dedicated to H. Gajewski on the occasion of his 65th birthday

Abstract. A one-dimensional coupled drift-diffusion dissipative Schrödinger model (hybrid model) is mathematically analyzed. The device domain is split into two parts: one in which the transport is well described by the drift-diffusion equations (classical zone) and another in which a quantum description via dissipative Schrödinger equations (quantum zone) is used. Both system are coupled such that the continuity of the current densities is guaranteed. The electrostatic potential is self-consistently determined by Poisson's equation on the whole device domain. We show that the hybrid model is well posed, and we prove existence of solutions and show their uniform boundedness, provided the distribution functions satisfy a so-called balance condition. The current densities are different from zero in the nonequilibrium case and are uniformly bounded.

Key words. quantum–classical coupling, hybrid models, drift-diffusion models, dissipative Schrödinger systems, Poisson equation, current coupling, semiconductors

AMS subject classifications. 47B44, 34L40, 47E05, 35J05

DOI. 10.1137/040611690

1. Introduction.

A basic model for carrier transport in semiconductors was created in 1950 by van Roosbroeck [51]. It describes the transport of electrons and holes by drift and diffusion processes in a self-consistent electrical field. This so-called drift-diffusion model was first used by Gummel [27] to calculate diodes. Since that time, drift-diffusion models have been intensively studied, and there exists extensive literature on them; see [23, 44, 52] and references therein. However, modern semiconductor devices inherently employ quantum effects in their operations, such as tunneling [15, 18, 54], which are well described by stationary or transient Wigner– or Schrödinger–Poisson systems [1, 9, 10, 12, 19, 20, 29, 30, 31, 36, 37, 38, 39, 41, 42, 46, 47, 48, 49, 53]. Unfortunately, the numerical treatment of a Wigner– or Schrödinger–Poisson system is fairly expensive compared to classical models, such as the drift-diffusion model. However, for several devices, such as resonant tunneling diodes [13, 14, 22, 45], the quantum effects occur only in some small spatial parts, while other parts admit a quite reasonable description by approved "classical models," such as drift-diffusion models, etc. Thus one looks for a model that combines a quantum description in parts where it is necessary with a classical description in other parts. The aim is to obtain a model which allows an effective and fast numerical treatment but describes the transport of electrons and holes in the semiconductor device with sufficient accuracy. Models of that type are usually called hybrid models; cf. [2, 6, 7, 16]. In [2, 16] coupled Schrödinger drift-diffusion models are used to simulate the current voltage

*Received by the editors July 15, 2004; accepted for publication (in revised form) March 23, 2005; published electronically December 6, 2005. The financial support of DAAD grant D/0205756 is gratefully acknowledged.

http://www.siam.org/journals/sima/37-3/61169.html

†Weierstrass-Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany (baro@wias-berlin.de, neidhard@wias-berlin.de, rehberg@wias-berlin.de). The research of the first author was supported by the DFG Research Center Matheon "Mathematics for Key Technologies" in Berlin. The research of the second author was supported by DFG grant RE 1480/2.

characteristic of a resonant tunneling diode. In the following we are interested in an analytical treatment of such models under quite general assumptions, but in a dissipative approximation.

In particular, we consider a one-dimensional stationary hybrid model which consists of a drift-diffusion model in the so-called "classical zone" and a dissipative Schrödinger model in the "quantum zone." Both systems are coupled by the conditions that the Fermi energies of the quantum zone are given by the electrochemical potentials of the classical zone, and the current densities are continuous at the interface points. Moreover, the electrostatic potential is determined self-consistently by a Poisson equation on the whole device domain.

The hybrid model approach, however, evokes several problems. In fact, if one is interested in a current density which is continuous over the whole device, then one has to consider an open quantum system. Indeed, for Schrödinger operators with self-adjoint boundary conditions, the current density is always zero since such operators commute with the complex conjugation. Thus, a continuous nontrivial net current flow through the interface between quantum and classical zones is impossible in this case. Consequently, hybrid models enforce at least non–self-adjoint boundary conditions for the Schrödinger operator to describe the particles in the quantum zone. Such models are introduced in [33, 34, 35]. Further, a nontrivial current density arises in the quantum zone only if the statistical behavior of the quantum system is described by a density matrix, which is different from those of the thermodynamical equilibrium. Hence, one has to find suitable nonequilibrium density matrices.

In more detail, we divide the one-dimensional device domain $\Omega = (a_0, b_0) \subseteq \mathbb{R}$ into a quantum zone $\Omega_q = (a, b)$, $a_0 < a < b < b_0$, and a classical region $\Omega_c = \Omega \setminus \Omega_q$. In the classical region the densities of electrons U^- and holes U^+, as well as the current densities for electrons J^- and holes J^+, respectively, are determined by means of the stationary drift-diffusion equations without generation or recombination [43] (temperature, elementary charge, and Boltzmann's constant scaled to one):

$$U^\pm(x) := N_0(x)^\pm \exp\Big(- V^\pm(x) \pm \phi^\pm(x)\Big) \qquad \text{(density relations)}, \tag{1.1}$$

$$J^\pm(x) := \mp\mu^\pm(x) U^\pm(x) \frac{d}{dx}\phi^\pm(x) \qquad \text{(current relations)}, \tag{1.2}$$

$$\frac{d}{dx} J^\pm = 0 \qquad \text{(continuity equations)}, \tag{1.3}$$

$x \in \Omega_c$. Here and in what follows, the superscript "+" refers to hole quantities and "−" to electron quantities. $N_0^\pm$ denotes the densities of states, $\mu^\pm$ the carrier mobilities, and $\phi^\pm$ the electrochemical potentials. The potentials $V^\pm$ contain the electrostatic potential φ and the band edge-offset potentials $V_h^\pm$, i.e.,

$$V^\pm := V_h^\pm \pm \varphi. \tag{1.4}$$

At the device boundary we impose inhomogeneous Dirichlet boundary conditions on the system (1.1)–(1.3), i.e.,

$$\phi^\pm(a_0) = \phi^\pm_{a_0} \quad \text{and} \quad \phi^\pm(b_0) = \phi^\pm_{b_0}. \tag{1.5}$$

To have a mathematically meaningful description, the system (1.1)–(1.3) requires boundary conditions at the end points a and b of the classical zones (a_0, a) and (b, b_0), too. However, for the hybrid model, a and b are not boundary points but interface points at which the coupling of the drift-diffusion system with the quantum subsystem is realized. Hence, the coupling conditions have to replace the boundary conditions at these interface points. We will develop these interface conditions later in the text.

In the quantum zone a dissipative Schrödinger system is adopted (cf. [3, 33]), which is derived from a quantum transmitting Schrödinger system; see [4, 8] and section B.1 in Appendix B. A dissipative Schrödinger model consists of two dissipative Schrödinger-type operators $H^\pm$ (cf. [34, 35] and section A.1 in Appendix A), arising from the differential expressions

$$H^\pm f = \left(-\frac{1}{2}\frac{d}{dx}\frac{1}{m^\pm}\frac{d}{dx} + V^\pm\right) f, \tag{1.6}$$

where $m^\pm$ denotes the effective mass of the particle under consideration and $V^\pm$ contain the band edge-offsets and the electrostatic potential; see (1.4). Supplemented by the boundary conditions

$$\frac{1}{2m^\pm(a)} f'(a) = -\varkappa_a^\pm f(a) \quad \text{and} \quad \frac{1}{2m^\pm(b)} f'(b) = \varkappa_b^\pm f(b), \tag{1.7}$$

$H^\pm$ are maximal dissipative, and completely non–self-adjoint operators on the Hilbert space $L^2(\Omega_q)$ if $\mathrm{Im}(\varkappa_a^\pm), \mathrm{Im}(\varkappa_b^\pm) > 0$. The coupling constants $\varkappa_a^\pm$ and $\varkappa_b^\pm$ are given by means of the potentials $V^\pm$ at the quantum–classical interface boundaries a and b by

$$\varkappa_a^\pm := i\sqrt{\frac{s^\pm - V^\pm(a)}{2m_a^\pm}} \quad \text{and} \quad \varkappa_b^\pm := i\sqrt{\frac{s^\pm - V^\pm(b)}{2m_b^\pm}}, \tag{1.8}$$

where i denotes the imaginary unit, i.e., $i^2 = -1$, and

$$s^\pm := V_{max}^\pm + \delta_0^\pm, \quad \delta_0^\pm > 0, \tag{1.9}$$

with

$$V_{max}^\pm := \max\{V^\pm(a), V^\pm(b)\}, \tag{1.10}$$

where $\delta_0^\pm$ are given positive constants. Note that $\mathrm{Im}(\varkappa_a^\pm) > 0$ and $\mathrm{Im}(\varkappa_b^\pm) > 0$. The operators $H^\pm$ are maximal dissipative, and the multiplicity of their minimal self-adjoint dilations is 2; cf. [35]. Moreover, the spectrum of minimal self-adjoint dilations coincides with the real line and is purely absolutely continuous. Thus, there exist two generalized eigenfunctions, denoted by $\psi_a^\pm(\lambda, x), \psi_b^\pm(\lambda, x)$, for each dilation operator; see [33] and section A.2 in Appendix A. The particle densities $u^\pm$ and current densities $j^\pm$ are then given by means of the generalized eigenfunctions $\psi_a^\pm(\lambda, x)$, $\psi_b^\pm(\lambda, x)$, assuming Boltzmann distribution, by

$$u^\pm(x) = \sum_{\nu=a,b} \int_{\Lambda^\pm} d\lambda\, n_0^\pm \exp(-\lambda \pm \epsilon_\nu) \, |\psi_\nu(\lambda, x)|^2, \quad x \in \Omega_q, \tag{1.11}$$

$$j^\pm = \sum_{\nu=a,b} \int_{\Lambda^\pm} d\lambda\, n_0^\pm \exp(-\lambda \pm \epsilon_\nu) \, \mathrm{Im}\left(\frac{1}{m(x)} \frac{\partial \psi_\nu(\lambda, x)}{\partial x} \overline{\psi_\nu(\lambda, x)}\right) \tag{1.12}$$

(see [33, 5]), where $n_0^\pm$ are the integrated density of states and

$$\Lambda^\pm := [V_{max}^\pm, V_{max}^\pm + \delta^\pm), \quad 0 < \delta_0^\pm < \delta^\pm \leq \infty. \tag{1.13}$$

The real parameters $\mathfrak{A} := \{\delta_0^\pm, \delta^\pm\}$, $0 < \delta_0^\pm < \delta^\pm \leq \infty$, are called the approximation parameters; see section B.2 in Appendix B.

The construction of the Schrödinger model in the quantum zone is to some extent artificial, since the physical interpretation of the generalized eigenfunctions ψ_a, ψ_b corresponding to the self-adjoint dilations of $H^\pm$ is not a priori clear. However, as pointed out in [4], the dissipative Schrödinger model can be seen as an approximation of the usual Schrödinger scattering model considered elsewhere (see, e.g., [8, 22]), with approximation parameters $\mathfrak{A}$. We will briefly outline this feature in section B.2 of Appendix B.

The Fermi levels $\epsilon_a^\pm$, $\epsilon_b^\pm$ in (1.11), (1.12) are determined by the drift-diffusion model by

$$\phi^\pm(a) = \epsilon_a^\pm \qquad \text{and} \qquad \phi^\pm(b) = \epsilon_b^\pm. \tag{1.14}$$

Thus, the conditions (1.14) couple the drift-diffusion equations (1.1)–(1.3) and the dissipative Schrödinger operators (1.6)–(1.12). This coupling will be called the Fermi coupling. Moreover, we impose the continuity of the drift-diffusion currents $J^\pm$ and the quantum current $j^\pm$, i.e.,

$$J^\pm(a) = j^\pm = J^\pm(b). \tag{1.15}$$

This condition is necessary in order to obtain a physically meaningful model.

In order to have a meaningful description of the semiconductor device, the electrostatic potential φ has to be computed self-consistently by a Poisson equation, i.e.,

$$-\frac{d}{dx}\varepsilon\frac{d}{dx}\varphi(x) = C(x) + \mathcal{U}^+(x) - \mathcal{U}^-(x), \quad x \in \Omega, \tag{1.16}$$

with boundary conditions

$$\varphi(a_0) = \varphi_{a_0} \qquad \text{and} \qquad \varphi(b_0) = \varphi_{b_0}, \tag{1.17}$$

where the carrier densities $\mathcal{U}^\pm$ are given by

$$\mathcal{U}^\pm(x) := \begin{cases} U^\pm(x) & \text{for } x \in \Omega_c, \\ u^\pm(x) & \text{for } x \in \Omega_q, \end{cases} \tag{1.18}$$

and ε, C denote the dielectric permittivity and the doping profile, respectively.

The coupled dissipative Schrödinger (1.6)–(1.12) and drift-diffusion model (1.1)–(1.3) with the coupling conditions (1.14)–(1.18) is called the dissipative hybrid model.

The aim of this paper is to show that the proposed dissipative hybrid model is well posed and admits a solution under natural assumptions for any choice of the approximation parameters $\mathfrak{A}$. The paper is organized as follows. In section 2 we investigate the stationary drift-diffusion system on the disconnected set $\Omega_c = (a_0, a) \cup (b, b_0)$, provided the current densities are given and equal on the different intervals (a_0, a) and (b, b_0). This leads to certain restrictions on the current densities; cf. Lemma 2.2. Section 3 is devoted to the rigorous setup of the dissipative Schrödinger system used in the quantum zone. The dissipative hybrid model is defined in section 4. In section 4.1 the so-called Fermi coupling is explained. Using the Fermi coupling, we show in section 4.2 that the stationary drift-diffusion and the dissipative Schrödinger models admit a current coupling. This result is in fact nontrivial and is based on Proposition 4.1. Using the results of section 4.2, in section 4.3 we rigorously introduce the dissipative hybrid model. Finally, in section 4.4 we consider the coupling with the Poisson equation. The problem of finding a solution to the dissipative hybrid model is reformulated in section 4.4 as a fixed point problem. In section 5 we show that the fixed point problem admits a solution. The existence proof is based on the Leray–Schauder fixed point theorem. Uniqueness is not shown and not expected for physical reasons; cf. [32, 50]. However, it turns out that all solutions are uniformly bounded by a bound that is determined by the data of the problem but independent of the choice of the approximation parameters $\mathfrak{A}$. Moreover, the current densities are nontrivial in the nonequilibrium case. We sum up with some comments in section 6. In section A we give an introduction to dissipative Schrödinger systems and prove some continuity results for the carrier and current density operators. The derivation of the dissipative Schrödinger model from the usual Schrödinger scattering model is exposed in section B.

Notation. By $L^p(\mathcal{O}, X, \mathfrak{m})$, $1 \leq p < \infty$, we denote the space of $\mathfrak{m}$-measurable and p-integrable functions over Borel sets $\mathcal{O} \subseteq \mathbb{R}$ with values in a Banach space X. By $L^\infty(\mathcal{O}, X, \mathfrak{m})$ the space of essentially bounded functions is denoted. If $\mathfrak{m}$ is the Lebesgue measure, then we write $L^p(\mathcal{O}) = L^p(\mathcal{O}, \mathbb{C}, \mathfrak{m})$ and $L^p_{\mathbb{R}}(\mathcal{O}) := L^p(\mathcal{O}, \mathbb{R}, \mathfrak{m})$, $1 \leq p \leq \infty$. For closed sets $\mathcal{O} \subseteq \mathbb{R}$ we denote by $C(\mathcal{O})$ and $C_{\mathbb{R}}(\mathcal{O})$ the spaces of continuous complex- or real-valued functions, respectively, on $\mathcal{O}$ equipped with the supremum norm.

The norm of a Banach or Hilbert space X is indicated by $\|\cdot\|_X$, or simply by $\|\cdot\|$; the scalar product of a Hilbert space X by $(\cdot,\cdot)_X$, or simply by $(\cdot,\cdot)$, where the first argument is the linear one. The dual space is indicated by X^*. By $\mathcal{B}(X, Y)$ the space of all linear bounded operators from the Banach space X to the Banach space Y is denoted with norm $\|\cdot\|_{\mathcal{B}(X,Y)}$. If $X = Y$, then $\mathcal{B}(X, X) = \mathcal{B}(X)$ and $\|\cdot\|_{\mathcal{B}(X,Y)} = \|\cdot\|_{\mathcal{B}(X)}$. If X is a Hilbert space, then $\mathcal{B}_1(X)$ denotes the space of trace class operators. For a densely defined linear operator $A : X \longrightarrow Y$ we denote by A^*, $\mathrm{spec}(A)$, and $\mathrm{res}(A)$ its adjoint, spectrum, and resolvent set, respectively.

Furthermore, for $\mathcal{O} = (a_0, b_0)$ or $\mathcal{O} = (a, b)$ we denote by $W^{1,2}(\mathcal{O})$ the usual Sobolev spaces of complex-valued functions on $\mathcal{O}$. The subspace of elements with homogeneous Dirichlet boundary conditions at the end points of the interval $\mathcal{O} \subseteq \mathbb{R}$ is denoted by $\overset{o}{W}{}^{1,2}(\mathcal{O})$. Its dual with respect to the L^2-pairing is denoted by $W^{-1,2}(\mathcal{O}) = (\overset{o}{W}{}^{1,2}(\mathcal{O}))^*$. If we have in mind only real-valued functions, then we write $W^{1,2}_{\mathbb{R}}(\mathcal{O})$ and $\overset{o}{W}{}^{1,2}_{\mathbb{R}}(\mathcal{O})$.

Moreover, the superscripts "+" and "−" always indicate quantities related to holes and electrons, respectively.

Throughout this paper φ will always denote the electrostatic potential. If needed, we will indicate the dependence of any quantity A on the electrostatic potential φ, or the current density J and the electrostatic potential φ, by writing $A[\varphi]$ or $A[J, \varphi]$, respectively.

2. Classical zone. In this section we consider the stationary drift-diffusion equations (1.1)–(1.3) on the disconnected set Ω_c with boundary conditions (1.5). The boundary conditions at a and b are replaced by the conditions that (i) the current densities at a and b are equal and (ii) these current densities are given. We show in this section that the system is well posed and admits solutions $\phi^\pm$, provided the given current densities are located in some interval around zero, which depends on the fixed electrostatic potential φ. Later on, the given current densities will be the quantum current densities for a fixed potential φ, which is finally determined self-consistently by the Poisson equation.

We assume that the carrier and current densities $U^{\pm}$ and $J^{\pm}$ are given by (1.1) and (1.2), respectively. We make the following assumptions.

ASSUMPTION 2.1.

(C.1) *The effective carrier mobilities $\mu^{\pm}$ are strictly positive and constant on Ω_c.*

(C.2) *The effective density of states $N_0^{\pm}$ are strictly positive constants on Ω_c.*

(C.3) *The boundary values ϕ_{a_0} and ϕ_{b_0} are given constants from $\mathbb{R}$.*

(C.4) *The band-edge offsets $V_h^{\pm}$ are real and continuous functions, i.e., $V_h^{\pm} \in C_{\mathbb{R}}(\overline{\Omega_c})$.*

(C.5) *Generation and recombination are absent.*

Moreover, we will assume throughout this section that the electrostatic potential φ is given and obeys $\varphi \in C_{\mathbb{R}}(\Omega)$. By Assumption (C.4) we get that the potentials $V^{\pm}$ defined by (1.4) satisfy $V^{\pm} \in C_{\mathbb{R}}(\overline{\Omega_c})$.

By (C.5) we obtain the continuity equations (1.3), which imply that the current density $J^{\pm}(x)$ is constant on each subinterval of Ω_c, i.e.,

$$J^{\pm}(x) = J^{\pm}(a), \quad x \in (a_0, a), \quad \text{and} \quad J^{\pm}(x) = J^{\pm}(b), \quad x \in (b, b_0). \tag{2.1}$$

Thus, we deduce

$$J^{\pm}(a) = \mu^{\pm} N_0^{\pm} \frac{e^{-\phi^{\pm}_{a_0}} - e^{-\phi^{\pm}(a)}}{\int_{a_0}^{a} dy\, e^{V^{\pm}(y)}} \tag{2.2}$$

and

$$J^{\pm}(b) = \mu^{\pm} N_0^{\pm} \frac{e^{-\phi^{\pm}(b)} - e^{-\phi^{\pm}_{b_0}}}{\int_{b}^{b_0} dy\, e^{V^{\pm}(y)}}. \tag{2.3}$$

LEMMA 2.2. *Let the electrostatic potential $\varphi \in C_{\mathbb{R}}(\Delta)$ and the current densities $J^{\pm}$ be given. There are solutions $\phi^{\pm} \in C^1_{\mathbb{R}}(\Omega_c)$ of the drift-diffusion equations* (1.1)–(1.3) *satisfying the boundary conditions* (1.5) *and the conditions $J^{\pm}(a) = J^{\pm}(b) = J^{\pm}$ if and only if $J^{\pm} \in (J^{\pm}_{min}, J^{\pm}_{max})$, where*

$$J^{\pm}_{min} := -\mu^{\pm} N_0^{\pm} \frac{e^{\pm\phi^{\pm}_{b_0}}}{\int_{b}^{b_0} dy\, e^{V^{\pm}(y)}}, \quad J^{\pm}_{max} := \mu^{\pm} N_0^{\pm} \frac{e^{\pm\phi^{\pm}_{a_0}}}{\int_{a_0}^{a} dy\, e^{V^{\pm}(y)}}, \tag{2.4}$$

with $V^{\pm}$ given by (1.4). *Moreover, these solutions $\phi^{\pm}$ are unique.*

Proof. Let us first assume that $\phi^{\pm}$ are solutions of the drift-diffusion equations such that $J^{\pm}(a) = J^{\pm}(b) = J^{\pm}$ holds. From (2.2) and (2.3), one gets

$$J^{\pm}(a) < \mu^{\pm} N_0^{\pm} \frac{e^{\pm\phi^{\pm}_{a_0}}}{\int_{a_0}^{a} dy\, e^{V^{\pm}(y)}} = J^{\pm}_{max} \tag{2.5}$$

and

$$J^{\pm}(b) > -\mu^{\pm} N_0^{\pm} \frac{e^{\pm\phi^{\pm}_{b_0}}}{\int_{b}^{b_0} dy\, e^{V^{\pm}(y)}} = J^{\pm}_{min}. \tag{2.6}$$

Thus, $J^{\pm} \in (J^{\pm}_{min}, J^{\pm}_{max})$.

Conversely, if $J^{\pm} \in (J^{\pm}_{min}, J^{\pm}_{max})$, then we can define

$$\phi^-(x) := \begin{cases} -\ln\left(e^{-\phi^-_{a_0}} - \frac{J^-}{\mu^- N^-_0}\int_{a_0}^{x} dy\, e^{V^-(y)}\right) & \text{for } x \in (a_0, a), \\ -\ln\left(e^{-\phi^-_{b_0}} + \frac{J^-}{\mu^- N^-_0}\int_{x}^{b_0} dy\, e^{V^-(y)}\right) & \text{for } x \in (b, b_0), \end{cases} \tag{2.7}$$

and similarly

$$\phi^+(x) := \begin{cases} \ln\left(e^{\phi^+_{a_0}} - \frac{J^+}{\mu^+ N^+_0}\int_{a_0}^{x} dy\, e^{V^+(y)}\right) & \text{for } x \in (a_0, a), \\ \ln\left(e^{\phi^+_{b_0}} + \frac{J^+}{\mu^+ N^+_0}\int_{x}^{b_0} dy\, e^{V^+(y)}\right) & \text{for } x \in (b, b_0). \end{cases} \tag{2.8}$$

A straightforward computation shows that $\phi^{\pm}$ defined by (2.7) and (2.8) are indeed solutions of the drift-diffusion equations (1.1)–(1.3) and satisfy the boundary conditions (1.5). The uniqueness of the solutions follows from the fact that (1.3) is a second order equation for $e^{\phi^{\pm}}$, where the boundary values at $\phi^{\pm}(a_0)$, $\phi^{\pm}(b_0)$ are fixed by (1.5), and $\frac{d}{dx}\phi^{\pm}(a_0)$, $\frac{d}{dx}\phi^{\pm}(b_0)$ are fixed by the conditions $J^{\pm}(a) = J^{\pm}(b) = J^{\pm}$. ∎

It is convenient to introduce the sets

$$\mathcal{E}^- := \left\{ (J^-, \varphi) \in \mathbb{R} \times C_{\mathbb{R}}(\Omega) : \begin{array}{l} 0 < e^{-\phi^-_{a_0}} - \frac{J^-}{\mu^- N^-_0}\int_{a_0}^{a} dx\, e^{V^-(x)} \\ 0 < e^{-\phi^-_{b_0}} + \frac{J^-}{\mu^- N^-_0}\int_{b}^{b_0} dx\, e^{V^-(x)} \end{array} \right\} \tag{2.9}$$

and

$$\mathcal{E}^+ := \left\{ (J^+, \varphi) \in \mathbb{R} \times C_{\mathbb{R}}(\Omega) : \begin{array}{l} 0 < e^{\phi^+_{a_0}} - \frac{J^+}{\mu^+ N^+}\int_{a_0}^{a} dx\, e^{V^+(x)} \\ 0 < e^{\phi^+_{b_0}} + \frac{J^+}{\mu^+ N^+}\int_{b}^{b_0} dx\, e^{V^+(x)} \end{array} \right\}. \tag{2.10}$$

Note that the definitions (2.7) and (2.8) make sense if $(J^{\pm}, \varphi) \in \mathcal{E}^{\pm}$. For pairs $(J^{\pm}, \varphi) \in \mathcal{E}^{\pm}$ we will indicate the dependence of $\phi^{\pm}$ from (2.7), (2.8) on the pairs by writing $\phi^{\pm}[J^{\pm}, \varphi]$.

For $(J^{\pm}, \varphi) \in \mathcal{E}^{\pm}$ the densities $U^{\pm}$ are given by means of (1.1), (2.7), and (2.8) by

$$U^-(x) = \begin{cases} N^-_0 e^{-V^-(x)} \left\{ e^{-\phi^-_{a_0}} - \frac{J^-}{\mu^- N^-_0}\int_{a_0}^{x} dy\, e^{V^-(y)} \right\} & \text{for } x \in (a_0, a), \\ N^-_0 e^{-V^-(x)} \left\{ e^{-\phi^-_{b_0}} + \frac{J^-}{\mu^- N^-_0}\int_{x}^{b_0} dy\, e^{V^-(y)} \right\} & \text{for } x \in (b, b_0), \end{cases} \tag{2.11}$$

and

$$U^+(x) = \begin{cases} N^+_0 e^{-V^+(x)} \left\{ e^{\phi^+_{a_0}} - \frac{J^+}{\mu^+ N^+_0}\int_{a_0}^{x} dy\, e^{V^+(y)} \right\} & \text{for } x \in (a_0, a), \\ N^+_0 e^{-V^+(x)} \left\{ e^{\phi^+_{b_0}} + \frac{J^+}{\mu^+ N^+_0}\int_{x}^{b_0} dy\, e^{V^+(y)} \right\} & \text{for } x \in (b, b_0). \end{cases} \tag{2.12}$$

Clearly the densities $U^{\pm}$ are positive if $(J^{\pm}, \varphi) \in \mathcal{E}^{\pm}$. The operators that assign the densities $U^{\pm}$ to pairs of $(J^{\pm}, \varphi)$, called the classical carrier density operators $\mathcal{D}^{\pm} : \mathcal{E}^{\pm} \longrightarrow L^1_{\mathbb{R}}(\Omega_c)$, are defined by

$$\mathcal{D}^{\pm}[J^{\pm}, \varphi] := U^{\pm}, \qquad (J^{\pm}, \varphi) \in \mathrm{dom}(\mathcal{D}^{\pm}) = \mathcal{E}^{\pm}, \tag{2.13}$$

where $U^{\pm}$ are given by (2.11) and (2.12). Of course, the carrier densities are not only from L^1 but in fact also are continuous functions. However, in section 3 we see that for the quantum densities the adequate function space is L^1. This suggests that we demand the same function space here.

3. Quantum zone. In this section we rigorously define the dissipative Schrödinger system described in the introduction of this paper; see (1.6)–(1.13). To that end we make the following general assumptions.

ASSUMPTION 3.1.

(Q.1) *The effective masses* $m^{\pm}$ *are positive and obey* $m^{\pm}$, $\frac{1}{m^{\pm}} \in L^{\infty}_{\mathbb{R}}(\Omega_q)$.

(Q.2) *The effective masses* $m_a^{\pm}$ *and* $m_b^{\pm}$ *entering the boundary coefficients* $\varkappa_a^{\pm}$ *and* $\varkappa_b^{\pm}$ *(see* (1.8)*) are strictly positive constants.*

(Q.3) *The approximation parameters* $\mathfrak{A} = \{\delta_0^{\pm}, \delta^{\pm}\}$ *entering* $\varkappa_a^{\pm}$ *and* $\varkappa_b^{\pm}$ *via* (1.9) *are strictly positive and obey* $\delta_0^{\pm} < \delta^{\pm} \leq \infty$.

(Q.4) *The band-edge offsets* $V_h^{\pm}$ *are essentially bounded over the whole device domain and continuous in the classical region* (*see also Assumption* (C.4)).

(Q.5) *The distribution functions* $f^{\pm} : \mathbb{R} \longrightarrow \mathbb{R}_+$ *are continuously differentiable and nonincreasing, i.e.,* $\frac{d}{dx} f^{\pm}(x) \leq 0$ *for* $x \in \mathbb{R}$, *such that*

$$D^{\pm}(s) := \sup_{\lambda \in [s,\infty)} f^{\pm}(\lambda)\sqrt{1+\lambda^2} < \infty, \quad s \in \mathbb{R}, \tag{3.1}$$

$$F^{\pm}(s) := \int_s^{\infty} d\lambda\, f^{\pm}(\lambda) < \infty, \quad s \in \mathbb{R}. \tag{3.2}$$

Moreover, the electrostatic potential $\varphi \in C_{\mathbb{R}}(\Omega)$ is assumed to be fixed in this section. The potentials $V^{\pm}$ are defined—as before—by (1.4).

Remark 3.2. Clearly, the reduced Boltzman distribution functions $f^{\pm}(\lambda) = n_0^{\pm}\, e^{-\lambda}$ and the reduced Fermi–Dirac distribution functions $f^{\pm}(\lambda) = n_0^{\pm}\, \ln(1+e^{-\lambda})$, where $n_0^{\pm}$ are the integrated density of states, satisfy assumption (Q.5).

We define the operators $H^{\pm}$ on the Hilbert space $L^2(\Omega_q)$ by

$$\mathrm{dom}(H^{\pm}) = \left\{ g \in W^{1,2}(\Omega_q) : \begin{array}{c} \frac{1}{m^{\pm}} g' \in W^{1,2}(\Omega_q), \\ \frac{1}{2m^{\pm}(a)} g'(a) = -\varkappa_a^{\pm} g(a), \\ \frac{1}{2m^{\pm}(b)} g'(b) = \varkappa_b^{\pm} g(b) \end{array} \right\} \tag{3.3}$$

and

$$(H^{\pm} g)(x) = -\frac{1}{2}\frac{d}{dx}\frac{1}{m^{\pm}(x)}\frac{d}{dx} g(x) + V^{\pm}(x) g(x), \tag{3.4}$$

where the coefficients $\varkappa_a^{\pm}$ and $\varkappa_b^{\pm}$ are given by (1.8). The operators $H^{\pm}$ are maximal dissipative and completely non–self-adjoint, since $\mathrm{Im}(\varkappa_a^{\pm}), \mathrm{Im}(\varkappa_b^{\pm}) > 0$; see [34].

The dissipative operators $H^{\pm}$ are regarded as pseudo-Hamiltonians of an open quantum system; see [17, 3]. By means of the dilation theorem (see, e.g., [21]), there exists a Hilbert space $\mathfrak{K}$ containing $L^2(\Omega_q)$ as a subspace and self-adjoint dilations $K^{\pm}$ on $\mathfrak{K}$ corresponding to the maximal dissipative operators $H^{\pm}$, i.e.,

$$(H^{\pm} - z)^{-1}\psi = P(K^{\pm} - z)^{-1}\psi \qquad \text{for all } \psi \in L^2(\Omega_q),$$

where P denotes the orthogonal projection from the dilation space $\mathfrak{K}$ onto the subspace $L^2(\Omega_q)$. Moreover, the minimality conditions

$$\mathfrak{K} = \operatorname*{clospan}_{z \in \mathbb{C}\setminus\mathbb{R}} \left\{ (K^{\pm} - z)^{-1}\psi,\ \psi \in L^2(\Omega_q) \right\}$$

are satisfied. The minimal self-adjoint dilations $K^{\pm}$ corresponding to the maximal dissipative operators $H^{\pm}$ can be obtained explicitly; see [35] and also section A.2. The spectrum of $K^{\pm}$ is absolutely continuous, is the whole real line, and has multiplicity 2. The incoming eigenfunctions of $K^{\pm}$ are denoted by $\psi_a^{\pm}(x,\lambda)$, $\psi_b^{\pm}(x,\lambda)$; see [35]. The macroscopic quantities such as carrier densities $u^{\pm}$ and current densities $j^{\pm}$ for the open quantum system are then determined by the incoming eigenfunctions of the quasi Hamiltonian and given statistics as in [33] and sections A.3 and A.4 by

$$u^{\pm}(x) = \sum_{\nu=a,b} \int_{\Lambda^{\pm}} d\lambda \; f^{\pm}(\lambda \mp \epsilon_\nu) |\psi_\nu^{\pm}(x,\lambda)|^2, \tag{3.5}$$

$$j^{\pm} = \sum_{\nu=a,b} \int_{\Lambda^{\pm}} d\lambda \; f^{\pm}(\lambda \mp \epsilon_\nu) \mathrm{Im}\left(\frac{1}{m(x)} \frac{\partial \psi_\nu^{\pm}(\lambda,x)}{\partial x} \overline{\psi_\nu^{\pm}(\lambda,x)} \right), \tag{3.6}$$

where $\epsilon_a^{\pm}$, $\epsilon_b^{\pm}$ are given Fermi levels and $\Lambda^{\pm}$ the sets given by (1.13). The current densities $j^{\pm}$ can be expressed by means of the transmission coefficients $0 \le t^{\pm}(\lambda) \le 1$ (see [35] and section A.4 in Appendix A) as

$$j^{\pm} = \frac{1}{2\pi} \int_{\Lambda^{\pm}} d\lambda \; t^{\pm}(\lambda) \Big(f^{\pm}(\lambda \mp \epsilon_a) - f^{\pm}(\lambda \mp \epsilon_b) \Big). \tag{3.7}$$

Clearly the carrier and current densities depend on the electrostatic potential φ since the eigenfunctions $\psi_a^{\pm}$, $\psi_b^{\pm}$ and the intervals $\Lambda^{\pm}$ depend on φ. Moreover, $u^{\pm}$ and $j^{\pm}$ depend on the Fermi energies $\epsilon_a^{\pm}$, $\epsilon_b^{\pm}$. Therefore, we define the carrier density operators $\mathcal{N}^{\pm} : \mathbb{R} \times \mathbb{R} \times C_{\mathbb{R}}(\Omega) \longrightarrow L^1(\Omega_q)$ and current density operators $j^{\pm} : \mathbb{R} \times \mathbb{R} \times C_{\mathbb{R}}(\Omega) \longrightarrow \mathbb{R}$, which assign to the Fermi levels and the electrostatic potential the corresponding carrier and current densities, i.e.,

$$\mathcal{N}^{\pm}[\epsilon_a^{\pm}, \epsilon_b^{\pm}, \varphi] := u^{\pm}(x), \quad (\epsilon_a^{\pm}, \epsilon_b^{\pm}, \varphi) \in \mathrm{dom}(\mathcal{N}^{\pm}) = \mathbb{R} \times \mathbb{R} \times C_{\mathbb{R}}(\Omega), \tag{3.8}$$

and

$$j^{\pm}[\epsilon_a^{\pm}, \epsilon_b^{\pm}, \varphi] := j^{\pm}, \quad (\epsilon_a^{\pm}, \epsilon_b^{\pm}, \varphi) \in \mathrm{dom}(j^{\pm}) = \mathbb{R} \times \mathbb{R} \times C_{\mathbb{R}}(\Omega). \tag{3.9}$$

The carrier density operators admit the estimate (cf. Proposition A.4 of section A.3)

$$\|\mathcal{N}^{\pm}[\epsilon_a^{\pm}, \epsilon_b^{\pm}, \varphi]\|_{L^1(\Omega_q)} \le C^{\pm}[\epsilon_a^{\pm}, \epsilon_b^{\pm}, \varphi] \left(\gamma_1^{\pm} + \gamma_2^{\pm} \sqrt{\|\varphi_{\mp}\|_{L^\infty(\Omega_q)}} \right), \tag{3.10}$$

where φ_+ and φ_- denote the positive and negative parts, respectively, of φ, i.e.,

$$\varphi_-(x) := \max\{0, -\varphi(x)\}, \quad \varphi_+(x) := \max\{0, \varphi(x)\}, \quad x \in \Omega_q. \tag{3.11}$$

The constants $\gamma_1^{\pm}$, $\gamma_2^{\pm}$ are independent of $\epsilon_a^{\pm}$, $\epsilon_b^{\pm}$, and φ. The constants $C^{\pm}[\epsilon_a^{\pm}, \epsilon_b^{\pm}, \varphi]$ are given by

$$C^{\pm}[\epsilon_a^{\pm}, \epsilon_b^{\pm}, \varphi] := \sup_{\lambda \in \Lambda^{\pm}[\varphi]} \max_{\nu=a,b} f^{\pm}(\lambda \mp \epsilon_\nu^{\pm}) \sqrt{1+\lambda^2}, \tag{3.12}$$

where $\Lambda^{\pm}[\varphi]$ denotes the dependence of the intervals $\Lambda^{\pm}$ given by (1.13) on the electrostatic potential φ.

4. Hybrid model. In this section we couple the drift-diffusion model of section 2 and the dissipative Schrödinger model of section 3, which leads to the dissipative hybrid model. This coupling is done in two steps: (i) the Fermi coupling and (ii) the current coupling.

4.1. Fermi coupling. Fermi coupling means to choose the Fermi energies $\epsilon_a^\pm$ and $\epsilon_b^\pm$ in the dissipative Schrödinger model (see (3.5), (3.6)) in an appropriate manner. Let us assume that the electrostatic potential φ and the current densities $J^\pm$ are given and obey $(J^\pm,\varphi) \in \mathcal{E}^\pm$, where $\mathcal{E}^\pm$ are given by (2.9), (2.10). Thus the Fermi potentials $\phi^\pm$ can be defined by means of (2.7) and (2.8). The Fermi levels $\epsilon_a^\pm$ and $\epsilon_b^\pm$ are then determined by

$$\epsilon_a^\pm = \phi^\pm(a) \quad \text{and} \quad \epsilon_b^\pm = \phi^\pm(b).$$

With this choice of the Fermi levels, we may define in accordance with (3.8) and (3.9) the mappings $\mathcal{N}^\pm : \mathcal{E}^\pm \longrightarrow L^1(\Omega_q)$,

$$\mathcal{N}^\pm[J^\pm,\varphi] := \mathcal{N}^\pm[\phi^\pm[J^\pm,\varphi](a), \phi[J^\pm,\varphi](b),\varphi], \quad (J^\pm,\varphi) \in \mathrm{dom}(\mathcal{N}^\pm) = \mathcal{E}^\pm, \tag{4.1}$$

and similar the mappings $j^\pm : \mathcal{E}^\pm \longrightarrow \mathbb{R}$ by

$$j^\pm[J^\pm,\varphi] := j^\pm[\phi^\pm[J^\pm,\varphi](a), \phi^\pm[J^\pm,\varphi](b),\varphi], \quad (J^\pm,\varphi) \in \mathrm{dom}(j^\pm) = \mathcal{E}^\pm. \tag{4.2}$$

4.2. Current coupling. The other condition for the coupling of the dissipative and drift-diffusion model is to impose the continuity of the current densities. More precisely, let the electrostatic potential $\varphi \in C_{\mathbb{R}}(\Omega)$ be given; then the current continuity condition reads

$$J^\pm = j^\pm[J^\pm,\varphi], \quad (J^\pm,\varphi) \in \mathcal{E}^\pm, \tag{4.3}$$

where $j^\pm$ are given by (4.2). Let us first show that condition (4.3) is well posed.

Proposition 4.1. *If Assumptions* 2.1 *and* 3.1 *are satisfied, then for any electrostatic potential $\varphi \in C_{\mathbb{R}}(\Omega)$, the equations* (4.3) *admit unique solutions $J^\pm$ such that $(J^\pm,\varphi) \in \mathcal{E}^\pm$.*

Proof. Since the considerations for holes and electrons are the same, we restrict ourselves in the following to holes and consider only the current continuity equation $J^+ = j^+[J^+,\varphi]$, $(J^+,\varphi) \in \mathcal{E}^+$. The transmission coefficient $t^+(z)$ is holomorphic and bounded by 1 in $z \in \mathbb{C}_-$; see section A.2, equation (A.23). We note that the limit $t^+(\lambda) = \lim_{\beta\uparrow 0} t^+(\lambda - i\beta)$ exists for all $\lambda \in \mathbb{R}$ and is bounded by 1, too. Using the uniqueness theorem for the H^∞-function (cf. Corollary II.4.2 of [26]), we find that this limit is different from zero for a.e. $\lambda \in \mathbb{R}$. Hence the function $t^+(\lambda)$ is different from zero and obeys the estimate

$$0 \le t^+(\lambda) \le 1, \quad \varphi \in C_{\mathbb{R}}(\Omega), \tag{4.4}$$

for a.e. $\lambda \in \mathbb{R}$. Inserting (2.7) and (2.8) into (3.7), we find

$$\begin{aligned} j^+[J^+,\varphi] = \frac{1}{2\pi}\int_{\Lambda^+[\varphi]} d\lambda\, t^+(\lambda)\Bigg\{ f^+\left(\lambda - \ln\left(e^{\phi^+_{a_0}} - \frac{J^+}{\mu^+ N_0^+}\int_{a_0}^a dy\, e^{V^+(y)}\right)\right) \\ -f^+\left(\lambda - \ln\left(e^{\phi^+_{b_0}} + \frac{J^+}{\mu_b^+ N_b^+}\int_b^{b_0} dy\, e^{V^+(y)}\right)\right)\Bigg\}. \end{aligned}$$

Since $\frac{d}{dx}f^+ \le 0$, one gets $\frac{\partial}{\partial J^+} j^+[J^+,\varphi] \le 0$. Hence, if φ is fixed, then the function $j^+[J^+,\varphi]$ is nonincreasing in J^+. By Lemma 2.2 one has $(J^+,\varphi) \in \mathcal{E}^+ \iff J^+ \in (J^+_{min}, J^+_{max})$. If $J^+ \uparrow J^+_{max}$, then $\phi^+[J^+,\varphi](a) \to -\infty$. Using the estimate

$$\int_{\Lambda^+} d\lambda\, t^+(\lambda) f^+(\lambda - \phi^+[J^+,\varphi](a)) \le \int_{V^+_{max} - \phi^+[J^+,\varphi](a)}^\infty d\lambda\, f^+(\lambda),$$

we find

$$\lim_{J^+ \uparrow J^+_{max}} \int_{\Lambda^+} d\lambda\, t^+(\lambda) f^+(\lambda - \phi^+[J^+,\varphi](a)) = 0,$$

which gives

$$(4.5)\quad \begin{aligned} j^+[J^+_{max},\varphi] &:= \lim_{J^+ \uparrow J^+_{max}} j^+[J^+,\varphi] \\ &= -\frac{1}{2\pi} \int_{\Lambda^+} d\lambda\, t^+(\lambda) f^+ \left(\lambda - \ln\left(e^{\phi^+_{b_0}} + \frac{J^+_{max}}{\mu^+ N^+_0} \int_b^{b_0} dy\, e^{V^+(y)}\right)\right). \end{aligned}$$

Similarly, if $J^+ \downarrow J^+_{min}$, then $\phi^+[J^+,\varphi](b) \to -\infty$. Hence

$$\lim_{J^+ \downarrow J^+_{min}} \int_{\Lambda^+} d\lambda\, t^+(\lambda) f^+(\lambda - \phi^+[J^+,\varphi](b)) = 0,$$

which yields

$$(4.6)\quad \begin{aligned} j^+[J^+_{min},\varphi] &:= \lim_{J^+ \downarrow J^+_{min}} j^+[J^+,\varphi] \\ &= \frac{1}{2\pi} \int_{\Lambda^+} d\lambda\, (\lambda) f^+ \left(\lambda - \ln\left(e^{\phi^+_{a_0}} - \frac{J^+_{min}}{\mu^+ N^+_0} \int_{a_0}^{a} dy\, e^{V^+(y)}\right)\right). \end{aligned}$$

Since $j^+[J^+,\varphi]$ is continuous and nonincreasing in J^+, as are $j^+[J^+_{min},\varphi] > 0$ and $j^+[J^+_{max},\varphi] < 0$, one immediately gets that the equation $J^+ = j^+[J^+,\varphi]$ admits a unique solution J^+ for each $\varphi \in C_{\mathbb{R}}(\Omega)$ such that $(J^+,\varphi) \in \mathcal{E}^+$. ∎

4.3. Dissipative hybrid system. If $J^\pm = J^\pm[\varphi]$ are the solutions of (4.3) for a given electrostatic potential $\varphi \in C_{\mathbb{R}}(\Omega)$, then it makes sense to introduce the following quantities of the dissipative hybrid system:

(4.7) $\quad \phi^\pm[\varphi] := \phi^\pm[J^\pm[\varphi],\varphi]$ Fermi potentials (cf. (2.7), (2.8)),

(4.8) $\quad \mathcal{N}^\pm[\varphi] := \mathcal{N}^\pm[J^\pm[\varphi],\varphi]$ dissipative particle density operators (cf. (4.1)),

(4.9) $\quad \mathcal{D}^\pm[\varphi] := \mathcal{D}^\pm[J^\pm[\varphi],\varphi]$ classical density operators (cf. (2.13)).

Moreover, we introduce the hybrid carrier density operators as follows.

DEFINITION 4.2. *Let Assumptions* 2.1 *and* 3.1 *be satisfied. The carrier density operator* $\mathcal{U}^\pm[\cdot] : C_{\mathbb{R}}(\Omega) \longrightarrow L^1_{\mathbb{R}}(\Omega)$ *of the dissipative hybrid system is defined by*

$$(4.10)\qquad \mathcal{U}^\pm[\varphi](x) := \begin{cases} \mathcal{D}^\pm[\varphi](x), & x \in \Omega_c, \\ \mathcal{N}^\pm[\varphi](x), & x \in \Omega_q. \end{cases}$$

In order to couple the hybrid system with the Poisson equation, we need to verify certain properties of the quantities (4.7)–(4.9). In the following lemma we give an L^∞-estimate of the quasi-Fermi potentials $\phi^\pm[\varphi]$ that is uniform in $\varphi \in C_{\mathbb{R}}(\Omega)$.

LEMMA 4.3. *If Assumptions* 2.1 *and* 3.1 *are satisfied, then for any* $\varphi \in C_{\mathbb{R}}(\Omega)$ *one has*

$$(4.11)\qquad \max_{x \in \Omega_c} \{|\phi^\pm[\varphi](x)|\} \le \eta^\pm,$$

where

$$\eta^{\pm} := \max\left\{\left|\phi^{\pm}_{a_0}\right|, \left|\phi^{\pm}_{b_0}\right|\right\}. \tag{4.12}$$

Proof. Assume that $J^+[\varphi] \geq 0$. Since $J^+[\varphi]$ solves (4.3), one gets from (3.7) and the monotonicity of the functions $f^{\pm}$ that

$$f^+(\lambda - \phi^+[\varphi](a)) \geq f^+(\lambda - \phi^+[\varphi](b)), \quad \lambda \in \Lambda^+.$$

Taking into account the monotonicity of f^+, we find that

$$-\phi^+[\varphi](a) \leq -\phi^+[\varphi](b).$$

If $J^+[\varphi] \geq 0$, then $-\phi^+[\varphi](x)$, $x \in \Omega_c$, is nondecreasing. That means we have

$$-\phi^+[\varphi](a_0) \leq -\phi^+[\varphi](a) \quad \text{and} \quad -\phi^+[\varphi](b) \leq -\phi^+[\varphi](b_0).$$

Hence

$$-\phi^+[\varphi](a_0) \leq -\phi^+[\varphi](a) \leq -\phi^+[\varphi](b) \leq -\phi^+[\varphi](b_0),$$

which shows that

$$\max\{|\phi^+[\varphi](a)|, |\phi^+[\varphi](b)|\} \leq \max\{|\phi^+[\varphi](a_0)|, |\phi^+[\varphi](b_0)|\}.$$

If $J^+[\varphi] \leq 0$, then from (3.7) one gets

$$f^+(\lambda - \phi^+[\varphi](a)) \leq f^+(\lambda - \phi^+[\varphi](b)), \quad \lambda \in \Lambda^+,$$

which yields

$$\phi^+[\varphi](a) \leq \phi^+[\varphi](b).$$

Since $\phi^+[\varphi](x)$ is nondecreasing on Ω_c, we find

$$\phi^+[\varphi](a_0) \leq \phi^+[\varphi](a) \leq \phi^+[\varphi](b) \leq \phi^+[\varphi](b_0),$$

which gives

$$\max\{|\phi^+[\varphi](a)|, |\phi^+[\varphi](b)|\} \leq \max\{|\phi^+[\varphi](a_0)|, |\phi^+[\varphi](b_0)|\} = \eta^+.$$

We complete the proof for holes with the remark that the quasi-Fermi potential $\phi^+[\varphi]$ is monotone on each subinterval (a_0, a) and (b, b_0). The proof for electrons is similar. ∎

With the help of Lemma 4.3, we prove an estimate for the carrier density operators.

LEMMA 4.4. *If Assumptions* 2.1 *and* 3.1 *are satisfied, then for any $\varphi \in C_{\mathbb{R}}(\Omega)$ the carrier density operators $\mathcal{N}^{\pm}[\cdot]$ admit the estimates*

$$\|\mathcal{N}^{\pm}[\varphi]\|_{L^1(\Omega_q)} \leq C^{\pm}(V^{\pm}_{max}[\varphi]) \left(\gamma^{\pm}_1 + \gamma^{\pm}_2 \sqrt{\|\varphi_{\mp}\|_{L^\infty(\Omega_q)}}\right), \tag{4.13}$$

where φ_+ and φ_- denote the positive and negative parts, respectively, of φ. The constants $\gamma^{\pm}_1$, $\gamma^{\pm}_2$ are independent of φ; $V^{\pm}_{\max}[\varphi]$ are given by (1.10), *and the functions $C^{\pm}(s)$ are given by*

$$C^{\pm}(s) := D^{\pm}(s - \eta^{\pm})(1 + \eta^{\pm}), \quad s \in \mathbb{R}, \tag{4.14}$$

with $D^{\pm}(\cdot)$ *and* $\eta^{\pm}$ *given by* (3.1) *and* (4.12), *respectively.*

Proof. From (3.10) and (3.12) we get

$$\|\mathcal{N}^{\pm}[\varphi]\|_{L^1(\Omega_q)} \le C^{\pm}[\phi^{\pm}[\varphi](a), \phi^{\pm}[\varphi](b), \varphi] \left(\gamma_1^{\pm} + \gamma_2^{\pm}\sqrt{\|\varphi_{\mp}\|_{L^\infty(\Omega_q)}}\right).$$

Thus, it suffices to show that constants $C^{\pm}[\phi^{\pm}[\varphi](a), \phi^{\pm}[\varphi](b), \varphi]$ are estimated by $C^{\pm}(V^{\pm}_{max}[\varphi])$. Let us consider the case of holes. We find by means of the definition of Λ^+ (see (1.13)) that

$$\sup_{\lambda\in\Lambda^+[\varphi]} \sqrt{1+\lambda^2} f^+(\lambda - \phi^+[\varphi](b)) \le D^+(V^+_{max}[\varphi] - \phi^+[\varphi](b)) \sup_{x\in\mathbb{R}} \left(\frac{1+(x+\phi^+[\varphi](b))^2}{1+x^2}\right)^{1/2}.$$

Since

$$\sup_{x\in\mathbb{R}} \left(\frac{1+(x+\phi^+[\varphi](b))^2}{1+x^2}\right)^{1/2} \le 1 + |\phi^+[\varphi](b)|,$$

we get

$$\sup_{\lambda\in\Lambda^+[\varphi]} \sqrt{1+\lambda^2} f^+(\lambda - \phi^+[\varphi](b)) \le D^+(V^+_{max}[\varphi] - \phi^+[\varphi](b))(1 + |\phi^+[\varphi](b)|).$$

In the same manner, we prove

$$\sup_{\lambda\in\Lambda^+[\varphi]} \sqrt{1+\lambda^2} f^+(\lambda - \phi^+[\varphi](a)) \le D^+(V^+_{max}[\varphi] - \phi^+[\varphi](a))(1 + |\phi^+[\varphi](a)|),$$

which yields

(4.15)

$$C^+[\phi^+[\varphi](a), \phi^+[\varphi](b), \varphi] \le \max_{\nu\in\{a,b\}} \left\{D^+(V^+_{max}[\varphi] - \phi^+[\varphi](\nu))(1 + |\phi^+[\varphi](\nu)|)\right\}.$$

Since $D^+(\cdot)$ is nonincreasing, we complete the proof using Lemma 4.3. The proof for electrons is similar. ∎

Like the carrier densities $\mathcal{N}^{\pm}[\varphi]$, the current densities $J^{\pm}[\varphi]$ also admit an estimate.

LEMMA 4.5. *If Assumptions* 2.1 *and* 3.1 *are satisfied, then for any* $\varphi \in C_{\mathbb{R}}(\Omega)$ *the estimates*

$$|J^{\pm}[\varphi]| \le \frac{1}{\pi} F^{\pm}(V^{\pm}_{max}[\varphi] - \eta^{\pm}) \tag{4.16}$$

are valid, where $V^{\pm}_{max}[\varphi]$ *are defined by* (1.10); $\eta^{\pm}$ *and the functions* $F^{\pm}(\cdot)$ *are given by* (4.12) *and by* (3.2), *respectively.*

Proof. We consider the case of holes. Since $J^+[\varphi]$ is a solution of (4.3), one has $J^+[\varphi] = j^+[J^+[\varphi], \varphi]$, where $j^+[J^+[\varphi], \varphi]$ is defined by (4.2). From (3.7) and the fact that the transmission coefficient $t^+[\varphi](\lambda)$ is uniformly bounded by one, we obtain

$$|J^+[\varphi]| \le \frac{1}{2\pi} \left\{ \int_{\Lambda^+[\varphi]} d\lambda \; f^+(\lambda - \phi^+[\varphi](b)) + f^+(\lambda - \phi^+[\varphi](a)) \right\},$$

which yields

$$|J^+[\varphi]| \leq \frac{1}{2\pi}\left\{F^+(V^+_{max}[\varphi] - \phi^+[\varphi](b)) + F^+(V^+_{max}[\varphi] - \phi^+[\varphi](a))\right\}.$$

By Lemma 4.3 we immediately get (4.16). The case of electrons is handled in a similar way. ∎

The next step is to show the continuity of the current density operator with respect to the electrostatic potential φ.

LEMMA 4.6. *Let Assumptions* 2.1 *and* 3.1 *be satisfied. If* $\varphi, \varphi_n \in C_{\mathbb{R}}(\Omega)$, $n \in \mathbb{N}$, *and* $\lim_{n\to\infty} \|\varphi_n - \varphi\|_{L^\infty(\Omega)} = 0$, *then* $\lim_{n\to\infty} J^\pm[\varphi_n] = J^\pm[\varphi]$.

Proof. We will prove the statement only for holes; the proof for electrons is similar. We set $J := J^+[\varphi]$ and $J_n := J^+[\varphi_n]$, $n \in \mathbb{N}$. If $J_n \not\to J$ as $n \to \infty$, then there is a subsequence $\{J_{n_k}\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} J_{n_k} = J_\infty \neq J$. This results from Lemma 4.5, which shows the uniform boundedness of $\{J_n\}^\infty_{n\in\mathbb{N}}$.

Let us show that $(J_\infty, \varphi) \in \mathcal{E}^+$. Since $(J_n, \varphi_n) \in \mathcal{E}^+$, $n \in \mathbb{N}$, and

$$\lim_{n\to\infty} J^+_{min}[\varphi_n] = J^+_{min}[\varphi] \quad \text{and} \quad \lim_{n\to\infty} J^+_{max}[\varphi_n] = J^+_{max}[\varphi],$$

one has $(J_\infty, \varphi) \notin \mathcal{E}^+$ if and only if either $J_\infty = J^+_{max}[\varphi]$ or $J_\infty = J^+_{min}[\varphi]$. However, this is impossible; namely, if $\lim_{k\to\infty} J_{n_k} = J^+_{max}[\varphi] > 0$, then $\lim_{k\to\infty} j^+[J_{n_k}, \varphi_{n_k}] = j^+[J_\infty, \varphi] \leq 0$; cf. (4.5). Similarly, if $\lim_{k\to\infty} J_{n_k} = J^+_{min}[\varphi] < 0$, then $\lim_{k\to\infty} j^+[J_{n_k}, \varphi_{n_k}] = j^+[J^+_\infty, \varphi] \geq 0$; cf. (4.6).

Since $(J_\infty, \varphi) \in \mathcal{E}^+$, the quantities $\phi^+[J_\infty, \varphi]$ are well defined. One gets $\lim_{k\to\infty} \phi^+[J_{n_k}, \varphi_{n_k}](b) = \phi^+[J_\infty, \varphi](b)$ and $\lim_{k\to\infty} \phi^+[J_{n_k}, \varphi_{n_k}](a) = \phi^+[J_\infty, \varphi](a)$, which yields

$$\lim_{k\to\infty} f^+(\lambda - \phi^+[J_{n_k}, \varphi_{n_k}](\nu)) = f^+(\lambda - \phi^+[J^\pm_\infty, \varphi](\nu)) \qquad \text{for a.e. } \lambda \in \mathbb{R},\ \nu = a, b.$$

From the uniform boundedness of the Fermi potentials (see Lemma 4.3), we obtain from Theorem A.9

$$\lim_{k\to\infty} j^+[J_{n_k}, \varphi_{n_k}] = j^+[J_\infty, \varphi].$$

By $J_{n_k} = j^+[J_{n_k}, \varphi_{n_k}]$ we find

$$J_\infty = \lim_{k\to\infty} J_{n_k} = j^+[J_\infty, \varphi].$$

Since the solution of this equation is unique, one gets $J_\infty = J^+[\varphi]$, which proves the continuity. ∎

Next, let us show that the carrier density operators $\mathcal{U}^\pm[\cdot]$ are continuous. To this end we first prove the continuity of the dissipative carrier density operators $\mathcal{N}^\pm[\cdot]$.

LEMMA 4.7. *Let Assumptions* 2.1 *and* 3.1 *be satisfied. If* $\varphi, \varphi_n \in C_{\mathbb{R}}(\Omega)$, $n \in \mathbb{N}$, *and* $\lim_{n\to\infty} \|\varphi_n - \varphi\|_{L^\infty(\Omega)} = 0$, *then* $\lim_{n\to\infty} \|\mathcal{N}^\pm[\varphi_n] - \mathcal{N}^\pm[\varphi]\|_{L^1(\Omega_q)} = 0$.

Proof. Let $\varphi, \varphi_n \in C_{\mathbb{R}}(\Omega)$, $n \in \mathbb{N}$, with $\lim_{n\to\infty} \|\varphi - \varphi_n\|_{L^\infty(\Omega)}$ be given. From Lemma 4.6 and (2.7), (2.8), we immediately obtain

$$\lim_{n\to\infty} \phi^\pm[\varphi_n](x) = \phi^\pm[\varphi](x) \quad \text{for every } x \in \Omega_c.$$

Hence, we get $\lim_{n\to\infty} f^\pm(\lambda \mp \phi^\pm[\varphi_n](\nu)) = f^\pm(\lambda \mp \phi^\pm[\varphi](\nu))$ for $\nu = a, b$. Taking into account the uniform boundedness of $\phi^\pm$ (cf. Lemma 4.3) and Theorem A.7, one completes the proof. ∎

PROPOSITION 4.8. *Let Assumptions* 2.1 *and* 3.1 *be satisfied. If* $\varphi, \varphi_n \in C_{\mathbb{R}}(\Omega)$, $n \in \mathbb{N}$, *and* $\lim_{n\to\infty} \|\varphi_n - \varphi\|_{L^\infty(\Omega)} = 0$, *then* $\lim_{n\to\infty} \|\mathcal{U}^\pm[\varphi_n] - \mathcal{U}^\pm[\varphi]\|_{L^1(\Omega)} = 0$.

Proof. Taking into account Lemma 4.7, it remains to show that

$$\lim_{n\to\infty} \|\mathcal{D}^\pm[\varphi_n] - \mathcal{D}^\pm[\varphi]\|_{L^1(\Omega_c)} = 0.$$

However, this follows immediately from Lemma 4.6 and (2.11), (2.12). ∎

4.4. Coupling to Poisson's equation: Dissipative hybrid model. In order to have a meaningful model for semiconductors, the electrostatic potential has to be computed self-consistently by a Poisson equation. In this section we pose the Poisson equation on the whole device domain Ω, where the right-hand side depends on the densities of the dissipative hybrid model. This leads to a nonlinear equation for the electrostatic potential φ, which will be reformulated as a fixed point problem.

Concerning the data for Poisson's equation, we make the following assumptions.

ASSUMPTION 4.9.

(P.1) *The doping profile* C *belongs to* $W^{-1,2}_{\mathbb{R}}(\Omega)$.

(P.2) *The dielectric permittivity* ε *is positive and obeys* $\varepsilon, \frac{1}{\varepsilon} \in L^\infty_{\mathbb{R}}(\Omega)$.

By $\widehat{\varphi}$ we denote the function which satisfies $\widehat{\varphi} \in W^{1,2}_{\mathbb{R}}(\Omega)$, $\varepsilon \frac{d}{dx}\widehat{\varphi} \in W^{1,2}_{\mathbb{R}}(\Omega)$, $\widehat{\varphi}(a_0) = \varphi_{a_0}$, and $\widehat{\varphi}(b_0) = \varphi_{b_0}$, and additionally

$$-\frac{d}{dx}\varepsilon(x)\frac{d}{dx}\widehat{\varphi}(x) = 0, \quad x \in \Omega.$$

DEFINITION 4.10. *We define the linear Poisson operator with zero boundary conditions* $\mathcal{P} : \overset{o}{W}{}^{1,2}_{\mathbb{R}}(\Omega) \longrightarrow W^{-1,2}_{\mathbb{R}}(\Omega)$ *by*

$$\langle \mathcal{P}\upsilon, \zeta \rangle := \int_a^b dx\, \varepsilon(x)\, \frac{d\upsilon}{dx}\frac{d\zeta}{dx}, \qquad \upsilon, \zeta \in \overset{o}{W}{}^{1,2}_{\mathbb{R}}(\Omega). \tag{4.17}$$

DEFINITION 4.11. *Assume* $w^\pm \in L^1(\Omega)$. *We say* $\varphi \in W^{1,2}_{\mathbb{R}}(\Omega)$ *satisfies Poisson's equation if* $\zeta := \varphi - \widehat{\varphi} \in \overset{o}{W}{}^{1,2}_{\mathbb{R}}(\Omega)$ *and, additionally, satisfies*

$$\mathcal{P}\zeta = C + w^+ - w^-, \tag{4.18}$$

where the function $w^\pm$ *must be understood in the sense of the embedding* $L^1(\Omega) \hookrightarrow W^{-1,2}_{\mathbb{R}}(\Omega)$

DEFINITION 4.12. *Let Assumptions* 2.1, 3.1, *and* 4.9 *be satisfied. We say an element* $\varphi \in W^{1,2}_{\mathbb{R}}(\Omega)$ *is a solution of the dissipative hybrid model if*

(i) *the carrier densities* $w^\pm \in L^1(\Omega)$ *are given by the hybrid densities, i.e.,* $w^\pm = \mathcal{U}^\pm[\varphi]$ *(cf.* (4.10)*), and*

(ii) *the potential* φ *satisfies Poisson's equation.*

We note that if $\varphi \in W^{1,2}_{\mathbb{R}}(\Omega)$ is a solution of the dissipative hybrid model, then the current densities are given by $J^\pm[\varphi]$.

Let us introduce for each fixed electrostatic potential $\varphi \in C_{\mathbb{R}}(\Omega)$ the map $\mathcal{R}[\varphi] : \overset{o}{W}{}^{1,2}_{\mathbb{R}}(\Omega) \longrightarrow W^{-1,2}_{\mathbb{R}}(\Omega)$,

$$\mathcal{R}[\varphi](\zeta) := \mathcal{P}\zeta + \left\{p^-[\varphi]e^{\zeta} - p^+[\varphi]e^{-\zeta}\right\}, \tag{4.19}$$

where

$$p^-[\varphi](x) := \begin{cases} N_0^- e^{\widehat{\varphi}(x) - V_h^-(x)} \left(e^{-\phi_{a_0}^-} - \frac{J^-[\varphi]}{\mu^- N_0^-} \int_{a_0}^x dy\, e^{V^-(y)} \right), & x \in (a_0, a), \\ 0, & x \in \Omega_q, \\ N_0^- e^{\widehat{\varphi}(x) - V_h^-(x)} \left(e^{-\phi_{b_0}^-} + \frac{J^-[\varphi]}{\mu^- N_0^-} \int_x^{b_0} dy\, e^{V^-(y)} \right), & x \in (b, b_0), \end{cases} \tag{4.20}$$

and

$$p^+[\varphi](x) := \begin{cases} N_0^+ e^{-(\widehat{\varphi}(x) + V_h^+(x))} \left(e^{\phi_{a_0}^+} - \frac{J^+[\varphi]}{\mu^+ N_0^+} \int_{a_0}^x dy\, e^{V^+(y)} \right), & x \in (a_0, a), \\ 0, & x \in \Omega_q, \\ N_0^+ e^{-(\widehat{\varphi}(x) + E_h^+(x))} \left(e^{\phi_{b_0}^+} + \frac{J^+[\varphi]}{\mu^+ N_0^+} \int_x^{b_0} dy\, e^{V^+(y)} \right), & x \in (b, b_0), \end{cases} \tag{4.21}$$

$\varphi \in C_{\mathbb{R}}(\Omega)$, where $V^\pm = V^\pm[\varphi]$ are given by (1.4). We note that $p^\pm[\varphi] \in L^\infty_{\mathbb{R}}(\Omega)$ and that drift-diffusion densities can be written as

$$U^\pm[J^\pm[\varphi], \varphi](x) = p^\pm[\varphi](x) e^{\mp\zeta(x)}, \quad x \in \Omega_c, \tag{4.22}$$

where $\varphi = \zeta + \widehat{\varphi}$ and $U^\pm[J^\pm[\varphi], \varphi]$ are given by (2.11) and (2.12), respectively.

Concerning the next lemma and its proof, we follow the terminology of [24]; in particular, the notions of strong monotonicity and boundedly Lipschitz continuity are used in accordance with Definitions III.1.1 and III.1.2 of [24].

LEMMA 4.13. *Let Assumptions* 2.1 *and* 4.9 *be satisfied. If* $\varphi \in C_{\mathbb{R}}(\Omega)$, *then the operator* $\mathcal{R}[\varphi]$ *is strongly monotone with monotonicity constant* $\frac{1}{\|1/\varepsilon\|_{L^\infty(\Omega)}}$ *and boundedly Lipschitz continuous.*

Proof. We note that the operator $\mathcal{P}$ is linear, is bounded, and obeys

$$\langle \mathcal{P}\zeta, \zeta \rangle \ \geq \ \frac{1}{\|1/\varepsilon\|_{L^\infty(\Omega)}} \|\zeta\|^2_{\overset{o}{W}{}^{1,2}_{\mathbb{R}}(\Omega)}. \tag{4.23}$$

Hence, $\mathcal{P}$ is a strongly monotone operator with monotonicity constant $\frac{1}{\|1/\varepsilon\|_{L^\infty(\Omega)}}$, which maps $\overset{o}{W}{}^{1,2}_{\mathbb{R}}(\Omega)$ onto $W^{-1,2}_{\mathbb{R}}(\Omega)$.

By Proposition 4.1 one has $(J^\pm[\varphi], \varphi) \in \mathcal{E}^\pm$; hence, $p^\pm[\varphi](x) \geq 0$ for $x \in \Omega$. Using this, one verifies that for each $\varphi \in C_{\mathbb{R}}(\Omega)$ the nonlinear operator $T[\varphi] : \overset{o}{W}{}^{1,2}_{\mathbb{R}}(\Omega) \longrightarrow W^{-1,2}_{\mathbb{R}}(\Omega)$,

$$T[\varphi](\zeta) := \left\{ p^-[\varphi] e^{\zeta} - p^+[\varphi] e^{-\zeta} \right\}, \tag{4.24}$$

is monotone. Hence, the sum $\mathcal{R}[\varphi] = \mathcal{P} + T[\varphi]$ is a strongly monotone operator with the same monotonicity constant as $\mathcal{P}$. Since $\mathcal{P}$ is bounded and linear, it is obviously Lipschitz continuous. A straightforward computation shows that $T[\varphi]$ is boundedly Lipschitz continuous, too. Hence the sum $\mathcal{R}[\varphi]$ is also boundedly Lipschitz continuous. ∎

Remark 4.14. From our Lemma 4.13, and Corollary III.2.3 of [24], we obtain that for $\varphi \in C_{\mathbb{R}}(\Omega)$ the operator $\mathcal{R}[\varphi]^{-1} : W^{-1,2}_{\mathbb{R}}(\Omega) \longrightarrow \overset{o}{W}{}^{1,2}_{\mathbb{R}}(\Omega)$ exists, is bounded, and is Lipschitz continuous with a Lipschitz constant not bigger than $\|1/\varepsilon\|_{L^\infty(\Omega)}$.

Let us introduce the mapping $\mathcal{Q} : C_{\mathbb{R}}(\Omega) \longrightarrow W^{1,2}_{\mathbb{R}}(\Omega)$ defined by

$$\mathcal{Q}(\varphi) := \widehat{\varphi} + \mathcal{R}[\varphi]^{-1}(C - \mathcal{N}^-[\varphi] + \mathcal{N}^+[\varphi]), \tag{4.25}$$

where $\varphi \in \mathrm{dom}(\mathcal{Q}) = C_{\mathbb{R}}(\Omega)$ and $\mathcal{N}^{\pm}[\varphi]$ have to be seen as elements from $W_{\mathbb{R}}^{-1,2}(\Omega)$. In what follows we simultaneously regard the mapping $\mathcal{Q}$ also as mapping from $C_{\mathbb{R}}(\Omega)$ into itself by means of the embedding $W_{\mathbb{R}}^{1,2}(\Omega) \hookrightarrow C_{\mathbb{R}}(\Omega)$.

PROPOSITION 4.15. *Let Assumptions* 2.1, 3.1, *and* 4.9 *be satisfied. An element* $\varphi \in W_{\mathbb{R}}^{1,2}(\Omega)$ *is a solution of the dissipative hybrid model if and only if* φ *is a fixed point of the mapping* $\mathcal{Q} : C_{\mathbb{R}}(\Omega) \longrightarrow C_{\mathbb{R}}(\Omega)$.

Proof. Let us assume that $\varphi \in W_{\mathbb{R}}^{1,2}(\Delta)$ is a solution of the dissipative hybrid model. From Definition 4.12 we get

$$\mathcal{P}\zeta = C - \mathcal{U}^{-}[\varphi] + \mathcal{U}^{+}[\varphi], \tag{4.26}$$

where $\varphi = \zeta + \widehat{\varphi}$. By means of the relation (4.22) we have

$$-\mathcal{D}^{-}[\varphi] + \mathcal{D}^{+}[\varphi] = -p^{-}[\varphi]e^{\zeta} + p^{+}[\varphi]e^{-\zeta}, \tag{4.27}$$

and thus, by definition of $\mathcal{U}^{\pm}$ (see (4.10)), we get

$$\mathcal{R}[\varphi](\zeta) = C - \mathcal{N}^{-}[\varphi] + \mathcal{N}^{+}[\varphi]. \tag{4.28}$$

Therefore we obtain

$$\varphi = \widehat{\varphi} + \mathcal{R}[\varphi]^{-1}(C - \mathcal{N}^{-}[\varphi] + \mathcal{N}^{+}[\varphi]), \tag{4.29}$$

which implies $\varphi = \mathcal{Q}[\varphi]$; i.e., φ is a fixed point of $\mathcal{Q}$. The converse statement is proven in a similar manner. ∎

5. Existence.

5.1. Preliminaries. Our final aim is to show that the dissipative hybrid model introduced in the previous section always admits a solution. By Proposition 4.15 this is equivalent to showing that the nonlinear mapping $\mathcal{Q} : C_{\mathbb{R}}(\Omega) \longrightarrow C_{\mathbb{R}}(\Omega)$ admits a fixed point. This will be done by applying the Leray–Schauder fixed point theorem [25, Theorem 11.3]. To this end we consider the nonlinear equation

$$\vartheta = t\mathcal{Q}(\vartheta), \quad \vartheta \in C_{\mathbb{R}}(\Omega), \quad t \in [0,1]. \tag{5.1}$$

Let us introduce the modified carrier density operators $\mathcal{U}_t^{\pm}[\cdot] : C_{\mathbb{R}}(\Omega) \longrightarrow L^1(\Omega)$, $t \in [0,1]$,

$$\mathcal{U}_t^{\pm}[\varphi](x) := \begin{cases} \mathcal{D}^{\pm}[t\varphi](x)e^{\mp(1-t)\varphi(x)}, & x \in \Omega_c, \\ \mathcal{N}^{\pm}[t\varphi](x), & x \in \Omega_q. \end{cases} \tag{5.2}$$

We note that $\mathcal{U}^{\pm}[\varphi] = \mathcal{U}_1^{\pm}[\varphi]$, $\varphi \in C_{\mathbb{R}}(\Omega)$; cf. (4.10).

LEMMA 5.1. *Let Assumptions* 2.1, 3.1, *and* 4.9 *be satisfied. If* $\vartheta \in C_{\mathbb{R}}(\Omega)$ *satisfies* (5.1) *for* $t \in [0,1]$, *then there is an element* $\varphi \in W_{\mathbb{R}}^{1,2}(\Omega)$ *such that* $\vartheta = t\varphi$, *and that* $\zeta := \varphi - \widehat{\varphi} \in \overset{o}{W}{}_{\mathbb{R}}^{1,2}(\Omega)$ *satisfies the modified Poisson equation*

$$\mathcal{P}\zeta = C + \mathcal{U}_t^{+}[\varphi] - \mathcal{U}_t^{-}[\varphi]. \tag{5.3}$$

Proof. The proof of the lemma is essentially the same as the proof of Proposition 4.15. ∎

Having in mind an application of the Leray–Schauder fixed point theorem, one has to show that the mapping $\mathcal{Q}$ is compact, i.e., continuous and maps every bounded set into a precompact one; cf. section 11.2 of [25]. This will be shown by the following lemmata.

LEMMA 5.2. *If Assumptions* 2.1, 3.1, *and* 4.9 *are satisfied, then the mapping* $\mathcal{Q} : C_{\mathbb{R}}(\Omega) \longrightarrow C_{\mathbb{R}}(\Omega)$ *is continuous.*

Proof. Let $\varphi, \varphi_n \in C_{\mathbb{R}}(\Omega)$, $n \in \mathbb{N}$, such that $\lim_{n\to\infty} \|\varphi - \varphi_n\|_{L^\infty(\Omega)} = 0$. We set

$$\psi := C - \mathcal{N}^-[\varphi] + \mathcal{N}^+[\varphi] \in W^{-1,2}_{\mathbb{R}}(\Omega)$$

and

$$\psi_n := C - \mathcal{N}^-[\varphi_n] + \mathcal{N}^+[\varphi_n] \in W^{-1,2}_{\mathbb{R}}(\Omega), \quad n \in \mathbb{N}.$$

By Lemma 4.7 we find $\lim_{n\to\infty} \|\mathcal{N}^\pm[\varphi_n] - \mathcal{N}^\pm[\varphi]\|_{L^1(\Omega)} = 0$, which yields

$$\lim_{n\to\infty} \|\psi_n - \psi\|_{W^{-1,2}(\Omega)} = 0. \tag{5.4}$$

Let us show that

$$\lim_{n\to\infty} \|\mathcal{R}[\varphi_n]^{-1}(\psi_n) - \mathcal{R}[\varphi]^{-1}(\psi)\|_{W^{1,2}(\Omega)} = 0. \tag{5.5}$$

Obviously one has

$$\psi_n - \mathcal{R}[\varphi_n](\mathcal{R}[\varphi]^{-1}(\psi)) = \psi_n - \psi - \left\{\mathcal{R}[\varphi_n](\mathcal{R}[\varphi]^{-1}(\psi)) - \psi\right\}. \tag{5.6}$$

The sequence $\mathcal{R}[\varphi_n] : \overset{o}{W}{}^{1,2}(\Omega) \longrightarrow W^{-1,2}(\Omega)$ converges strongly to $\mathcal{R}[\varphi]$; i.e., for each $\zeta \in \overset{o}{W}{}^{1,2}(\Omega)$ one has $\mathcal{R}[\varphi_n](\zeta) \to \mathcal{R}[\varphi](\zeta)$ as $n \to \infty$. Hence

$$\lim_{n\to\infty} \|\mathcal{R}[\varphi_n](\mathcal{R}[\varphi]^{-1}(\psi)) - \psi\|_{W^{-1,2}(\Omega)} = 0. \tag{5.7}$$

From (5.4), (5.6), and (5.7) we get

$$\lim_{n\to\infty} \|\psi_n - \mathcal{R}[\varphi_n](\mathcal{R}[\varphi]^{-1}(\psi))\|_{W^{-1,2}(\Omega)} = 0. \tag{5.8}$$

Using the representation

$$\mathcal{R}[\varphi_n]^{-1}(\psi_n) - \mathcal{R}[\varphi]^{-1}(\psi) = \mathcal{R}[\varphi_n]^{-1}(\psi_n) - \mathcal{R}[\varphi_n]^{-1}(\mathcal{R}[\varphi_n](\mathcal{R}[\varphi]^{-1}(\psi)))$$

and (5.8), we obtain from Remark 4.14 the relation (5.5), which yields the continuity of $\mathcal{Q}$. ∎

LEMMA 5.3. *If Assumptions* 2.1, 3.1, *and* 4.9 *are satisfied, then the mapping* $\mathcal{Q} : C_{\mathbb{R}}(\Omega) \longrightarrow C_{\mathbb{R}}(\Omega)$ *is compact.*

Proof. By Lemma 5.2 it remains to show that $\mathcal{Q}$ maps a bounded set into a precompact set. To this end we are going to verify that the mapping $\mathcal{Q} : C_{\mathbb{R}}(\Omega) \longrightarrow W^{1,2}(\Omega)$ defined by (4.25) maps bounded sets into bounded sets. The compactness of the embedding $W^{1,2}(\Omega) \hookrightarrow C_{\mathbb{R}}(\Omega)$ then implies the asserted precompactness.

Using the definition (4.25), we get the estimate

$$\|\mathcal{Q}(\varphi)\|_{W^{1,2}(\Omega)} \le \left\{\|\widehat{\varphi}\|_{W^{1,2}(\Omega)} + \|\mathcal{R}[\varphi]^{-1}(C - \mathcal{N}^-[\varphi] + \mathcal{N}^+[\varphi])\|_{\overset{o}{W}{}^{1,2}(\Omega)}\right\}. \tag{5.9}$$

Since, by Lemma 4.13, for each $\varphi \in C_{\mathbb{R}}(\Omega)$ the operator $\mathcal{R}[\varphi]$ is strongly monotone with monotonicity constant $\frac{1}{\|1/\varepsilon\|_{L^\infty(\Omega)}}$, we obtain from Theorem 2.17 of [38] (see also [24]) the estimate

$$\begin{aligned}(5.10)\quad &\|\mathcal{R}[\varphi]^{-1}(C-\mathcal{N}^-[\varphi]+\mathcal{N}^+[\varphi])\|_{\overset{\circ}{W}{}^{1,2}(\Omega)}\\ &\qquad\le \|1/\varepsilon\|_{L^\infty(\Omega)}\|\mathcal{R}[\varphi](0)-(C-\mathcal{N}^-[\varphi]+\mathcal{N}^+[\varphi])\|_{W^{-1,2}(\Omega)}.\end{aligned}$$

Thus we get by (4.19)

$$\mathcal{R}[\varphi](0)=p^-[\varphi]-p^+[\varphi].$$

Hence

$$\begin{aligned}&\|\mathcal{R}[\varphi]^{-1}(C-\mathcal{N}^-[\varphi]+\mathcal{N}^+[\varphi])\|_{\overset{\circ}{W}{}^{1,2}(\Omega)}\\ &\qquad\le \|1/\varepsilon\|_{L^\infty(\Omega)}\left\|\{p^-[\varphi]-p^+[\varphi]\}-(C-\mathcal{N}^-[\varphi]+\mathcal{N}^+[\varphi])\right\|_{W^{-1,2}(\Omega)},\end{aligned}$$

which yields the estimate

$$\begin{aligned}(5.11)\quad &\|\mathcal{R}[\varphi]^{-1}(C-\mathcal{N}^-[\varphi]+\mathcal{N}^+[\varphi])\|_{\overset{\circ}{W}{}^{1,2}(\Omega)}\\ &\qquad\le \beta\left(1+\sum_{s=\pm}\left\{\|p^s[\varphi]\|_{L^1(\Omega)}+\|\mathcal{N}^s[\varphi]\|_{L^1(\Omega_q)}\right\}\right),\end{aligned}$$

where β is some positive constant depending only on the doping profile C, the permittivity function ε, and the norm of the embedding $L^1_{\mathbb{R}}(\Omega)\hookrightarrow W^{-1,2}_{\mathbb{R}}(\Omega)$.

Using the definitions (2.7), (2.8), as well as (4.20) and (4.21), we find the representations

$$p^\pm[\varphi](x)=\begin{cases}N_0^\pm e^{\mp(\widehat{\varphi}(x)\pm V_h^\pm(x))}e^{\pm\phi^\pm[\varphi](x)}, & x\in\Omega_c,\\ 0, & x\in\Omega_q,\end{cases}$$

and taking into account Lemma 4.3, we obtain the estimate

$$\|p^\pm[\varphi]\|_{L^1(\Omega_c)}\le\Gamma_c^\pm,$$

with

$$\Gamma_c^\pm:=N_0^\pm e^{\eta^\pm}\int_{\Omega_c}dx\; e^{-V_h^\pm\mp\widehat{\varphi}(x)}.$$

Hence,

$$(5.12)\qquad \|p^-[\varphi]\|_{L^1(\Omega_c)}+\|p^+[\varphi]\|_{L^1(\Omega_c)}\le\Gamma_c^-+\Gamma_c^+=:\Gamma_c.$$

By Lemma 4.4 we find the estimate

$$\begin{aligned}(5.13)\quad &\|\mathcal{N}^-[\varphi]\|_{L^1(\Omega_q)}+\|\mathcal{N}^+[\varphi]\|_{L^1(\Omega_q)}\\ &\le C^-(V^-_{max}[\varphi])\left(\gamma_1^-+\gamma_2^-\|\varphi_+\|_{L^\infty(\Omega_q)}\right)+C^+(V^+_{max}[\varphi])\left(\gamma_1^++\gamma_2^+\|\varphi_-\|_{L^\infty(\Omega_q)}\right),\end{aligned}$$

where $V^{\pm}_{max}[\varphi]$ are given by (1.10). If $\varphi \in \mathcal{B}_{C_{\mathbb{R}}(\Omega)}(r) := \{\varphi \in C_{\mathbb{R}}(\Omega) : \|\varphi\|_{L^\infty(\Omega)} \le r\}$, then $V^{\pm}_{max}[\varphi] \ge -r - c_h$, where

$$c_h := \max\{\|V_h^+\|_{L^\infty(\Omega)}, \|V_h^-\|_{L^\infty(\Omega)}\}. \tag{5.14}$$

Since the functions $C^{\pm}(\cdot)$ are nonincreasing, we find $C^{\pm}(V^{\pm}_{max}[\varphi]) \le C^{\pm}(-r - c_h)$, which yields

$$\|\mathcal{N}^-[\varphi]\|_{L^1(\Omega_q)} + \|\mathcal{N}^+[\varphi]\|_{L^1(\Omega_q)} \le \Gamma_q(r) \qquad \text{for all } \varphi \in \mathcal{B}_{C_{\mathbb{R}}(\Omega)}(r),\ r > 0, \tag{5.15}$$

where the constant $\Gamma_q(r)$ is given by

$$\Gamma_q(r) := C^-(-r - c_h)\big(\gamma_1^- + \gamma_2^- \sqrt{r}\big) + C^+(-r - c_h)\big(\gamma_1^+ + \gamma_2^+ \sqrt{r}\big).$$

Finally, one gets by (5.9)–(5.15) the estimate

$$\|\mathcal{Q}(\varphi)\|_{W^{1,2}(\Omega)} \le r_0,$$

with

$$r_0 := \|\widehat{\varphi}\|_{\overset{o}{W}{}^{1,2}(\Omega)} + \beta\Big(1 + \Gamma_c + \Gamma_q(r)\Big).$$

Hence $\mathcal{Q}(\mathcal{B}_{C_{\mathbb{R}}(\Omega)}(r)) \subseteq \mathcal{B}_{W^{1,2}(\Omega)}(r_0)$; i.e., the mapping $\mathcal{Q}$ maps bounded sets into bounded sets of $W^{1,2}(\Omega)$. Since the embedding $W^{1,2}(\Omega) \hookrightarrow C_{\mathbb{R}}(\Omega)$ is compact, we deduce the precompactness of the set $\mathcal{Q}(\mathcal{B}_{C_{\mathbb{R}}(\Omega)}(r))$, which completes the proof. $\square$

5.2. A priori estimates. Our next aim is to investigate solutions of (5.3) and to verify certain a priori estimates for them.

LEMMA 5.4. *Let Assumptions* 2.1, 3.1, *and* 4.9 *be satisfied. If* $\varphi \in W^{1,2}_{\mathbb{R}}(\Omega)$, *and if* $\zeta := \varphi - \widehat{\varphi} \in \overset{o}{W}{}^{1,2}_{\mathbb{R}}$ *satisfies* (5.3) *for some* $t \in [0, 1]$, *then there exists a constant* M *independent of* φ *and* t *such that*

$$\varphi(x) \le M \left\{1 + C^+(t\varphi_{max} - c_h)\sqrt{1 + \|\varphi_-\|_{L^\infty(\Omega_q)}}\right\}, \tag{5.16}$$

$$-\varphi(x) \le M \left\{1 + C^-(-t\varphi_{min} - c_h)\sqrt{1 + \|\varphi_+\|_{L^\infty(\Omega_q)}}\right\} \tag{5.17}$$

for all $x \in \Omega$, *where*

$$\varphi_{max} := \max\{\varphi(a), \varphi(b)\} \quad \text{and} \quad \varphi_{min} := \min\{\varphi(a), \varphi(b)\}, \tag{5.18}$$

$\varphi_\pm$ *denotes the positive and negative parts of* φ, *and* c_h *is given by* (5.14).

Proof. We set $d := \mathcal{P}^{-1}C \in \overset{o}{W}{}^{1,2}_{\mathbb{R}}(\Omega)$. Since $\zeta := \varphi - \widehat{\varphi}$ is a solution of (5.3), one has

$$\mathcal{P}(\zeta - d) = \mathcal{U}_t^+[\varphi] - \mathcal{U}_t^-[\varphi].$$

Since the right-hand side of the above equation is in $L^1(\Omega)$, one gets that $g := \zeta - d \in \overset{o}{W}{}^{1,2}_{\mathbb{R}}(\Omega)$, $\varepsilon g' \in W^{1,1}_{\mathbb{R}}(\Omega)$, and

$$-\frac{d}{dx}\varepsilon(x)\frac{d}{dx}g(x) = \mathcal{U}_t^+[\varphi](x) - \mathcal{U}_t^-[\varphi](x)$$

for a.e. $x \in \Omega$. We remark that

$$\mathcal{P}(\zeta - d) = -\frac{d}{dx}\varepsilon(x)\frac{d}{dx}g.$$

Let $\Omega_0 = (x_0, x_1) \subseteq \Omega$ be given such that $\zeta(x_0) = \zeta(x_1) = 0$ and $\zeta(x) > 0$ for $x \in \Omega_0$. We set

$$g^+(x) = \int_{x_0}^{x} dy\, \frac{1}{\varepsilon(y)} \int_{x_0}^{y} dz\, \mathcal{U}_t^+[\varphi](z), \quad x \in \Omega_0.$$

Obviously, one has

$$\frac{d}{dx}\varepsilon(x)\frac{d}{dx}h(x) = \mathcal{U}_t^-[\varphi](x)$$

for a.e. $x \in \Omega_0$, where $h(x) := g(x) + g^+(x) = \zeta(x) - d(x) + g^+(x)$, $x \in \Omega_0$. Using the maximum principle [25, Theorem 8.1] we obtain

$$\sup_{x \in \Omega_0} h(x) \leq \max\{h(x_0), h(x_1)\},$$

which yields

$$\zeta(x) \leq d(x) + \max\{-d(x_0), -d(x_1) + g^+(x_1)\}, \quad x \in \Omega_0.$$

Thus

$$\zeta(x) \leq 2\|d\|_{L^\infty(\Omega)} + g^+(x_1), \quad x \in \Omega_0. \tag{5.19}$$

For $x \in \Omega_c \cap \Omega_0$ there is, by (5.2),

$$\mathcal{U}_t^+[\varphi](x) = N_0^+ e^{\phi^+[t\varphi](x)} e^{-\widehat{\varphi}(x) - V_h^+(x)} e^{-\zeta(x)}.$$

By Lemma 4.3 and $\zeta(x) \geq 0$, $x \in \Omega_0$, we get

$$\mathcal{U}_t^+[\varphi](x) \leq N_0^+ e^{\eta^+} e^{-\widehat{\varphi}(x) - V_h^+(x)}, \quad x \in \Omega_c \cap \Omega_0,$$

which yields the estimate

$$\int_{\Omega_0 \cap \Omega_c} dz\, \mathcal{U}_t^+[\varphi](z) \leq N_0^+ e^{\eta^+} \int_{\Omega_0 \cap \Omega_c} dz\, e^{-\widehat{\varphi}(z) - V_h^+(z)}.$$

Hence, we obtain the estimate

$$\begin{aligned} g^+(x_1) &\leq \|1/\varepsilon\|_{L^1(\Omega_0)} \int_{\Omega_0} dz\, \mathcal{U}_t^+[\varphi](z) \\ &\leq \|1/\varepsilon\|_{L^1(\Omega_0)} \left(\int_{\Omega_0 \cap \Omega_q} dz\, \mathcal{U}_t^+[\varphi](z) + e^{\eta^+} N_0^+ \int_{\Omega_0 \cap \Omega_c} dz\, e^{-\widehat{\varphi}(z) - V_h^+(z)} \right). \end{aligned}$$

Inserting this estimate into (5.19), one finds positive constants M_1^+, M_2^+ independent of φ and t such that the inequality

$$\zeta(x) \leq M_1^+ + M_2^+ \int_{\Omega_q} dz\, \mathcal{U}_t^+[\varphi](z) \tag{5.20}$$

holds for $x \in \Omega_+ := \{x \in \Omega : \zeta(x) > 0\}$.

In the same manner we find positive constants M_1^-, M_2^-, which are independent of φ and t, such that

$$-\zeta(x) \le M_1^- + M_2^- \int_{\Omega_q} dz\, \mathcal{U}_t^-[\varphi](z) \tag{5.21}$$

for $x \in \Omega_- := \{x \in \Omega : \zeta(x) < 0\}$.

Since

$$\int_{\Omega_q} dz\, \mathcal{U}_t^\pm[\varphi](z) = \|\mathcal{N}^\pm[t\varphi]\|_{L^1(\Omega_q)},$$

we obtain from Lemma 4.4 and (5.20), (5.21) the existence of positive constants M^+, M^-, which do not depend on φ and t, such that

$$\zeta(x) \le M^+ \left(1 + C^+(V_{max}^+[t\varphi])\sqrt{1 + \|\varphi_-\|_{L^\infty(\Omega_q)}}\right) \tag{5.22}$$

for all $x \in \Omega_+$ and

$$-\zeta(x) \le M^- \left(1 + C^-(V_{max}^-[t\varphi])\sqrt{1 + \|\varphi_+\|_{L^\infty(\Omega_q)}}\right) \tag{5.23}$$

for all $x \in \Omega_-$.

Since $\zeta(x) \le 0$ for $x \in \Omega \setminus \Omega_+$ and $-\zeta(x) \le 0$ for $x \in \Omega \setminus \Omega_-$, we obtain from (5.22) and (5.23) that in fact these relations are valid for each $x \in \Omega$.

Taking into account the estimates $V_{max}^+[t\varphi] \ge t\varphi_{max} - c_h$, $V_{max}^-[t\varphi] \ge -t\varphi_{min} - c_h$ (cf. (5.14) and (5.18)), we get—since $C^\pm(\cdot)$ is nonincreasing—the estimates

$$C^+(V^+[t\varphi]) \le C^\pm(t\varphi_{max} - c_h) \quad \text{and} \quad C^-(V^-[t\varphi]) \le C^\pm(-t\varphi_{min} - c_h).$$

Finally, using the estimates $\pm\varphi(x) \le \pm\zeta(x) + \|\widehat{\varphi}\|_{L^\infty(\Omega)}$, $x \in \Omega$, and $t \in [0,1]$, we immediately obtain from (5.22) and (5.23) the estimates (5.16), (5.17) for some constant M. ∎

COROLLARY 5.5. *Let Assumptions* 2.1, 3.1, *and* 4.9 *be satisfied. If* $\varphi \in W_{\mathbb{R}}^{1,2}(\Omega)$ *and if* $\zeta := \varphi - \widehat{\varphi} \in \overset{o}{W}{}_{\mathbb{R}}^{1,2}(\Omega)$ *satisfies* (5.3) *for some* $t \in [0,1]$, *then*

$$\|\varphi_+\|_{L^\infty(\Omega_q)} \le M \left(1 + C^+(t\varphi_{max} - c_h)\sqrt{1 + \|\varphi_-\|_{L^\infty(\Omega_q)}}\right) \tag{5.24}$$

and

$$\|\varphi_-\|_{L^\infty(\Omega_q)} \le M \left(1 + C^-(-t\varphi_{max} - c_h)\sqrt{\|1 + \|\varphi_+\|_{L^\infty(\Omega_q)}}\right). \tag{5.25}$$

Proof. Since $\varphi_{min} \le \varphi_{max}$, one has $-t\varphi_{max} \le -t\varphi_{min}$, $t \in [0,1]$. Taking into account the fact that the functions $C^\pm(\cdot)$ are nonincreasing, we obtain from Lemma 5.4 the estimates (5.24) and (5.25). ∎

5.3. Main theorem. Using Corollary 5.5, we aim to show that all solutions of (5.3) are included in a uniform ball; that is, there is an $r_0 > 0$ such that

$$\mathfrak{L} := \left\{\vartheta \in C_{\mathbb{R}}(\Omega) : \vartheta = t\mathcal{Q}_\infty(\vartheta), \quad t \in [0,1]\right\} \subseteq \mathcal{B}_{C_{\mathbb{R}}(\Omega)}(r_0). \tag{5.26}$$

For this last step we need the following additional balance condition.

ASSUMPTION 5.6 (balance condition). *Let the distribution functions $f^{\pm}$ satisfy the assumption* (Q.5). *We say that the distribution functions $f^{\pm}$ obey the balance condition if*

$$G(x,y) := \sup_{s \geq 0} \left\{ D^{+}(s+x) D^{-}(-s+y)^{1/2} + D^{+}(-s+x)^{1/2} D^{-}(s+y) \right\} < \infty$$

for $x, y \in \mathbb{R}$, where $D^{\pm}(\cdot)$ are defined by (3.1).

Remark 5.7. The Boltzmann and the Fermi–Dirac distributions satisfy the balance condition; see Remark 3.2.

Indeed, let $f^{\pm}(\cdot)$ be the Boltzmann distribution function, i.e., $f^{\pm}(\lambda) = e^{-\lambda}$, $\lambda \in \mathbb{R}$. Using the definition (3.1), we get that $D^{\pm}(\lambda) = e^{-\lambda}\sqrt{1+\lambda^2}$, $\lambda \in \mathbb{R}$. A straightforward computation shows that

$$G(x,y) = \sup_{s \geq 0} \left\{ e^{-\frac{s}{2} - x - y} g(x,y,s) \right\}$$

for $x, y \in \mathbb{R}$, where

$$g(x,y,s) := \left(\sqrt{1+(s+x)^2}\sqrt{1+(s-y)^2} + \sqrt{1+(s-x)^2}\sqrt{1+(s+y)^2} \right).$$

Obviously, one has $G(x,y) < \infty$ for $x, y \in \mathbb{R}$. Hence, the balance condition holds.

The verification of the balance condition for Fermi–Dirac distribution functions is easier than for Boltzmann distributions since their growth at minus infinity is linear and not exponential, as for Boltzmann distributions.

THEOREM 5.8. *Let Assumptions* 2.1, 3.1, *and* 4.9 *be satisfied. If the balance condition, i.e., Assumption* 5.6, *is valid, then for any choice of the approximation parameters $\mathfrak{A} = \{\delta_0^{\pm}, \delta^{\pm}\}$, $0 < \delta_0^{\pm} < \delta^{\pm} \leq \infty$,*

(i) *a solution $\varphi \in W^{1,2}_{\mathbb{R}}(\Omega)$ of dissipative hybrid model in the sense of Definition* 4.12 *exists, and*

(ii) *there is an $r_0 \in (0,\infty)$ independent of the approximation parameters $\{\delta_0^{\pm}, \delta^{\pm}\}$ such that any solution $\varphi \in W^{1,2}_{\mathbb{R}}(\Omega)$ of the dissipative hybrid model obeys $\|\varphi\|_{L^\infty(\Omega)} \leq r_0$.*

The corresponding current densities $J^{\pm}[\varphi]$ of a solution $\varphi \in W^{1,2}_{\mathbb{R}}(\Omega)$ of the dissipative hybrid model are different from zero if and only if the boundary values of the quasi-Fermi potentials are different, i.e., $\phi^{\pm}_{a_0} \neq \phi^{\pm}_{b_0}$, provided the distribution functions $f^{\pm}(\cdot)$ are strictly decreasing.

Proof. To prove (i) it is enough to show that $\mathcal{Q} : C_{\mathbb{R}}(\Omega) \longrightarrow C_{\mathbb{R}}(\Omega)$ has a fixed point; see Proposition 4.15. To prove this we use the Leray–Schauder fixed point theorem. Since by Lemmas 5.2 and 5.3 the mapping $\mathcal{Q}$ is continuous and compact, it remains to show that the set $\mathfrak{L}$ defined by (5.26) is uniformly bounded in $t \in [0,1]$. If $\varphi \in \mathfrak{L}$, then by Lemma 5.1 it satisfies (5.3). If $\varphi \in \mathfrak{L}$ satisfies (5.3), then the estimates of Corollary 5.5 hold.

Let us assume that $\varphi_{max} \geq 0$. Using the estimate $1 + \|\varphi_{\pm}\|^{1/2}_{L^\infty(\Omega_q)} \leq \sqrt{2}\left(1 + \|\varphi_{\pm}\|_{L^\infty(\Omega_q)}\right)^{1/2}$, we obtain from Corollary 5.5 the estimates

$$1 + \|\varphi_{+}\|_{L^\infty(\Omega_q)} \leq r_1 \left(1 + C^{+}\sqrt{1 + \|\varphi_{-}\|_{L^\infty(\Omega_q)}}\right), \tag{5.27}$$

$$1 + \|\varphi_{-}\|_{L^\infty(\Omega_q)} \leq r_1 \left(1 + C^{-}\sqrt{1 + \|\varphi_{+}\|_{L^\infty(\Omega_q)}}\right), \tag{5.28}$$

where r_1 is a positive constant independent of $\varphi \in \mathfrak{L}$. Moreover, we used abbreviations $C^{\pm} := C^{\pm}(\pm t\varphi_{max} - c_h)$. Inserting (5.28) into (5.27), we find a positive constant r_2, which does not depend on φ and t, such that

$$\|\varphi_+\|_{L^\infty(\Omega_q)} \leq r_2 \left(1 + C^+\sqrt{C^-}\right)^{4/3},$$

where we used the fact that $C^+ \leq C^+(-c_h)$, since $\varphi_{max} \geq 0$.

Further, by definition (4.14) we get

$$\begin{aligned} C^+\sqrt{C^-} &= C^+(t\varphi_{max} - c_h)\sqrt{C^-(-t\varphi_{max} - c_h)} \\ &= (1+\eta^+)\sqrt{1+\eta^-}D^+(t\varphi_{max} - c_h - \eta^+)D^-(-t\varphi_{max} - c_h - \eta^-)^{1/2}. \end{aligned}$$

Taking into account the balance condition (Assumption 5.6) and $\varphi_{max} \geq 0$, we get

$$C^+\sqrt{C^-} \leq (1+\eta^+)\sqrt{1+\eta^-}G(-c_h - \eta^+, -c_h - \eta^-),$$

which shows that there is a positive constant r_3 independent of $\varphi \in \mathfrak{L}$ and $t \in [0,1]$ such that

$$\|\varphi_+\|_{L^\infty(\Omega_q)} \leq r_3.$$

Since $0 \leq \varphi_{max} \leq \|\varphi_+\|_{L^\infty(\Omega_q)}$, we have $-t\varphi_{max} \geq -t\|\varphi_+\|_{L^\infty(\Omega_q)} \geq -\|\varphi_+\|_{L^\infty(\Omega_q)} \geq -r_3$, $t \in [0,1]$. By the monotonicity of $C^-(\cdot)$ we obtain $C^-(-t\varphi_{max} - c_h) \leq C^-(-r_3)$. Using Corollary 5.5, we finally get the existence of a positive constant r_4, which does not depend on $\varphi \in \mathfrak{L}$ and t, such that

$$\|\varphi_-\|_{L^\infty(\Omega_q)} \leq r_4.$$

Hence $\|\varphi\|_{L^\infty(\Omega_a)} \leq r_+ := \max\{r_3, r_4\}$, which shows that the $\mathfrak{L}_q := \{\varphi \restriction \Omega_q : \varphi \in \mathfrak{L}\} \subseteq \mathcal{B}_{C_{\mathbb{R}}(\Omega_q)}(r_+)$, provided $\varphi_{max} \geq 0$.

In a similar manner one shows that there exists a positive constant r_- independent of $t \in [0,1]$ such that $\mathfrak{L}_q := \{\varphi \restriction \Omega_q : \varphi \in \mathfrak{L}\} \subseteq \mathcal{B}_{C_{\mathbb{R}}(\Omega_q)}(r_-)$, provided $\varphi_{max} \leq 0$.

Thus, we get that $\mathfrak{L}_q \subseteq \mathcal{B}_{C_{\mathbb{R}}(\Omega_q)}(r_q)$, $r_q := \max\{r_+, r_-\}$. In particular, we have $-r_q \leq \varphi_{min} \leq \varphi_{max} \leq r_q$. Using Lemma 5.4, we find

$$\varphi(x) \leq M\left(1 + C^+(-r_q - c_h)\left(1 + r_q^{1/2}\right)\right) =: r_5$$

and

$$-\varphi(x) \leq M\left(1 + C^-(-r_q - c_h)\left(1 + r_q^{1/2}\right)\right) =: r_6$$

for all $x \in \Omega$. Setting $r_0 = \max\{r_5, r_6\}$, we conclude that $\|\varphi\|_{L^\infty(\Omega)} \leq r_0$. Hence, $\mathfrak{L} \subseteq B_{C_{\mathbb{R}}(\Omega)}(r_0)$. Since r_0 depends only on quantities entering Assumptions 2.1, 3.1, 4.9, and 5.6, but is independent of $t \in [0,1]$, the uniform boundedness of the set $\mathfrak{L}$ is verified. Hence the Leray–Schauder fixed point theorem implies the existence of a solution of a dissipative hybrid model. Assertion (ii) is verified.

To prove the last assertion, we note that in accordance with (3.7) the current density for holes $J^+[\varphi]$ satisfies the equation

$$J^+[\varphi] = \frac{1}{2\pi}\int_{\Lambda^+[\varphi]} d\lambda\, t^+[\varphi](\lambda)\left\{f^+(\lambda - \phi^+[\varphi](a)) - f^+(\lambda - \phi^+[\varphi](b))\right\},$$

where $t^+[\varphi](\lambda)$ is the transmission coefficient. Notice that $t^+[\varphi](\lambda) \geq 0$ for a.e. $\lambda \in \mathbb{R}$. If $J^+[\varphi] \geq 0$, then $-\phi^+[\varphi](x)$, $x \in \Omega_q$, is nondecreasing such that $-\phi^+[\varphi](a) \leq -\phi^+[\varphi](b)$. Hence

$$f^+(\lambda - \phi^+[\varphi](a)) - f^+(\lambda - \phi^+[\varphi](b)) \geq 0.$$

If $J^+[\varphi] = 0$, then

$$f^+(\lambda - \phi^+[\varphi](a)) - f^+(\lambda - \phi^+[\varphi](b)) = 0$$

for a.e. $\lambda \in \Lambda^+[\varphi]$. Since the distribution functions $f^\pm(\cdot)$ are strictly decreasing, one gets $\phi^+[\varphi](a) = \phi^+[\varphi](b)$. If $J^+[\varphi] = 0$, then $\phi^+[\varphi](a_0) = \phi^+[\varphi](a)$ and $\phi^+[\varphi](b) = \phi^+[\varphi](b_0)$, which yields $\phi^+_{a_0} = \phi^+[\varphi](a_0) = \phi^+[\varphi](b_0) = \phi^+_{b_0}$. We proceed similarly if $J^+[\varphi] \leq 0$. Conversely, if $\phi^+_{a_0} = \phi^+_{b_0}$ and $J^+[\varphi] \geq 0$, then $-\phi^+_{a_0} = -\phi^+[\varphi](a_0) \leq -\phi^+[\varphi](a) \leq -\phi^+_b[\varphi](b) \leq -\phi^+[\varphi](b_0) = -\phi^+_{b_0} = -\phi^+_{a_0}$, which yields $J^+[\varphi] = 0$. We act similarly if $J^+[\varphi] \leq 0$. The proof for electrons is similar. $\square$

6. Comments. We analyzed a dissipative hybrid model, which consists of a coupled drift-diffusion model and dissipative Schrödinger operators. The electrostatic potential is determined by a Poisson equation on the whole device domain. We showed that the coupled system is well posed and always admits a solution, provided Assumptions 2.1, 3.1, 4.9, and 5.6 are satisfied; see Theorem 5.8. The proof is based on a Leary–Schauder fixed point argument, which does not give uniqueness of solution. In fact, uniqueness is in general not expected.

Let us comment on the results as follows:

1. The dissipative Schrödinger model considered in section 3 is to some extent artificial since the physical interpretation of the generalized eigenfunctions $\psi^\pm_a$, $\psi^\pm_b$ of the minimal self-adjoint dilations $K^\pm$ is not a priori clear. However, as has been shown in [4], for fixed energy there are two channels which fit into the picture of the more commonly used quantum transmitting Schrödinger equation (B.1); see (B.8) and (B.9). Using this fact, one can consider the dissipative Schrödinger model as an approximation by means of a proper choice of the distribution functions for the dissipative system. The approximation we consider here is outlined in section B.2 and depends on the approximation parameters $\mathfrak{A} = \{\delta_0, \delta\}$; see (B.10) and (B.11). We note that other approximations can be chosen, which then leads to a different dissipative hybrid model.
   Moreover, the dissipative Schrödinger model allows us to use the Lax–Philip scattering techniques to prove the continuity of the carrier and current density operator; see section A.2. The Lax–Phillips scattering theory is simpler than that of the Schrödinger operators corresponding to the Schrödinger operator (B.1); see also [4].
   In the dissipative approximation presented in section B.2, we considered only wave functions with energy larger than V_{max} since they are the current carrying states. The inclusion of the states with energies smaller than V_{max} is so far an open problem.
2. The drift-diffusion model we consider here is based on Boltzmann statistics (see (1.1)–(1.3)), which leads to explicit expressions for the electrochemical potentials $\phi^\pm$; see (2.7) and (2.8). The case of Fermi–Dirac statistics may be considered in the same fashion. However, one loses the explicit expressions.

3. Here we treated the bipolar case, where electrons and holes interact via the Poisson equation only. To prove the existence of solutions, the statistical distribution functions $f^{\pm}$ have to satisfy the balance condition (Assumption 5.6), which is satisfied for the common distribution functions; see Remark 5.7. The unipolar case, where only one particle species is considered, is even simpler. In this case the balance condition is redundant, and the existence of a solution is easily obtained by the estimate (3.10) of the density and applying the maximum principle to Poisson's equation.
4. Theorem 5.8 shows that the solutions of a dissipative hybrid model are contained in a ball with radius r_0. Following the proof of the theorem, one can derive this radius explicitly.
5. In Theorem 5.8 we showed that the current densities of the dissipative hybrid model are equal to zero if and only if $\phi^{\pm}_{a_0} = \phi^{\pm}_{b_0}$, i.e., in thermal equilibrium. In this case the electrochemical potentials $\phi^{\pm}$ are constant, and thus the statistical operators $\varrho^{\pm}$ are given by functions of the corresponding self-adjoint dilation $K^{\pm}$, i.e., $\varrho^{\pm} = \mathfrak{f}^{\pm}(K^{\pm})$ with $\mathfrak{f}^{\pm}(\lambda) = f^{\pm}(\lambda \mp \phi^{\pm})\chi_{\Lambda^{\pm}}(\lambda)$, where $\phi^{\pm}$ are the constant electrochemical potentials. One may expect that the solution of the dissipative hybrid model in thermal equilibrium is unique. However, this is so far an open problem.

Appendix A. Dissipative Schrödinger systems. Let us give a short introduction to the theory of dissipative Schrödinger systems; for details see [3, 5, 35, 33]. We start with some facts on Schrödinger-type operators.

A.1. Schrödinger-type operators. Let the conditions $0 < m \in L^{\infty}_{\mathbb{R}}(\Omega_q)$, $\frac{1}{m} \in L^{\infty}_{\mathbb{R}}(\Omega_q)$, be satisfied, where $\Omega_q = (a,b)$. Moreover, let $V \in L^{\infty}_{\mathbb{R}}(\Omega_q)$ and $\varkappa_a, \varkappa_b \in \mathbb{C}_+$, where $\mathbb{C}_+ := \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$, be given. The Schrödinger-type operator H is defined by

$$(Hg)(x) := -\frac{1}{2}\frac{d}{dx}\frac{1}{m}\frac{d}{dx}g(x) + V(x)g(x), \quad x \in \Omega_q, \quad g \in \mathrm{dom}(H), \tag{A.1}$$

where its domain is given by

$$\mathrm{dom}(H) := \left\{ g \in W^{1,2}(\Omega_q) : \begin{array}{c} \frac{1}{m}g' \in W^{1,2}(\Omega_q), \\ \frac{1}{2m(a)}g'(a) = -\varkappa_a g(a), \\ \frac{1}{2m(b)}g'(b) = \varkappa_b g(b) \end{array} \right\}. \tag{A.2}$$

The operator H is maximal dissipative and completely non–self-adjoint on the Hilbert space $L^2(\Omega_q)$; that is, the operator does not possess self-adjoint parts. Its spectrum consists only of discrete eigenvalues in the lower half-plane [34].

In the following we are going to prove some continuity results for the operator H with respect to the potential V and the boundary coefficients $\varkappa_a$, $\varkappa_b$. To this end we introduce

$$\mathcal{T} := \mathbb{C}_+ \times \mathbb{C}_+ \times L^{\infty}(\Omega_q),$$

and write—if needed—$H[\tau]$ for $\tau = (\varkappa_a, \varkappa_b, V) \in \mathcal{T}$ to indicate the dependence of the maximal dissipative operator on the potential V and the boundary coefficients $\varkappa_a$, $\varkappa_b$. Moreover, for $\tau, \tau_n \in \mathcal{T}$, $n \in \mathbb{N}$, we write $\lim_{n\to\infty} \tau_n = \tau$ if

$$\lim_{n\to\infty} \left(|\varkappa_{a,n} - \varkappa_a| + |\varkappa_{b,n} - \varkappa_b| + \|V_n - V\|_{L^{\infty}(\Omega)} \right) = 0. \tag{A.3}$$

Remark A.1. Let the electrostatic potential $\varphi \in C_{\mathbb{R}}(\Omega)$ be given and assume that the band-edge offset potentials $V_h^{\pm}$ satisfy Assumption 3.1 (Q.4). Then the triples $\tau^{\pm} = (\varkappa_a^{\pm}, \varkappa_b^{\pm}, V^{\pm})$, where $\varkappa_a^{\pm}$, $\varkappa_b^{\pm}$ are given by (1.8) and $V^{\pm}$ by (1.4), are well defined, and there is $\tau^{\pm} \in \mathcal{T}$. Thus the operators $H^{\pm}$ defined by (3.3), (3.4) are of the form (A.1), (A.2). Moreover, $\varphi, \varphi_n \in C_{\mathbb{R}}(\Omega)$, $n \in \mathbb{N}$, with $\lim_{n\to\infty} \|\varphi_n - \varphi\|_{L^\infty(\Omega)} = 0$ implies $\lim_{n\to\infty} \tau_n^{\pm} = \tau^{\pm}$, where

$$\tau_n^{\pm} := (\varkappa_a^{\pm}[\varphi_n], \varkappa_b^{\pm}[\varphi_n], V^{\pm}[\varphi_n]) \quad \text{and} \quad \tau^{\pm} := (\varkappa_a^{\pm}[\varphi], \varkappa_b^{\pm}[\varphi], V^{\pm}[\varphi]),$$

where $\varkappa_a^{\pm}[\varphi]$, $\varkappa_b^{\pm}[\varphi]$, $V^{\pm}[\varphi]$ indicates the dependence of the coupling constants $\varkappa_a^{\pm}$, $\varkappa_b^{\pm}$ and the potential $V^{\pm}$ on the electrostatic potential φ; cf. (1.8) and (1.4).

A.2. Dilation and Lax–Phillips scattering. The constants $\varkappa_a, \varkappa_b \in \mathbb{C}_+$ can be written as

$$\varkappa_a = q_a + i\frac{\alpha_a^2}{2} \quad \text{and} \quad \varkappa_b = q_b + i\frac{\alpha_b^2}{2}, \tag{A.4}$$

where $\alpha_a, \alpha_b > 0$.

Since the operator H is maximal dissipative, it admits a minimal self-adjoint dilation K on some dilation space $\mathfrak{K}$; see [21]. We choose the dilation space $\mathfrak{K}$,

$$\mathfrak{K} := L^2(\mathbb{R}_-, \mathbb{C}^2) \oplus L^2(\Omega_q) \oplus L^2(\mathbb{R}_+, \mathbb{C}^2). \tag{A.5}$$

To describe the minimal dilation K in $\mathfrak{K}$ we set

$$\vec{g} := g_- \oplus g \oplus g_+, \tag{A.6}$$

where $g \in L^2(\Omega_q)$, and where

$$g_-(x) := \begin{pmatrix} g_-^b(x) \\ g_-^a(x) \end{pmatrix} \in L^2(\mathbb{R}_-, \mathbb{C}^2) \quad \text{and} \quad g_+(x) := \begin{pmatrix} g_+^b(x) \\ g_+^a(x) \end{pmatrix} \in L^2(\mathbb{R}_+, \mathbb{C}^2) \tag{A.7}$$

for $x \in \mathbb{R}_-$ and $x \in \mathbb{R}_+$, respectively.

THEOREM A.2. *Let $(\varkappa_a, \varkappa_b, V) \in \mathcal{T}$ be given. Then the operator K defined by*

$$\mathrm{dom}(K) := \left\{ \vec{g} \in \mathcal{K} : \begin{array}{c} g_\pm \in W^{1,2}(\mathbb{R}_\pm, \mathbb{C}^2),\ g, \frac{1}{m}g' \in W^{1,2}(\Omega_q), \\ \frac{1}{2m(b)}g'(b) - q_b g(b) = \alpha_b \frac{g_-^b(0) + g_+^b(0)}{2}, \\ i\alpha_b g(b) = g_+^b(0) - g_-^b(0), \\ \frac{1}{2m(a)}g'(a) + q_a g(a) = \alpha_a \frac{g_-^a(0) + g_+^a(0)}{2}, \\ i\alpha_a g(a) = g_-^a(0) - g_+^a(0) \end{array} \right\} \tag{A.8}$$

and

$$K\vec{g} := -i\frac{d}{dx}g_- \oplus \left(-\frac{1}{2}\frac{d}{dx}\frac{1}{m}\frac{d}{dx} + V \right) g \oplus -i\frac{d}{dx}g_+, \qquad \vec{g} \in \mathrm{dom}(K), \tag{A.9}$$

is self-adjoint.

Proof. The proof is given in [35, Theorem 4.1]. □

The operator K is a minimal self-adjoint dilation of H, i.e.,

$$(H - z)^{-1}\psi = P(K - z)^{-1}\psi \qquad \text{for all } \psi \in L^2(\Omega_q) \text{ and } z \in \mathbb{C}_+, \tag{A.10}$$

and

$$\mathfrak{K} = \operatorname*{clospan}_{z\in\mathbb{C}\setminus\mathbb{R}} \left\{(K-z)^{-1}\psi,\ \psi \in L^2(\Omega_q)\right\}, \tag{A.11}$$

where P is the orthogonal projection from the dilation space $\mathfrak{K}$ onto the subspace $L^2(\Omega_q)$.

We introduce the unclosed operator $\alpha : L^2(\Omega_q) \to \mathbb{C}^2$,

$$\alpha f = \begin{pmatrix} \alpha_b f(b) \\ -\alpha_a f(a) \end{pmatrix}, \qquad f \in \operatorname{dom}(\alpha) = W^{1,2}(\Omega_q), \tag{A.12}$$

and the boundary operators $T : \operatorname{res}(H) \longrightarrow \mathcal{B}(L^2(\Omega_q), \mathbb{C}^2)$, $T_* : \operatorname{res}(H^*) \longrightarrow \mathcal{B}(L^2(\Omega_q), \mathbb{C}^2)$ defined by

$$\begin{aligned} T(z)g &:= \alpha(H-z)^{-1}g, \quad g \in L^2(\Omega_q), \\ T_*(z)g &:= \alpha(H^*-z)^{-1}g, \quad g \in L^2(\Omega_q). \end{aligned} \tag{A.13}$$

The 2×2 matrix-valued function $\Theta : \operatorname{res}(H^*) \longrightarrow \mathcal{B}(\mathbb{C}^2)$ defined by

$$\Theta(z) = I_{\mathbb{C}^2} - i\alpha T(\overline{z})^* \tag{A.14}$$

is holomorphic, contractive on $\mathbb{C}_-$, and unitary on $\mathbb{R}$. The matrix-valued Θ is called the characteristic function of H; see [21, 35] for motivation and details of this definition.

The incoming generalized eigenfunctions of K, i.e., functions $\vec{\psi}_a = \psi_{a,-} \oplus \psi_a \oplus \psi_{a,+}$ and $\vec{\psi}_b = \psi_{b,-} \oplus \psi_b \oplus \psi_{b,+}$, which satisfy

$$\begin{aligned} -i\frac{d}{dx}\psi_{\nu,-}(\lambda,x) \oplus \left(-\frac{1}{2}\frac{d}{dx}\frac{1}{m}\frac{d}{dx} + V(x)\right)\psi_\nu(,\lambda,x) \oplus -i\frac{d}{dx}\psi_{\nu,+}(\lambda,x) \\ = \lambda\Big(\psi_{\nu,-}(\lambda,x) \oplus \psi_\nu(\lambda,x) \oplus \psi_{\nu,+}(\lambda,x)\Big), \qquad \lambda \in \mathbb{R},\ \nu = a,b, \end{aligned}$$

and the boundary conditions in (A.8), are given by

$$\begin{aligned} \vec{\psi}_\nu(\lambda,x) &= \psi_{\nu,-} \oplus \psi_\nu \oplus \psi_{\nu,+} \\ &= \frac{1}{\sqrt{2\pi}}\Big(e^{i\lambda x}e_\nu \oplus \big(T_*(\lambda)^* e_\nu\big)(x) \oplus e^{i\lambda x}\Theta(\lambda)e_\nu\Big), \qquad \lambda \in \mathbb{R},\ \nu = a,b, \end{aligned} \tag{A.15}$$

$e_b = \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)$, $e_a = \left(\begin{smallmatrix}0\\1\end{smallmatrix}\right)$; see [35, 5].

The incoming Fourier transform $\Phi_- : \mathfrak{K} \longrightarrow \widehat{\mathfrak{K}_0} := L^2(\mathbb{R};\mathbb{C}^2)$ is then defined by

$$\begin{aligned} \Big(\Phi_- \vec{f}\Big)(\lambda) := &\begin{pmatrix} \int_{-\infty}^0 dx\ \left(f_-(x), \psi_{b,-}(\lambda,x)\right)_{\mathbb{C}^2} \\ \int_{-\infty}^0 dx\ \left(f_-(x), \psi_{a,-}(\lambda,x)\right)_{\mathbb{C}^2} \end{pmatrix} \\ &+ \begin{pmatrix} \int_a^b dx\ \overline{\psi_b(\lambda,x)}\, f(x) \\ \int_a^b dx\ \overline{\psi_a(\lambda,x)}\, f(x) \end{pmatrix} + \begin{pmatrix} \int_0^\infty dx\ \left(f_+(x), \psi_{b,+}(\lambda,x)\right)_{\mathbb{C}^2} \\ \int_0^\infty dx\ \left(f_+(x), \psi_{a,+}(\lambda,x)\right)_{\mathbb{C}^2} \end{pmatrix}, \end{aligned}$$

$\lambda \in \mathbb{R}$, for all $\vec{f} = f_- \oplus f \oplus f_+ \in \mathfrak{K}$. The incoming Fourier transform Φ_- establishes a unitary equivalence between the dilation K and the multiplication operator M with the independent variable λ; see [35, 5].

We introduce the identification operators $J_\pm : \mathfrak{K}_0 \longrightarrow \mathfrak{K}$, $\mathfrak{K}_0 = L^2(\mathbb{R};\mathbb{C}^2)$,

$$\begin{aligned} J_-\vec{f} &:= P_-\vec{f} \oplus 0 \oplus 0, \\ J_+\vec{f} &:= 0 \oplus 0 \oplus P_+\vec{f}, \end{aligned} \qquad \vec{f} \in \mathfrak{K}_0, \tag{A.16}$$

where $P_\pm$ denote the orthogonal projections from $\mathfrak{K}_0$ onto $L^2(\mathbb{R}_\pm, \mathbb{C}^2)$. Let K_0 be the differentiation operator $K_0 = -i\frac{d}{dx}$, defined on $\mathfrak{K}_0$. The Lax–Phillips wave operators,

$$W_\pm := s - \lim_{t\to\pm\infty} e^{itK} J_\pm e^{-itK_0}, \tag{A.17}$$

always exist (see [3, 33]) and are unitary.

By $\mathcal{F} : \mathfrak{K}_0 \longrightarrow \widehat{\mathfrak{K}}_0 = L^2(\mathbb{R}, \mathbb{C}^2)$ we denote the usual Fourier transform

$$(\mathcal{F}\vec{f})(\lambda) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} dx\, e^{-ix\lambda} \vec{f}(x), \quad \vec{f} \in \mathfrak{K}_0, \quad \lambda \in \mathbb{R}. \tag{A.18}$$

The incoming Fourier transform Φ_- of K is then expressed by

$$\Phi_- = \mathcal{F}W_-^* : \mathfrak{K} \longrightarrow \widehat{\mathfrak{K}}_0 \tag{A.19}$$

(see [3, 35, 33]).

The Lax–Phillips scattering operator $S : \mathfrak{K}_0 \longrightarrow \mathfrak{K}_0$ is defined by

$$S := W_+^* W_-. \tag{A.20}$$

The scattering operator commutes with K_0, which yields that the operator $\widehat{S} : \widehat{\mathfrak{K}}_0 \longrightarrow \widehat{\mathfrak{K}}_0$,

$$\widehat{S} := \mathcal{F} S \mathcal{F}^*, \tag{A.21}$$

commutes with M. Hence the operator $\widehat{S}$ can be represented as a multiplication operator with a 2×2 matrix-valued function $\{\widehat{S}(\lambda)\}_{\lambda\in\mathbb{R}}$, which is called the scattering matrix, in particular, the Lax–Phillips scattering matrix. It turns out that the scattering matrix can be computed directly from the operator H by means of the characteristic function Θ. In fact, there is

$$S(\lambda) = \Theta(\lambda)^* \tag{A.22}$$

for a.e. $\lambda \in \mathbb{R}$.

The transmission coefficient $t(\lambda)$, $\lambda \in \mathbb{R}$, is then given by

$$t(\lambda) := |(S(\lambda)e_b, e_a)_{\mathbb{C}^2}|^2 = |(e_b, \Theta(\lambda)e_a)_{\mathbb{C}^2}|^2, \tag{A.23}$$

where $e_b, e_a \in \mathbb{C}$ are given by $e_b = \binom{1}{0}$, $e_a = \binom{0}{1}$. Moreover, there is

$$t(\lambda) = |(S(\lambda)e_a, e_b)_{\mathbb{C}^2}|^2$$

(see [35]).

Clearly, all the quantities defined above depend on the triple $\tau = (\varkappa_a, \varkappa_b, V) \in \mathcal{T}$. If needed, we will denote the dependence of any quantity A on $\tau \in \mathcal{T}$ by writing $A[\tau]$.

A.3. Carrier density operator. Let $\tau \in \mathcal{T}$ be given and H be the associated maximal dissipative Schrödinger operator. By means of a so-called density matrix ρ, one can assign a particle density u to each operator H in a unique way. The mapping $\tau \mapsto u$ is then defined as the carrier density operator. In what follows we will discuss this procedure in detail.

A density matrix is an element of the Banach space $L^\infty(\mathbb{R}, \mathcal{B}(\mathbb{C}^2))$ such that its values are self-adjoint and nonnegative 2×2 matrices; i.e., for a.e. $\lambda \in \mathbb{R}$ there is $\rho(\lambda) = \rho(\lambda)^*$ and $\rho(\lambda) \geq 0$. With ρ one associates a bounded multiplication operator $\widehat{\rho} : \widehat{\mathfrak{K}}_0 \longrightarrow \widehat{\mathfrak{K}}_0$ on the Hilbert space $\mathfrak{K}_0 = L^2(\mathbb{R}, \mathbb{C}^2)$ defined by

$$(\widehat{\rho}\vec{f})(\lambda) := \rho(\lambda)\vec{f}(\lambda). \tag{A.24}$$

Using the transformation (A.19), one defines by

$$\varrho := \Phi_-^* \widehat{\rho} \Phi_- \tag{A.25}$$

an operator on the dilation space $\mathfrak{K}$, which is self-adjoint, is nonnegative, and commutes with the dilation K. The operator ϱ is called a density operator or a steady state albeit that ϱ is not a trace class operator. However, it turns out that if the additional condition

$$C_\rho := \sup_{\lambda \in \mathbb{R}} \sqrt{\lambda^2 + 1} \|\rho(\lambda)\|_{\mathcal{B}(\mathbb{C}^2)} < \infty \tag{A.26}$$

is satisfied, then the product ϱP (where P denotes again the projection from $\mathfrak{K}$ onto $L^2(\Omega_q)$) belongs to the trace class; cf. [33]. Using this observation, the following is shown in [3, 35]: for a fixed density matrix ρ satisfying (A.26) and given $\tau \in \mathcal{T}$, there is exactly one function $u \in L^1(\Omega_q)$ satisfying

$$\operatorname{tr}(\varrho M(h)) = \int_{\Omega_q} dx\, u(x)h(x) \quad \text{for all } h \in L^\infty_{\mathbb{R}}(\Omega_q), \tag{A.27}$$

where $M(h)$ is the multiplication operator defined by

$$M(h)\vec{g} := 0 \oplus hg \oplus 0, \quad \vec{g} = g_- \oplus g \oplus g_+ \in \mathfrak{K}, \quad (hg)(x) = h(x)g(x), \quad x \in \Omega_q$$

(see [35]). We note that condition (A.26) implies $\varrho M(h) \in \mathcal{B}_1(\mathfrak{K})$ for each $h \in L^\infty_{\mathbb{R}}(\Omega_q)$.

Remark A.3. Let $f^\pm$ be distribution functions obeying Assumption 3.1 (Q.5). We define the density matrices $\rho^\pm$ by

$$\rho^\pm(\lambda) = \begin{pmatrix} f^\pm(\lambda \mp \epsilon_b) & 0 \\ 0 & f^\pm(\lambda \mp \epsilon_a) \end{pmatrix} \chi_{\Lambda^\pm}(\lambda),$$

where ϵ_a, ϵ_b are given Fermi levels and $\chi_{\Lambda^\pm}$ denotes the characteristic function of the sets $\Lambda^\pm$; see (1.13). Clearly $\rho^\pm$ satisfy the condition (A.26). In [33] it is shown that the densities $u^\pm$ corresponding to $\rho^\pm$ and $\tau^\pm$, where $\tau^\pm$ are given as in Remark A.1, defined by means of the relation (A.27) coincide with (3.5).

The mapping that assigns the density u to each triple $\tau \in \mathcal{T}$, if the density matrix ρ is given, is called the carrier density operator $\mathcal{N}_\rho$; i.e., $\mathcal{N}_\rho : \mathcal{T} \longrightarrow L^1(\Omega_q)$ is defined by

$$\int_{\Omega_q} dx\, \mathcal{N}_\rho[\tau](x)h(x) = \operatorname{tr}(\varrho[\tau]M(h)) \qquad \text{for all } h \in L^\infty(\Omega_q) \text{ and each } \tau \in \mathcal{T},$$

where $\varrho[\tau]$ indicates the dependence of ϱ given by (A.25) on τ.

PROPOSITION A.4. *Let $\tau = \{\varkappa_a, \varkappa_b, V\} \in \mathcal{T}_+$. If the density matrix $\rho \in L^\infty(\mathbb{R}, \mathcal{B}(\mathbb{C}^2))$ satisfies the condition* (A.26) *and the boundary coefficients obey* $\mathrm{Re}(\varkappa_a) \le 0$ *and* $\mathrm{Re}(\varkappa_b) \le 0$*, then*

$$\|\mathcal{N}_\rho[\tau]\|_{L^1(\Omega_q)} \le C_\rho \left(3 + \left[8 + 4\sqrt{\|m\|_{L^\infty(\Omega_q)}}(b-a)\right]\sqrt{1 + \|V_-\|_{L^\infty(\Omega_q)}}\right), \tag{A.28}$$

where $V_-(x) := \max\{0, -V(x)\}$, $x \in \Omega_q$, and C_ρ is given by (A.26).

Proof. In [4, Lemma 6.2] the estimate

$$\|\mathcal{N}_\rho[\tau]\|_{L^1(\Omega_q)} \le C_\rho \left(3 + \left[8 + 4\sqrt{\|m\|_{L^\infty(\Omega_q)}}(b-a)\right]\sqrt{1 + \|V\|_{L^\infty(\Omega_q)}}\right)$$

is proven. The improved estimate (A.28) can be obtained by carefully checking the proof of Lemma 6.2 of [4]. Indeed, by doing so one obtains that the nonnegative part of the potential v moves the spectrum of the operator H to the right-hand side, which yields that it can be neglected. $\square$

We are going to verify the continuity of the carrier density operator in its dependence of τ. To this end we need the following.

PROPOSITION A.5. *Let $\tau, \tau_n \in \mathcal{T}$, $n \in \mathbb{N}$. If $\tau_n \to \tau$ as $n \to \infty$, then*

$$\lim_{n\to\infty} \|(K[\tau_n] - z)^{-1} - (K[\tau] - z)^{-1}\|_{\mathcal{B}_1(\mathfrak{K})} = 0 \tag{A.29}$$

for $z \in \mathbb{C} \setminus \mathbb{R}$.

Proof. At first we show that for each $\vec{g} \in \mathrm{dom}(K[\tau])$ there is a sequence $\{\vec{g}_n\}_{n\in\mathbb{N}}$ such that $\vec{g}_n \in \mathrm{dom}(K[\varkappa_{a,n}, \varkappa_{b,n}, V])$, $\lim_{n\to\infty} \vec{g}_n = \vec{g}$, and $\lim_{n\to\infty} K[\varkappa_{a,n}, \varkappa_{b,n}, V]\vec{g}_n = K[\tau]\vec{g}$ in the sense of $\mathfrak{K}$. Let

$$\vec{g}_n = \vec{g} + \vec{h}_n, \quad n \in \mathbb{N},$$

where

$$\vec{h}_n := 0 \oplus h_n \oplus h_{+,n}, \quad n \in \mathbb{N}.$$

Furthermore, let $\theta(\cdot) : \mathbb{R} \longrightarrow [0,1]$ be a smooth function equal to one in a neighborhood of zero and equal to zero in neighborhood of $y_0 := 2\int_a^b dt\, m(t)$. We set

$$h_n(x) := \theta\left(2\int_a^x dt\, m(t)\right) h_{a,n}(x) + \theta\left(2\int_x^b dt\, m(t)\right) h_{b,n}(x), \quad x \in \Omega_q, \quad n \in \mathbb{N},$$

where

$$h_{a,n}(x) := 2C_{a,n}\int_a^x m(t)dt, \qquad h_{b,n}(x) := -2C_{b,n}\int_x^b m(t)dt, \quad x \in [a,b],$$

and

$$\begin{aligned} C_{a,n} &:= (\alpha_{a,n} - \alpha_a)g_-^a(0) - (\varkappa_{a,n} - \varkappa_a)g(a), \\ C_{b,n} &:= (\alpha_{b,n} - \alpha_b)g_-^b(0) + (\varkappa_{b,n} - \varkappa_b)g(b). \end{aligned}$$

Notice that $\lim_{n\to\infty} C_{a,n} = \lim_{n\to\infty} C_{b,n} = 0$. Further, we set

$$h^a_{+,n}(x) := h^a_{+,n}(0)e^{-x}, \quad x \in \mathbb{R}_+, \quad \text{and} \quad h^b_{+,n}(x) := h^b_{+,n}(0)e^{-x}, \quad x \in \mathbb{R}_+,$$

where

$$h^a_{+,n}(0) := -i(\alpha_{a,n} - \alpha_a)g(a) \quad \text{and} \quad h^b_{+,n}(0) := i(\alpha_{b,n} - \alpha_b)g(b), \quad n \in \mathbb{N}.$$

A straightforward computation shows that $\vec{g}_n \in \mathrm{dom}(K[\varkappa_{a,n}, \varkappa_{b,n}, V])$, $\lim_{n\to\infty} \vec{g}_n = \vec{g}$, and $\lim_{n\to\infty} K[\varkappa_{a,n}, \varkappa_{b,n}, V]\vec{g}_n = K[\tau]\vec{g}$.

Since the sequence $\{\vec{g}_n\}_{n\in\mathbb{N}}$ exists for each $\vec{g} \in \mathrm{dom}(K[\tau])$, one gets by [40, Theorem 2.1] that

$$s - \lim_{n\to\infty} (K[\varkappa_{a,n}, \varkappa_{b,n}, V] - z)^{-1} = (K[\tau] - z)^{-1}. \tag{A.30}$$

The operators $K[\varkappa_{a,n}, \varkappa_{b,n}, V]$, $n \in \mathbb{N}$, and $K[\tau]$ are self-adjoint extensions of the symmetric operator $K_\bullet[V]$ given by

$$\mathrm{dom}(K_\bullet[V]) := \left\{ \vec{g} \in \mathfrak{K} : \begin{array}{c} g_\pm \in W^{1,2}(\mathbb{R}_\pm, \mathbb{C}^2) \quad g, \frac{1}{m}g' \in W^{1,2}(\Omega_q), \\ g(a) = g(b) = 0, \\ \frac{1}{m(a)}g'(a) = \frac{1}{m(b)}g'(b) = 0, \\ g_-(0) = g_+(0) = 0 \end{array} \right\}$$

and

$$K_\bullet[V]\vec{g} := -i\frac{d}{dx}g_- \oplus \left(-\frac{1}{2}\frac{d}{dx}\frac{1}{m}\frac{d}{dx} + V\right) g \oplus -i\frac{d}{dx}g_+, \quad \vec{g} \in \mathrm{dom}(K_\bullet[V]),$$

which has the deficiency indices $\{4, 4\}$. This fact immediately improves the strong convergence (A.30) to the trace class convergence, i.e.,

$$\lim_{n\to\infty} \|(K[\varkappa_{a,n}, \varkappa_{b,n}, V] - z)^{-1} - (K[\tau] - z)^{-1}\|_{\mathcal{B}_1(\mathfrak{K})} = 0.$$

Since $P(K[\tau] - z)^{-1} \in \mathcal{B}_1(\mathfrak{K})$, one gets

$$\lim_{n\to\infty} \|(K[\tau_n] - z)^{-1}(V - V_n)P(K[\tau] - z)^{-1}\|_{\mathcal{B}_1(\mathfrak{K})} = 0.$$

Using the representation

$$\begin{aligned} &(K[\tau_n] - z)^{-1} - (K[\varkappa_{a,n}, \varkappa_{b,n}, V] - z)^{-1} \\ &\quad = (K[\tau_n] - z)^{-1}(V - V_n)P(K[\tau] - z)^{-1} \\ &\qquad + (K[\tau_n] - z)^{-1}(V - V_n)P\left((K[\varkappa_{a,n}, \varkappa_{b,n}, V] - z)^{-1} - (K[\tau] - z)^{-1}\right), \end{aligned}$$

we find

$$\lim_{n\to\infty} \|(K[\tau_n] - z)^{-1} - (K[\varkappa_{a,n}, \varkappa_{b,n}, V] - z)^{-1}\|_{\mathcal{B}_1(\mathfrak{K})} = 0.$$

Finally, taking into account the representation

$$\begin{aligned} &(K[\tau_n] - z)^{-1} - (K[\tau] - z)^{-1} \\ &= (K[\tau_n]-z)^{-1} - (K[\varkappa_{a,n}, \varkappa_{b,n}, V]-z)^{-1} + (K[\varkappa_{a,n}, \varkappa_{b,n}, V]-z)^{-1} - (K[\tau]-z)^{-1}, \end{aligned}$$

we complete the proof. ∎

Proposition A.5 immediately implies the continuity of the incoming Fourier transform.

Proposition A.6. *Let $\tau, \tau_n \in \mathcal{T}$, $n \in \mathbb{N}$. If $\tau_n \longrightarrow \tau$ as $n \to \infty$, then*

$$s - \lim_{n\to\infty} \Phi_-[\tau_n] = \Phi_-[\tau]. \tag{A.31}$$

Proof. Let

$$W_-[\tau_n, \tau] := s - \lim_{t\to-\infty} e^{itK[\tau_n]}e^{-itK[\tau]}.$$

From Proposition A.5 we obtain that

$$s - \lim_{n\to\infty} W_-[\tau_n, \tau] = I_{\mathfrak{K}},$$

which yields

$$w - \lim_{n\to\infty} W_-[\tau_n, \tau]^* = I_{\mathfrak{K}}.$$

Since $W_-[\tau_n, \tau]$ is unitary for each $n \in \mathbb{N}$, we find

$$s - \lim_{n\to\infty} W_-[\tau_n, \tau]^* = I_{\mathfrak{K}}.$$

By the chain rule for wave operators, we obtain

$$W_-[\tau_n] = W_-[\tau_n, \tau]W_-[\tau], \quad n \in \mathbb{N},$$

which gives

$$s - \lim_{n\to\infty} W_-[\tau_n]^* = W_-[\tau]^*. \tag{A.32}$$

Taking into account the representation (A.19), we complete the proof. ∎

Using Proposition A.6, we will now verify the continuity of the carrier density operator.

Theorem A.7. *Let $\tau, \tau_n \in \mathcal{T}$, $n \in \mathbb{N}$. Further, suppose that there is a density matrix $\rho \in L^\infty(\mathbb{R}, \mathcal{B}(\mathbb{C}^2))$ such that $C_\rho < \infty$ and a sequence of density matrices $\{\rho_n\}_{n\in\mathbb{N}}$, $\rho_n \in L^\infty(\mathbb{R}, \mathcal{B}(\mathbb{C}^2))$, such that $\sup_{n\in\mathbb{N}} C_{\rho_n} < \infty$. If $\tau_n \to \tau$ as $n \to \infty$ and*

$$\lim_{n\to\infty} \rho_n(\lambda) = \rho(\lambda)$$

for a.e. $\lambda \in \mathbb{R}$, then

$$\lim_{n\to\infty} \|\mathcal{N}_{\rho_n}[\tau_n] - \mathcal{N}_\rho[\tau]\|_{L^1(\Omega_q)} = 0. \tag{A.33}$$

Proof. We set $\Phi_n := \Phi_-[\tau_n]$ and $\Phi := \Phi_-[\tau]$ as well as

$$\iota_n(\lambda) := (\lambda - i)\rho_n(\lambda) \quad \text{and} \quad \iota(\lambda) := (\lambda - i)\rho(\lambda), \quad \lambda \in \mathbb{R}, \quad n \in \mathbb{N}.$$

From (A.25) we find the representation

$$\varrho_n[\tau_n] - \varrho[\tau] = \Phi_n^* \,\widehat{\iota}_n\, \Phi_n(K_n - i)^{-1} - \Phi^* \,\widehat{\iota}\, \Phi(K - i)^{-1}, \quad n \in \mathbb{N},$$

where $K_n := K[\tau_n]$ and $K := K[\tau]$. Note that $(K-i)^{-1}P \in \mathcal{B}_1(\mathfrak{K})$. Hence we find the estimate

$$\begin{aligned}
&\left\|(\varrho_n[\tau_n] - \varrho[\tau])P_{\mathfrak{h}}^{\mathfrak{K}}\right\|_{\mathcal{B}_1(\mathfrak{K})} \\
&\quad \le C_{\rho_n}\left\|(K_n - i)^{-1} - (K-i)^{-1}\right\|_{\mathcal{B}_1(\mathfrak{K})} + C_{\rho_n}\left\|(\Phi_n - \Phi)(K-i)^{-1}P\right\|_{\mathcal{B}_1(\mathfrak{K},\mathfrak{K}_0)} \\
&\qquad + \left\|(\widehat{\iota}_n - \widehat{\iota})\Phi(K-i)^{-1}P\right\|_{\mathcal{B}_1(\mathfrak{K},\mathfrak{K}_0)} + \left\|(\Phi_n^* - \Phi^*)\,\widehat{\iota}\,\Phi(K-i)^{-1}P\right\|_{\mathcal{B}_1(\mathfrak{K})}.
\end{aligned}$$

The first term on the right-hand side goes to zero by Proposition A.5. The second term tends to zero by Proposition A.6 and $(K-i)^{-1}P \in \mathcal{B}_1(\mathfrak{K})$. By $s-\lim_{n\to\infty}\widehat{\iota}_n = \widehat{\iota}$ and $(K-i)^{-1}P \in \mathcal{B}_1(\mathfrak{K})$ the third term goes to zero. Finally, from Proposition A.6 and the isometry of the incoming Fourier transform we get $s-\lim_{n\to\infty}\Phi_n = \Phi$, which yields that the fourth term converges to zero. Hence we find

$$\lim_{n\to\infty}\|(\varrho_n[\tau_n] - \varrho[\tau])P\|_{\mathcal{B}_1(\mathfrak{K})} = 0.$$

Taking into account (A.27), this proves (A.33). $\square$

A.4. Current density operator. Similar to the carrier density operator, it is possible to introduce a current density operator $j_\rho : \mathcal{T} \longrightarrow \mathbb{R}$ for a given maximal dissipative operator $h[\tau]$ and a density matrix $\rho \in L^\infty(\mathbb{R}, \mathcal{B}(\mathbb{C}^2))$, provided the density matrix satisfies the additional condition

$$L_\rho := \int_{\mathbb{R}} d\lambda\, \mathrm{tr}(\rho(\lambda)) < \infty \tag{A.34}$$

(cf. [3, 5, 35, 33]). In [33] it is shown that the current density operator admits a representation by the so-called current density observable $C[\tau](\lambda)$ at energy $\lambda \in \mathbb{R}$, which is defined by

$$C[\tau](\lambda) := \frac{1}{2\pi}\left(P_a\Theta[\tau](\lambda)P_b - P_b\Theta[\tau](\lambda)P_a\right)\Theta[\tau](\lambda)^*, \quad \tau \in \mathcal{T}, \tag{A.35}$$

where $\Theta[\tau]$ is the characteristic function of the maximal dissipative operator $H[\tau]$ (cf. (A.14) and [5, 33]) and the projections $P_a := (\cdot, e_a)_{\mathbb{C}^2}e_a$, $P_b := (\cdot, e_b)_{\mathbb{C}^2}e_b$ with $e_a = \binom{0}{1}$, $e_b = \binom{1}{0}$. Indeed, if the density matrix ρ satisfies the condition (A.34), then the current density operator $j_\rho[\cdot] : \mathcal{T}_+ \longrightarrow \mathbb{R}$ admits the representation

$$j_\rho[\tau] := \int_{\mathbb{R}} d\lambda\, \mathrm{tr}(\rho(\lambda)C[\tau](\lambda)). \tag{A.36}$$

Since $\|C[\tau](\lambda)\|_{\mathcal{B}(\mathbb{C}^2)} \le \frac{1}{2\pi}$, one gets the estimate

$$|j_\rho[\tau]| \le \frac{1}{2\pi}\int_{\mathbb{R}} d\lambda\, \mathrm{tr}(\rho(\lambda)). \tag{A.37}$$

Remark A.8. In the case when $\rho^\pm$ are given as in Remark A.3, we obtain from (A.36) the expression (3.7), where the transmission coefficients are given by (A.23).

Let us now prove the continuity of the current density operator.

THEOREM A.9. *Let $\tau, \tau_n \in \mathcal{T}$, $n \in \mathbb{N}$. Further, suppose that there is a density matrix $\rho \in L^\infty(\mathbb{R}, \mathcal{B}(\mathbb{C}^2))$ such that $L_\rho < \infty$ (cf.* (A.34)*) and a sequence of density* *matrices* $\{\rho_n\}_{n\in\mathbb{N}}$, $\rho_n \in L^\infty(\mathbb{R},\mathcal{B}(\mathbb{C}^2))$, *such that* $\sup_{n\in\mathbb{N}} L_{\rho_n} < \infty$. *If* $\tau_n \longrightarrow \tau$ *as* $n\to\infty$, $\lim_{n\to\infty}\rho_n(\lambda) = \rho(\lambda)$ *for a.e.* $\lambda\in\mathbb{R}$, *and*

$$\lim_{n\to\infty}\int_{\mathbb{R}} d\lambda\ (\rho_n(\lambda)e,e)_{\mathbb{C}^2} = \int_{\mathbb{R}} d\lambda\ (\rho(\lambda)e,e)_{\mathbb{C}^2} \tag{A.38}$$

for each $e\in\mathbb{C}^2$, *then* $\lim_{n\to\infty} j_{\rho_n}[\tau_n] = j_\rho[\tau]$.

Proof. By Proposition A.5 we obtain that $s-\lim_{n\to\infty} S[\tau_n] = S[\tau]$. This yields

$$\lim_{n\to\infty}\int_{\mathbb{R}} d\lambda\ \left\|(S[\tau_n](\lambda) - S[\tau](\lambda))\vec{f}(\lambda)\right\|^2_{\mathbb{C}^2} = 0$$

for each $\vec{f}\in\widehat{\mathfrak{K}}_0 = L^2(\mathbb{R},\mathbb{C}^2)$. Taking into account (A.22), we obtain

$$\lim_{n\to\infty}\int_{\mathbb{R}} d\lambda\ \left\|(\Theta[\tau_n](\lambda)^* - \Theta[\tau](\lambda)^*)\vec{f}(\lambda)\right\|^2_{\mathbb{C}^2} = 0$$

and

$$\lim_{n\to\infty}\int_{\mathbb{R}} d\lambda\ \left\|(\Theta[\tau_n](\lambda) - \Theta[\tau](\lambda))\vec{f}(\lambda)\right\|^2_{\mathbb{C}^2} = 0$$

for each $\vec{f}\in\widehat{\mathfrak{K}}_0$. Hence

$$\lim_{n\to\infty}\int_{\mathbb{R}} d\lambda\ \left\|(C[\tau_n](\lambda) - C[\tau](\lambda))\vec{f}(\lambda)\right\|^2_{\mathbb{C}^2} = 0 \tag{A.39}$$

for each $\vec{f}\in\widehat{\mathfrak{K}}_0$. Further, we have

$$\begin{aligned}
&\mathrm{tr}(\rho_n(\lambda)C[\tau_n](\lambda)) - \mathrm{tr}(\rho(\lambda)C[\tau](\lambda))\\
&\quad = \mathrm{tr}(\rho_n(\lambda)^{1/2}C[\tau_n]\rho_n(\lambda)^{1/2})_{\mathbb{C}^2} - \mathrm{tr}(\rho^{1/2}(\lambda)C[\tau]\rho^{1/2}(\lambda))_{\mathbb{C}^2}\\
&\quad = \sum_{\nu=a,b}\left\{(C[\tau_n](\lambda)\rho_n(\lambda)^{1/2}e_\nu,\rho_n(\lambda)^{1/2}e_\nu)_{\mathbb{C}^2} - (C[\tau](\lambda)\rho(\lambda)^{1/2}e_\nu,\rho(\lambda)^{1/2}e_\nu)_{\mathbb{C}^2}\right\},
\end{aligned}$$

$\lambda\in\mathbb{R}$. Setting $\vec{f}_{\nu,n}(\lambda) := \rho_n(\lambda)^{1/2}e_\nu$ and $\vec{f}_\nu(\lambda) := \rho(\lambda)^{1/2}e_\nu$, $\nu = a,b$, we get $\vec{f}_{\nu,n},\vec{f}_\nu\in\mathfrak{K}_0$. Using $\|C[\tau](\lambda)\|_{\mathcal{B}(\mathbb{C}^2)}\le\frac{1}{2\pi}$, $\lambda\in\mathbb{R}$, we obtain the estimate

$$\begin{aligned}
&\left|\mathrm{tr}(\rho_n(\lambda)C[\tau_n](\lambda)) - \mathrm{tr}(\rho(\lambda)C[\tau](\lambda))\right|\\
&\quad\le\sum_{\nu=a,b}\frac{1}{2\pi}\left\{\|\vec{f}_{\nu,n}(\lambda)\|_{\mathbb{C}^2} + \|\vec{f}_\nu(\lambda)\|_{\mathbb{C}^2}\right\}\|\vec{f}_{\nu,n}(\lambda) - \vec{f}_\nu(\lambda)\|_{\mathbb{C}^2}\\
&\qquad + \sum_{\nu=a,b}\|\vec{f}_\nu(\lambda)\|_{\mathbb{C}^2}\|(C[\tau_n](\lambda) - C[\tau](\lambda))\vec{f}_\nu(\lambda)\|_{\mathbb{C}^2}.
\end{aligned}$$

We note that $\vec{f}_{\nu,n},\vec{f}_\nu\in\widehat{\mathfrak{K}}_0$, $\nu=a,b$, $n\in\mathbb{N}$. Hence

$$
\begin{aligned}
|j_{\rho_n}[\tau_n] - j_\rho[\tau]| &\le \int_{\mathbb{R}} d\lambda \ |\mathrm{tr}(C[\tau_n](\lambda)\rho_n(\lambda)) - \mathrm{tr}(C[\tau](\lambda)\rho(\lambda))| \\
&\le \frac{1}{2\pi}\left\{\left(\sum_{\nu=a,b}\|\vec{f}_{\nu,n}\|^2_{\widehat{\mathfrak{K}}_0}\right)^{1/2} + \left(\sum_{\nu=a,b}\|\vec{f}_{\nu}\|^2_{\widehat{\mathfrak{K}}_0}\right)^{1/2}\right\}\left(\sum_{\nu=a,b}\|\vec{f}_{\nu,n}-\vec{f}_{\nu}\|^2_{\widehat{\mathfrak{K}}_0}\right)^{1/2} \\
&\quad + \left(\sum_{\nu=a,b}\|\vec{f}_{\nu}\|^2_{\widehat{\mathfrak{K}}_0}\right)^{1/2}\left(\sum_{\nu=a,b}\int_{\mathbb{R}} d\lambda \ \|(C[\tau_n](\lambda)-C[\tau](\lambda))\vec{f}_\nu(\lambda)\|^2_{\mathbb{C}^2}\right)^{1/2},
\end{aligned}
$$

which yields

$$
\begin{aligned}
\text{(A.40)}\quad \left|j_{\rho_n}[\tau_n] - j_\rho[\tau]\right| &\le \left\{L^{1/2}_{\rho_n} + L^{1/2}_{\rho}\right\}\left(\sum_{\nu=a,b}\|\vec{f}_{\nu,n}-\vec{f}_{\nu}\|^2_{\widehat{\mathfrak{K}}_0}\right)^{1/2} \\
&\quad + L^{1/2}_{\rho}\left(\sum_{\nu=a,b}\int_{\mathbb{R}} d\lambda \ \|(C[\tau_n](\lambda)-C[\tau](\lambda))\vec{f}_\nu(\lambda)\|^2_{\mathbb{C}^2}\right)^{1/2}.
\end{aligned}
$$

Since $\rho_n(\lambda) \to \rho(\lambda)$ for a.e. $\lambda \in \mathbb{R}$ as $n \to \infty$, we find $\vec{f}_{\nu,n}(\lambda) \to \vec{f}_\nu(\lambda)$ for a.e. $\lambda \in \mathbb{R}$ as $n \to \infty$. From (A.38) we get that $\lim_{n\to\infty}\|\vec{f}_{\nu,n}\|_{\widehat{\mathfrak{K}}_0} = \|\vec{f}_\nu\|_{\widehat{\mathfrak{K}}_0}$, $\nu = a,b$. By Theorem 13.44 of [28] we get that $\vec{f}_{\nu,n}$ converges weakly to $\vec{f}_\nu$, which implies $\lim_{n\to\infty}\|\vec{f}_{\nu,n}-\vec{f}_\nu\|_{\widehat{\mathfrak{K}}_0} = 0$, $\nu = a,b$. Thus the first term of the right-hand side tends to zero as $n \to \infty$. Taking into account (A.39), we show that the second term of the right-hand side goes to zero as $n \to \infty$. □

Appendix B. Approximation. In the following we will outline in which sense the dissipative Schrödinger system can be regarded as an approximation of the frequently used Schrödinger scattering model in [8, 22] (also known as the quantum transmitting Schrödinger model [4]).

B.1. The quantum transmitting Schrödinger model. Let us briefly review the quantum transmitting Schrödinger model considered, e.g., in [8, 4, 11, 22]. The wave functions $\widetilde{\psi}$ are solutions of the one-dimensional Schrödinger equation

$$
\text{(B.1)}\qquad \left(-\frac{1}{2}\frac{d}{dx}\frac{d}{dx} + \widetilde{V}\right)\widetilde{\psi}(x) = \lambda\widetilde{\psi}(x), \quad x \in \mathbb{R},
$$

where the effective mass m and the potential $\widetilde{V}$ satisfy

$$
m(x) = \begin{cases} m_a, & x \in (-\infty, a),\\ m(x), & x \in \Omega_q,\\ m_b, & x \in (b,\infty),\end{cases} \quad\text{and}\quad \widetilde{V}(x) = \begin{cases} V(a), & x \in (-\infty, a),\\ V(x), & x \in \Omega_q,\\ V(b), & x \in (b,\infty).\end{cases}
$$

Due to the flatness of the potential in the asymptotic regions, the solutions of the Schrödinger equation (B.1) are superpositions of plane waves. In what follows we will refer to quantities in the left asymptotic region $(-\infty, a)$ by the subscript a and, similarly, the subscript b is used for quantities in the right asymptotic region (b, ∞).

For energies $\lambda > V(a)$ and $\lambda > V(b)$ there will be two independent solutions $\widetilde{\psi_a}$, $\widetilde{\psi_b}$ of the Schrödinger equation representing particles incident from the left and right, respectively. In the asymptotic regions they are of the form

$$\widetilde{\psi_a}(\lambda, x) = \frac{1}{\sqrt{2\pi\, v_a(\lambda)}} \begin{cases} e^{ik_a(\lambda)(x-a)} + S_{aa}(\lambda) e^{-ik_a(\lambda)(x-a)}, & x < a, \\ S_{ab}(\lambda) e^{ik_b(\lambda)(x-b)}, & x > b, \end{cases}$$

for $\lambda > V(a)$ and

$$\widetilde{\psi_b}(\lambda, x) = \frac{1}{\sqrt{2\pi\, v_b(\lambda)}} \begin{cases} S_{ba}(\lambda) e^{-ik_a(\lambda)(x-a)}, & x < a, \\ e^{-ik_b(\lambda)(x-b)} + S_{bb}(\lambda) e^{-ik_b(\lambda)(x-b)}, & x > b, \end{cases}$$

for $\lambda > V(b)$, where $k_\nu(\lambda) = [2m_\nu(\lambda - V(\nu))]^{1/2}$, $\nu = a, b$, are the wave vectors; $v_\nu(\lambda) = k_\nu(\lambda)/m_\nu$, $\nu = a, b$, are the group velocities; S_{ab}, S_{ba} the transmission amplitudes; and S_{aa}, S_{bb} the reflection amplitudes.

We note that the Schrödinger equation (B.1) may have bounded states with energies $\lambda < \min\{V(a), V(b)\}$. However, we will neglect theses states in the present considerations.

It is well known that the solutions $\widetilde{\psi}_a$, $\widetilde{\psi}_b$ of the Schrödinger equation (B.1) can be obtained by solving (B.1) on the bounded domain Ω_q with appropriate boundary conditions; see, e.g., [8, 22]. Moreover, the Schrödinger operator induced by (B.1) on $L^2(\mathbb{R};\mathbb{C})$ is completely described by the so-called quantum transmitting boundary family, which is a maximal dissipative operator family on $L^2(\Omega_q, \mathbb{C})$; see [4].

The macroscopic quantities such as carrier density and current density are given by

$$\tilde{u}(x) = \sum_{\nu=a,b} \int_{V(\nu)}^{\infty} d\lambda\, f(\lambda - \epsilon_\nu) |\widetilde{\psi_\nu}(\lambda, x)|^2, \quad x \in \Omega_q, \tag{B.2}$$

$$\tilde{j} = \sum_{\nu=a,b} \int_{V(\nu)}^{\infty} d\lambda\, f(\lambda - \epsilon_\nu) \mathrm{Im} \left(\frac{1}{m(x)} \frac{\partial \widetilde{\psi_\nu}(\lambda, x)}{\partial x} \overline{\widetilde{\psi_\nu}(\lambda, x)} \right), \tag{B.3}$$

where $\epsilon_a, \epsilon_b \in \mathbb{R}$ are given Fermi levels and f is a distribution function such as Boltzmann or Fermi–Dirac; see Remark 3.2.

Let us introduce the transmission coefficient $T(\lambda)$ given by

$$T(\lambda) := \frac{v_b(\lambda)}{v_a(\lambda)} |S_{ba}(\lambda)|^2, \qquad \lambda \in (V_{max}, \infty), \tag{B.4}$$

where $V_{max} = \max\{V(a), V(b)\}$. The current density $\tilde{j}$ can then be written as

$$\tilde{j} = \frac{1}{2\pi} \int_{V_{max}}^{\infty} d\lambda\, T(\lambda) \Big(f(\lambda - \epsilon_a) - f(\lambda - \epsilon_b) \Big) \tag{B.5}$$

(see, e.g., [4, 8, 22]). From (B.5) we observe that only the wave functions $\widetilde{\psi_\nu}(\lambda, x)$, $\nu = a, b$, with energies larger than V_{max}, contribute to the current.

B.2. Dissipative approximation. In [4] it is shown that the Schrödinger model considered in the previous section and the dissipative Schrödinger model coincide for fixed energies. In this section we will use this correspondence to derive a dissipative approximation.

Let $\widetilde{\psi_\nu}$, $\nu = a, b$, be the scattering solutions of the Schrödinger equation (B.1). As already mentioned, only the states with energy above V_{max} contribute to the current density. In what follows we are interested in only these states. Assume that δ is a strictly positive constant. We introduce the discrete energies λ_k, $k = 0, 1, \dots$, by

$$\lambda_k := V_{max} + k\,\delta, \qquad k = 0, 1, \dots .$$

Moreover, we introduce

$$\Lambda_k := [\lambda_{k-1}, \lambda_k), \qquad k = 1, 2, \dots .$$

Therefore, the carrier $\tilde{u}$ and current density $\tilde{j}$ (see (B.2), (B.5)) may be written as

$$\tilde{u}(x) = \sum_{k=1}^{\infty} \tilde{u}_k(x), \quad x \in \Omega, \quad \text{and} \quad \tilde{j} = \sum_{j=1}^{\infty} \tilde{j}_k,$$

where

$$\text{(B.6)} \quad \tilde{u}_k(x) := \int_{\Lambda_k} d\lambda \sum_{\nu=a,b} f(\lambda - \epsilon_\nu) |\widetilde{\psi_\nu}(\lambda, x)|^2, \qquad k = 1, 2, \dots, \quad x \in \Omega,$$

$$\text{(B.7)} \qquad \tilde{j}_k \; := \frac{1}{2\pi} \int_{\Lambda_k} d\lambda\, T(\lambda) \left(f(\lambda - \epsilon_a) - f(\lambda - \epsilon_b) \right), \qquad k = 1, 2, \dots .$$

Let δ_0 be a strictly positive constant with $\delta_0 < \delta$. The constants s_k, $k = 1, 2, \dots$, are given by

$$s_k := \lambda_{k-1} + \delta_0, \qquad k = 1, 2, \dots .$$

Note that $s_k \in \Lambda_j$ for each $k = 1, 2, \dots$. Moreover, we define the complex constants

$$\varkappa_a(s_k) := 2i\, v_a(s_k) = i\sqrt{\frac{s_k - V(a)}{2m_a}}, \qquad \varkappa_a(s_k) := 2i\, v_b(s_k) = i\sqrt{\frac{s_k - V(b)}{2m_b}},$$

where $v_\nu(\lambda)$, $\nu = a, b$, are the group velocities defined in section B.1; see also (1.8). Note that $\mathrm{Im}(\varkappa_a(s_k)), \mathrm{Im}(\varkappa_b(s_k)) > 0$ for all $k = 1, 2, \dots$. The maximal dissipative Schrödinger operators defined by (A.1) and (A.2) by means of $\varkappa_a(s_k)$ and $\varkappa_b(s_k)$ are denoted by $H(s_k)$, $k = 1, 2, \dots$, where the potential V is given and fixed. To each maximal dissipative operator $H(s_k)$ there corresponds a minimal self-adjoint dilation $K(s_k)$ (see (A.8), (A.9) in section A.2) and incoming eigenfunctions $\psi_\nu^{(s_k)}(\lambda, x)$, $\nu = a, b$, $k = 1, 2, \dots$; see (A.15). Moreover, the transmission coefficient (A.23) corresponding to each maximal dissipative operator $H(s_k)$ is denoted by $t^{(s_k)}(\lambda)$, $k = 1, 2, \dots$.

In [4] the following relations between the eigenfunctions $\widetilde{\psi_\nu}$ of section B.1 and $\psi_\nu^{(s_k)}$, $\nu = a, b$, are proven:

$$\text{(B.8)} \quad |\widetilde{\psi_\nu}(s_k, x)|^2 = |\psi_\nu^{(s_k)}(s_k, x)|^2 \quad \text{for all } x \in \Omega_q \text{ and } k = 1, 2, \dots, \text{ where } \nu = a, b.$$

Moreover, for the transmission coefficient T from (B.4) and the transmission coefficients $t^{(s_k)}$, $k = 1, 2, \dots$, there is

$$\text{(B.9)} \qquad T(s_k) = t^{(s_k)}(s_k) \quad \text{for all } k = 1, 2, \dots .$$

Using these relations, we may approximate the carrier densities $\tilde{u}_k$ and the current density $\tilde{j}_k$, $k = 1, 2, \dots$, defined by (B.6) and (B.7), respectively, by

$$\widetilde{u}_k(x) \approx u^{(s_k)}(x) \quad \text{and} \quad \widetilde{j}_k \approx j^{(s_k)}, \quad k = 1, 2, \dots,$$

where

$$\begin{aligned} u^{(s_k)}(x) \quad &:= \int_{\Lambda_k} d\lambda \sum_{\nu=a,b} f(\lambda - \epsilon_\nu)|\psi^{(s_k)}(\lambda, x)|^2, \\ j^{(s_k)} &:= \int_{\Lambda_k} d\lambda\, t^{(s_k)}(\lambda)\left(f(\lambda - \epsilon_a) - f(\lambda - \epsilon_b)\right). \end{aligned}$$

Hence, for large $N \in \mathbb{N}$ we obtain

$$\tilde{u}(x) \approx \sum_{k=1}^{N} u^{s_k}(x) \quad \text{and} \quad \tilde{j} = \sum_{k=1}^{N} j^{(s_k)},$$

where $\tilde{u}$ and $\tilde{j}$ are the carrier and current densities given by (B.2) and (B.3), respectively.

Thus, carrier and current densities are approximated by means of a sequence of carrier densities $u^{(s_k)}$ and currents densities $j^{(s_k)}$, which are determined by a sequence of maximal dissipative operators and given statistics. The quantities δ, δ_0, and N are the approximation parameters.

In the case when we choose the rather rough approximation $N = 1$, we obtain

$$\tilde{u}(x) \approx \int_{\Lambda} d\lambda \sum_{\nu=a,b} f(\lambda - \epsilon_\nu)|\psi_\nu(\lambda, x)|^2 \quad \text{and} \quad \tilde{j} \approx \int_{\Lambda} d\lambda\, t(\lambda)(f(\lambda - \epsilon_a) - f(\lambda - \epsilon_b)),$$

where

$$s := V_{max} + \delta_0, \quad t(\lambda) := t^{(s)}(\lambda), \quad \psi_\nu := \psi_\nu^{(s)}, \tag{B.10}$$

and

$$\Lambda := [V_{max}, V_{max} + \delta), \tag{B.11}$$

for some strictly positive δ, δ_0 with $\delta_0 < \delta$. This is the approximation we consider in this paper; see (1.11)–(1.13) and (3.5)–(3.7).

Acknowledgments. We thank N. Ben Abdallah, P. Degond, and H.-Chr. Kaiser for fruitful discussions and thank the Institute “Mathematiques pour l’Industrie et la Physique” in Toulouse for hospitality and financial support. Moreover, we acknowledge the helpful suggestions and comments received from the referees, which improved the quality of the present article.

REFERENCES

[1] A. Arnold and F. Nier, *The two-dimensional Wigner-Poisson problem for an electron gas in the charge neutral case*, Math. Methods Appl. Sci., 14 (1991), pp. 595–613.

[2] M. Baro, N. Ben Abdallah, P. Degond, and A. El Ayyadi, *A 1-D coupled Schrödinger drift-diffusion model including collisions*, J. Comput. Phys., 203 (2005), pp. 129–153.

[3] M. Baro, H.-Chr. Kaiser, H. Neidhardt, and J. Rehberg, *Dissipative Schrödinger-Poisson systems*, J. Math. Phys., 45 (2004), pp. 21–43.

[4] M. Baro, H.-Chr. Kaiser, H. Neidhardt, and J. Rehberg, *A quantum transmitting Schrödinger-Poisson system*, Rev. Math. Phys., 16 (2004), pp. 281–330.
[5] M. Baro and H. Neidhardt, *Dissipative Schrödinger-type operators as a model for generation and recombination*, J. Math. Phys., 44 (2003), pp. 2373–2401.
[6] N. Ben Abdallah, *A hybrid kinetic-quantum model for stationary electron transport*, J. Statist. Phys., 90 (1998), pp. 627–662.
[7] N. Ben Abdallah, P. Degond, and I. Gamba, *Coupling one-dimensional time-dependent classical and quantum transport models*, J. Math. Phys., 43 (2002), pp. 1–24.
[8] N. Ben Abdallah, P. Degond, and P. A. Markowich, *On a one-dimensional Schrödinger-Poisson scattering model*, Z. Angew. Math. Phys., 48 (1997), pp. 135–155.
[9] N. Ben Abdallah and O. Pinaud, *A mathematical model for the transient evolution of a resonant tunneling diode*, C. R. Math. Acad. Sci. Paris, 334 (2002), pp. 283–288.
[10] F. Brezzi and P. A. Markowich, *The three-dimensional Wigner-Poisson problem: Existence, uniqueness and approximation*, Math. Methods Appl. Sci., 14 (1991), pp. 35–61.
[11] V. Buslaev and V. Fomin, *An inverse scattering problem for the one-dimensional Schrödinger equation on the entire axis*, Vestnik Leningrad Univ., 17 (1962), pp. 56–64 (in Russian).
[12] F. Castella, *L^2-solutions to the Schrödinger-Poisson system: Existence, uniqueness, time behaviour, and smoothing effects*, Math. Models Methods Appl. Sci., 7 (1997), pp. 1051–1083.
[13] Ph. Caussignac, J. Descloux, and A. Yamnahakki, *Simulation of some quantum models for semiconductors*, Math. Models Methods Appl. Sci., 12 (2002), pp. 1049–1074.
[14] F. Chevoir and B. Vinter, *Scattering-assisted tunneling in double-barrier diodes: Scattering rates and valley current*, Phys. Rev. B, 47 (1993), pp. 7260–7274.
[15] J. H. Davies, *The Physics of Low Dimensional Semiconductors*, Cambridge University Press, Cambridge, UK, 1997.
[16] P. Degond and A. El Ayyadi, *A coupled Schrödinger drift-diffusion model for quantum semiconductor device simulations*, J. Comput. Phys., 181 (2002), pp. 222–259.
[17] P. Exner, *Open Quantum System and Feynman Integrals*, Fundamental Theories of Physics, D. Reidel, Dordrecht, 1985.
[18] D. K. Ferry and S. M. Goodnick, *Transport in Nanostructures*, Cambridge University Press, Cambridge, UK, 1997.
[19] M. V. Fischetti, *Theory of electron transport in small semiconductor devices using the Pauli master equation*, J. Appl. Phys., 83 (1998), pp. 270–291.
[20] M. V. Fischetti, *Master-equation approach to the study of electronic transport in small semiconductor devices*, Phys. Rev. B, 59 (1999), pp. 4901–4917.
[21] C. Foias and B. Sz.-Nagy, B, *Harmonic Analysis of Operators on Hilbert Space*, Akadémia Kiadó, Budapest, North-Holland, Amsterdam, London; American Elsevier, New York, 1970.
[22] W. R. Frensley, *Boundary conditions for open quantum systems driven far from equilibrium*, Rev. Modern Phys., 62 (1990), pp. 745–791.
[23] H. Gajewski, *Analysis and numerics of carrier transport in semiconductors*, Mitt. Ges. Angew. Math. Mech., 16 (1993), pp. 35–57 (in German).
[24] H. Gajewski, K. Gröger, and K. Zacharias, *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, Mathematische Lehrbücher und Monographien, II. Abteilung, Mathematische Monographien, Band 38, Akademie-Verlag, Berlin, 1974.
[25] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Reprint of the 1998 edition. Classics in Math., Springer-Verlag, Berlin, 2001.
[26] J. B. Garnett, *Bounded Analytic Functions*, Pure and Appl. Math. 96, Academic Press, New York, 1981.
[27] H. K. Gummel, *A self-consistent iterative scheme for one-dimensional steady state transistor calculations*, IEEE Trans. Education 11 (1964), pp. 455–465.
[28] E. Hewitt and K. Stromberg, *Real and Abstract Analysis. A Modern Treatment of the Theory of Functions of a Real Variable*, Springer-Verlag, New York, 1965.
[29] R. Illner, *The Wigner-Poisson and Schrödinger-Poisson systems*, Transport. Theory Statist. Phys., 21 (1992), pp. 753–767.
[30] R. Illner, *Existence, uniqueness and asymptotic behavior of Wigner-Poisson and Vlasov-Poisson systems: A survey*, Transport. Theory Statist. Phys., 26 (1997), pp. 195–207.
[31] R. Illner, P. F. Zweifel, and H. Lange, *Global existence, uniqueness and asymptotic behaviour of solutions of the Wigner-Poisson and Schrödinger-Poisson systems*, Math. Methods Appl. Sci., 17 (1994), pp. 349–376.
[32] G. Jona-Lasinio, C. Presilla, and J. Sjöstrand, *On Schrödinger equations with concentrated nonlinearities*, Ann. Phys., 240 (1995), pp. 1–21.

[33] H.-Chr. Kaiser, H. Neidhardt, and J. Rehberg, *Density and current of a dissipative Schrödinger operator*, J. Math. Phys., 43 (2002), pp. 5325–5350.

[34] H.-Chr. Kaiser, H. Neidhardt, and J. Rehberg, *Macroscopic current induced boundary conditions for Schrödinger-type operators*, Integral Equations Operator Theory, 45 (2003), pp. 39–63.

[35] H.-Chr. Kaiser, H. Neidhardt, and J. Rehberg, *On 1-dimensional dissipative Schrödinger-type operators, their dilations and eigenfunction expansions*, Math. Nachr., 252 (2003), pp. 51–69.

[36] H.-Chr. Kaiser and J. Rehberg, *On stationary Schrödinger-Poisson equations*, Z. Angew. Math. Mech., 75 (1995), Suppl. 2, pp. 467–468.

[37] H.-Chr. Kaiser and J. Rehberg, *On stationary Schrödinger-Poisson equations modelling an electron gas with reduced dimension*, Math. Methods Appl. Sci., 20 (1997), pp. 1283–1312.

[38] H.-Chr. Kaiser and J. Rehberg, *About a one-dimensional stationary Schrödinger-Poisson system with Kohn-Sham potential*, Z. Angew. Math. Phys., 50 (1999), pp. 423–458.

[39] H.-Chr. Kaiser and J. Rehberg, *About a stationary Schrödinger-Poisson system with Kohn-Sham potential in a bounded two- or three-dimensional domain*, Nonlinear Anal. Ser. A Theory Methods, 41 (2000), pp. 33–72.

[40] Th. G. Kurtz, *Extensions of Trotter's operator semigroup approximation theorems*, J. Funct. Anal., 3 (1969), pp. 354–375.

[41] H. Lange, B. Toomire, and P. F. Zweifel, *An overview of Schrödinger-Poisson problems*, Rep. Math. Phys., 36 (1995), pp. 331–345.

[42] P. A. Markowich, *Boltzmann distributed quantum steady states and their classical limit*, Forum Math., 6 (1994), pp. 1–33.

[43] P. A. Markowich, *The Stationary Semiconductor Device Equations*, Springer-Verlag, Vienna, 1986.

[44] M. S. Mock, *Analysis of Mathematical Models of Semiconductor Devices*, Boole Press, Dún Laoghaire, Ireland, 1983.

[45] P. Mounaix, O. Vanbesien, and D. Lippens, *Effect of cathode space layer on the current-voltage characteristic of resonant tunneling diodes*, Appl. Phys. Lett., 57 (1990), pp. 1517–1519.

[46] F. Nier, *A stationary Schrödinger-Poisson system arising from the modelling of electronic devices*, Forum Math., 2 (1990), pp. 489–510.

[47] F. Nier, *A variational formulation of Schrödinger-Poisson systems in dimension $d \leq 3$*, Comm. Partial Differential Equations, 18 (1993), pp. 1125–1147.

[48] F. Nier, *Schrödinger-Poisson systems in dimension $d \leqq 3$: The whole-space case*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 1179–1201.

[49] F. Nier, *The dynamics of some quantum open systems with short range nonlinearities*, Nonlinearity, 11 (1998), pp. 1127–1172.

[50] C. Presilla and J. Sjöstrand, *Transport properties in resonant tunneling heterostructures*, J. Math. Phys., 37 (1996), pp. 4816–4844.

[51] W. van Roosbroeck, *Theory of flow electrons and holes in germanium and other semiconductors*, Bell Systems Tech. J., 29 (1950), pp. 560–607.

[52] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer-Verlag, Vienna, 1984.

[53] H. Steinrück, *The one-dimensional Wigner–Poisson problem and its relation to the Schrödinger–Poisson problem*, SIAM J. Math. Anal., 22 (1991), pp. 957–972.

[54] C. Weissbuch and B. Vinter, *Quantum Semi-conductor Strucutres*, Academic Press, New York, 1991.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 982–995

ON THE MINIMUM PROBLEM FOR NONCONVEX SCALAR FUNCTIONALS*

SANDRO ZAGATTI†

Abstract. We study the minimum problem for scalar nonconvex functionals defined on Sobolev maps satisfying a Dirichlet boundary condition and refine well-known existence results under standard regularity assumptions.

Key words. minimum problem, weak continuity, extremality

AMS subject classification. 49J45

DOI. 10.1137/040612506

1. Introduction.

We consider functionals of the type

$$\mathcal{F}(u) = \int_\Omega f(x, Du(x))dx,$$

where Ω is an open bounded subset of $\mathbb{R}^n$, the Lagrangian $f = f(x, \xi)$ maps $\Omega \times \mathbb{R}^n$ into $\mathbb{R}$, and the competing real valued maps u vary in suitable Sobolev spaces and satisfy a Dirichlet boundary condition.

Our interest is focused on nonsemicontinuous functionals, corresponding to integrands f which are nonconvex in the second variable. In this case the so-called direct method of the calculus of variations does not guarantee the existence of minimizers and the study of the minimum problem associated to $\mathcal{F}$ requires further techniques based on the theory of differential inclusions and on the study of Hamilton–Jacobi equations.

We describe briefly the idea of the procedure: introduce the lower convex envelope of f with respect to the second variable, usually denoted by f^{**}, and the corresponding relaxed functional

$$\overline{\mathcal{F}}(u) = \int_\Omega f^{**}(x, Du(x))dx;$$

if standard growth conditions on f are assumed, so that the functional is coercive on a suitable Sobolev space, the set S of minimizers of $\overline{\mathcal{F}}$ is nonempty and the solution of the minimum problem for the nonconvex functional $\mathcal{F}$ consists in finding an element $\tilde{u} \in S$ such that

$$f^{**}(x, D\tilde{u}(x)) = f(x, D\tilde{u}(x)) \quad \text{for almost every} \quad x \in \Omega. \tag{1}$$

The realization of this program has seen considerable improvements in the last years and we mention the result contained in [14], [6], [12], [7], [1], [3], and the monograph [8] for a review. In the homogeneous case, corresponding to a Lagrangian f which does not depend explicitly on the variable x, the program sketched above has been performed by different methods. Assuming that coercivity holds, the main hypotheses are the following:

*Received by the editors July 29, 2004; accepted for publication (in revised form) May 9, 2005; published electronically December 6, 2005.
http://www.siam.org/journals/sima/37-3/61250.html

†S.I.S.S.A. Via Beirut 2/4 I-34014 Trieste, Italy (zagatti@sissa.it).

(i) the bipolar f^{**} is affine on each connected component of the set

$$X \doteq \{\xi \in \mathbb{R}^n : f(\xi) > f^{**}(\xi)\}\,; \tag{2}$$

(ii) all the minimizers of $\overline{\mathcal{F}}$ are (classically) differentiable at almost every point $x \in \Omega$.

Condition (i) cannot be avoided. As shown in [4] (see also Theorem 2 in section 3 of this paper), if local affinity on X is violated the nonconvex functional fails to have minimizers for arbitrary boundary data and only in special cases, determined by certain compatibility conditions between the datum and the Lagrangian, the minimum problem can be solved. On the other hand it is an open question to establish if condition (ii) can be weakened or suppressed.

The paper [11] treats the nonhomogeneous case (where f depends explicitly on x), and in order to prove the existence of minimizers the authors impose the conditions listed below.

(a) The map $\xi \mapsto f(x,\xi)$ is affine on each connected component of the set

$$X(x) \doteq \{\xi \in \mathbb{R}^n : f(x,\xi) > f^{**}(x,\xi)\}\,;$$

that is to say, assuming for simplicity that $X(x)$ is connected for every x, there exists a C^1 field m and a continuous function q such that

$$\xi \mapsto f^{**}(x,\xi) = \langle m(x), \xi\rangle + q(x) \quad \text{for every } \xi \in X(x).$$

(b) The boundary of the set on which the divergence of the field m vanishes is zero.

(c) The bipolar f^{**} satisfies a strict convexity condition outside $X(x)$; i.e., there exists a nonnegative increasing function ω, vanishing only in zero, such that

$$f^{**}\left(x, \frac{\xi+\eta}{2}\right) \le \frac{1}{2} f^{**}(x,\xi) + \frac{1}{2} f^{**}(x,\eta) - \omega(|\xi-\eta|),$$

for every $x \in \Omega$ such that $X(x)$ is nonempty, for every $\xi \in E(x)$, and for every $\eta \in \mathbb{R}^n \setminus E(x)$, where $E(x) = \{\xi \in \mathbb{R}^n : f^{**}(x,\xi) = \langle m(x), \xi\rangle + q(x)\}$.

(d) All the minimizers of the relaxed functional are continuous on Ω and classically differentiable almost everywhere on Ω.

In addition, in order to guarantee condition (d), a *uniform strict convexity* condition at infinity is imposed so that, in particular, the set valued map $x \mapsto X(x)$ is necessarily uniformly bounded on Ω.

In this paper we propose a new approach to the problem by using the notion of *integro-maximal minimizer*. We now give a brief description of this method applied to the homogeneous case, where it is more transparent. Assume the coercivity of the functional and that conditions (i) and (ii) hold, and consider the nonempty set S of minimizers of $\overline{\mathcal{F}}$ which is compact with respect to strong convergence in $L^1(\Omega)$. The functional

$$S \ni u \mapsto \int_\Omega u(x)dx$$

is continuous with respect to the same topology and then, by the Weierstrass theorem, there exists an element $\overline{u}$ which maximizes it on S, that is to say,

$$\int_\Omega \overline{u}(x)dx \ge \int_\Omega u(x)dx \quad \forall u \in S.$$

The map $\overline{u}$ minimizes $\mathcal{F}$. Roughly speaking, the maximization of the integral *forces* the gradient of $\overline{u}$ to avoid the interior of the set X defined in (2) and to take values at its boundary so that (1) is satisfied.

In the general nonhomogeneous case, the argument is more elaborate, but essentially analogous.

By this method we are able to refine the results contained in the aforementioned papers. We treat separately the homogeneous and the nonhomogeneous case (sections 3 and 4, respectively) since in the first one, which can be taken as a guide for the general one, the hypotheses are minimal. We stress that in section 3 we give essentially a new proof of results already achieved in other papers and, for completeness, we recall also a well-known result [4] which shows that the local affinity of f^{**} is not only sufficient but also necessary for the existence of minimizers for arbitrary boundary data.

In the nonhomogeneous case (section 4), we improve the result contained in [11] by weakening the assumptions specified above. We maintain (d), and in (a)–(b) we require that m is a continuous Sobolev field satisfying conditions which strictly subsume (b) while condition (c) is suppressed. In addition, we do not need strict convexity at infinity so that the set valued map $\Omega \ni x \mapsto X(x)$ is allowed to be not necessarily uniformly bounded.

2. Preliminaries and notations. We denote, respectively, by $\langle\cdot,\cdot\rangle$ and $|\cdot|$ the inner product and the euclidean norm in $\mathbb{R}^n$. For $x \in \mathbb{R}^n$ and $r > 0$, $B(x,r)$ is the open ball in $\mathbb{R}^n$ of center x and of radius r; $\mu(E)$ denotes the Lebesgue measure of a (Lebesgue measurable) subset E of $\mathbb{R}^n$. Throughout the paper Ω is an open bounded subset of $\mathbb{R}^n$. We use the spaces $L^p(\Omega)$, $W^{1,p}(\Omega,\mathbb{R})$, and $W_0^{1,p}(\Omega,\mathbb{R})$, for $1 \le p \le \infty$, with their usual (strong and weak) topologies and use the precise representative of Sobolev functions as defined in [10]; the conjugate exponent of $p \ge 1$ is written $p' \doteq p/(p-1)$. By 1_E we denote the characteristic function of a subset E of $\mathbb{R}^n$, i.e.,

$$1_E(x) \doteq \begin{cases} 1 \text{ for } & x \in E, \\ 0 \text{ for } & x \in \mathbb{R}^n \setminus E. \end{cases}$$

In the proof of our main result we shall need the following well-known argument (see [9] and [14]).

LEMMA 1. *Let U be an open subset of $\mathbb{R}^n$, $p \in [1,\infty]$, $u \in W^{1,p}(U,\mathbb{R})$, $r > 0$. Assume that there exists a point $x_0 \in U$ such that u is differentiable at x_0 (with differential denoted by $Du(x_0)$). Then there exists $\rho > 0$ and maps $u^+, u^- \in W^{1,p}(U,\mathbb{R})$ with the following properties:*

$$\overline{B(x_0,\rho)} \subseteq U; \tag{3}$$

$$u^\pm - u \in W_0^{1,p}(U,\mathbb{R}); \tag{4}$$

$$u(x) \le u^+(x) \textit{ for a.e. } x \in B(x_0,\rho); \tag{5}$$

$$u(x) \ge u^-(x) \textit{ for a.e. } x \in B(x_0,\rho); \tag{6}$$

$$A^+ \doteq \{x \in U\colon u^+(x) > u(x)\} \textit{ is nonempty and } A^+ \subseteq B(x_0,\rho); \tag{7}$$

$$A^- \doteq \{x \in U\colon u^-(x) < u(x)\} \textit{ is nonempty and } A^- \subseteq B(x_0,\rho); \tag{8}$$

$$\begin{cases} |Du^\pm(x) - Du(x_0)| = r & \textit{for a.e. } x \in A^\pm; \\ Du^\pm(x) = Du(x) & \textit{for a.e. } x \in U \setminus A^\pm; \end{cases} \tag{9}$$

$$\int_\Omega u^+ dx > \int_\Omega u\, dx; \tag{10}$$

(11) $$\int_\Omega u^- dx < \int_\Omega u\, dx;$$

for every field $l \in W^{1,p'}(B(x_0,\rho),\mathbb{R}^n)$, *we have*

(12) $$\int_{B(x_0,\rho)} \langle l, (Du^\pm - Du)\rangle dx = -\int_{B(x_0,\rho)} \operatorname{div}\, l \cdot (u^\pm - u) dx.$$

Proof. We give the proof for the map u^+; it will be clear that the construction of u^- is analogous.

Choose a positive ρ such that (3) is satisfied. Since u is differentiable at x_0 there exists $s > 0$ such that

(13) $$\frac{s}{r} \le \rho$$

and

(14) $$\left| \frac{u(x) - u(x_0) - \langle Du(x_0), x - x_0\rangle}{x - x_0} \right| \le \frac{r}{2} \qquad \forall x \in B\left(x_0, \frac{s}{r}\right).$$

Inserting (13) into (14) we have, in particular,

(15) $$u(x) - u(x_0) - \langle Du(x_0), x - x_0\rangle \ge -\frac{s}{2} \qquad \forall x \in B\left(x_0, \frac{s}{r}\right).$$

Define a map w on $B\left(x_0, \frac{s}{r}\right)$ by setting

(16) $$w(x) \doteq \max\left\{ u(x), u(x_0) + \langle Du(x_0), x - x_0\rangle + \frac{s}{4} - r|x - x_0| \right\}$$

and introduce the set

(17) $$A^+ \doteq \left\{ x \in B\left(x_0, \frac{s}{r}\right) : w(x) > u(x) \right\}.$$

The function u is differentiable and then continuous at the point x_0; since the map

(18) $$B(x_0,\rho) \ni x \mapsto u(x_0) + \langle Du(x_0), x - x_0\rangle + \frac{s}{4} - r|x - x_0|$$

is also continuous at the same point and its value is strictly greater than $u(x_0)$, it follows that the set A^+ is nonempty. In addition the map defined in (18) is Lipschitz continuous and then belongs to $W^{1,\infty}\left(B\left(x_0, \frac{s}{r}\right), \mathbb{R}\right)$, which is contained in $W^{1,p}\left(B\left(x_0, \frac{s}{r}\right), \mathbb{R}\right)$; hence, by Stampacchia's theorem, w belongs to the space $W^{1,p}\left(B\left(x_0, \frac{s}{r}\right), \mathbb{R}\right)$ and we have

(19) $$\begin{cases} Dw(x) = Du(x_0) - rD|x - x_0| \quad \text{for a.e. } x \in A, \\ Dw(x) = Du(x) \quad \text{for a.e. } x \in B\left(x_0, \frac{s}{r}\right) \setminus A. \end{cases}$$

We observe that, for any $x \in B\left(x_0, \frac{s}{r}\right)$ such that $|x - x_0| > \frac{3}{4}\frac{s}{r}$, we have, by (15),

$$\begin{aligned} &u(x_0) + \langle Du(x_0), x - x_0\rangle + \frac{s}{4} - r|x - x_0| \\ &\quad < u(x_0) + \langle Du(x_0), x - x_0\rangle + \frac{s}{4} - \frac{3}{4}s \\ &\quad = u(x_0) + \langle Du(x_0), x - x_0\rangle - \frac{s}{2} \\ &\quad \le u(x). \end{aligned}$$

Hence

$$A^+ \subseteq B\left(x_0, \frac{s}{r}\right) \tag{20}$$

and the map w coincides with u on $B\left(x_0, \frac{s}{r}\right) \setminus \overline{B\left(x_0, \frac{3}{4}\frac{s}{r}\right)}$. By definition (16) and by standard notions on Sobolev functions (see, for example, [10]), this implies that

$$(w-u)\,|_{B\left(x_0, \frac{s}{r}\right)} \in W_0^{1,p}\left(B\left(x_0, \frac{s}{r}\right), \mathbb{R}\right). \tag{21}$$

Now we set

$$u^+(x) \doteq \begin{cases} w(x) & \text{for } x \in B\left(x_0, \frac{s}{r}\right), \\ u(x) & \text{for } x \in U \setminus B\left(x_0, \frac{s}{r}\right). \end{cases} \tag{22}$$

From (16), (20) (and the consideration which follows), (21), and (22) we obtain that u^+ lies in $W^{1,p}(U, \mathbb{R})$. We see now that conditions (4)–(12) hold. Property (4) is a trivial consequence of (16), (21), and (22); (5) follows trivially from definitions (16) and (22). Condition (7) has been already proved and, in order to show (9), we observe that, by (19) and (22), we have that $|Du^+(x) - Du(x_0)| = rD|x - x_0| = r$ for almost every $x \in A$ and that $Du^+(x) = Du(x)$ for almost every $x \in U \setminus A$. Inequality (10) is a consequence of (5) and of the fact that the open set A, being nonempty, has positive measure. Formula (12) follows from (22), (21), and the divergence theorem.

The map u^- can be defined by changing the sign of the second argument in definition (16) and, by analogous arguments, obtaining properties (6), (8), and (11) in place of (5), (7), and (10).

Hence the proof is finished.

3. The homogeneous case. We consider a lower semicontinuous function $f\colon \mathbb{R}^n \to \mathbb{R}$, consider its bipolar f^{**}, and define the set

$$X \doteq \left\{\xi \in \mathbb{R}^d \colon f(\xi) > f^{**}(\xi)\right\}. \tag{23}$$

The lower semicontinuity of f and the continuity of f^{**} imply that X is open.

For $p \in [1, \infty]$ and given $u_0 \in W^{1,p}(\Omega, \mathbb{R})$, we introduce the functionals

$$\mathcal{F}(u) = \int_\Omega f(Du(x))dx, \qquad u \in u_0 + W_0^{1,p}(\Omega, \mathbb{R});$$

$$\overline{\mathcal{F}}(u) = \int_\Omega f^{**}(Du(x))dx, \qquad u \in u_0 + W_0^{1,p}(\Omega, \mathbb{R}).$$

We need the following hypotheses.

Hypothesis H1. The function f^{**} is locally affine on the set X. In other words, for every $\eta \in X$ there exists $r(\eta) > 0$, $m(\eta) \in \mathbb{R}^n$, and $q(\eta) \in \mathbb{R}$ such that $B(\eta, r(\eta)) \subset X$,

$$f^{**}(\xi) = \langle m(\eta), \xi\rangle + q(\eta) \quad \forall \xi \in B(\eta, r(\eta)) \quad \text{and} \tag{24}$$

$$f^{**}(\xi) \geq \langle m(\eta), \xi\rangle + q(\eta) \quad \forall \xi \in \mathbb{R}^n. \tag{25}$$

In our result we require that the set of minimizers of the functional $\overline{\mathcal{F}}$ is nonempty and sequentially compact with respect to the strong topology of $L^1(\Omega)$. In addition we need all minimizers to be classically differentiable almost everywhere in Ω. These facts are ensured by the following conditions (see [3]).

Hypothesis H2.

$$\begin{array}{lll} 1 < p \le n, & \exists\, a,b,c,d > 0: & a|\xi|^p - b \le f(\xi) \le c|\xi|^p + d \quad \forall \xi \in \mathbb{R}^n, \\ n < p < \infty, & \exists\, a,b > 0: & f(\xi) \ge a|\xi|^p - b \quad \forall \xi \in \mathbb{R}^n, \\ p = \infty & \exists\, K > 0: & f(\xi) = +\infty \quad \forall \xi \in \mathbb{R}^n: \ |\xi| \ge K. \end{array}$$

Remark 1. It is well known that the convexity of f^{**} is equivalent to the sequential lower semicontinuity of $\overline{\mathcal{F}}$ with respect to the weak (weak* when $p = \infty$) topology of $W^{1,p}(\Omega, \mathbb{R})$. Hence Hypothesis H2 guarantees that the functional $\overline{\mathcal{F}}$ admits at least one minimizer.

Remark 2. Let $u_0 \in W^{1,p}(\Omega, \mathbb{R})$. If f satisfies H2, by previous considerations the set S of minimizers of $\overline{\mathcal{F}}$ is nonempty. Suppose that there exists $\tilde{u}$ such that $D\tilde{u}(x) \in \mathbb{R}^n \setminus X$ for almost every $x \in \Omega$. Then $\tilde{u}$ is a minimizer of $\mathcal{F}$. Indeed, since $f(\xi) \ge f^{**}(\xi)$ for every $\xi \in \mathbb{R}^n$, we have that $\inf \mathcal{F} \le \min \overline{\mathcal{F}}$. Since $f(D\tilde{u}(x)) = f^{**}(D\tilde{u}(x))$ for almost every $x \in \Omega$, it turns out that

$$\mathcal{F}(\tilde{u}) = \int_\Omega f(D\tilde{u}(x))dx = \int_\Omega f^{**}(D\tilde{u}(x))dx = \min \overline{\mathcal{F}} \le \inf \mathcal{F}.$$

We are ready to prove the main result of this section.

THEOREM 1. *Let $p \in [1, \infty]$ and $u_0 \in W^{1,p}(\Omega, \mathbb{R})$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a lower semicontinuous function satisfying* H1. *Assume that the set S of minimizers of $\overline{\mathcal{F}}$ is nonempty and sequentially compact with respect to the strong topology of $L^1(\Omega)$. Assume that all the elements of S are differentiable almost everywhere in Ω. Then the functional $\mathcal{F}$ attains its minimum.*

Proof. The functional

$$S \ni u \mapsto \int_\Omega u(x)dx$$

is continuous with respect to the strong topology of $L^1(\Omega)$. Hence, by the Weierstrass theorem, there exists $\overline{u} \in S$ such that

$$\int_\Omega \overline{u}(x)dx \ge \int_\Omega u(x)dx \quad \forall u \in S. \tag{26}$$

Claim. The map $\overline{u}$ minimizes $\mathcal{F}$.

We prove that $D\overline{u}(x)$ lies in the set $\mathbb{R}^n \setminus X$ for every $x \in \Omega$ such that $\overline{u}(x)$ is differentiable at x, so that the claim will be proved by Remark 2.

We argue by contradiction: assume that there exists a point $x_0 \in \Omega$ such that $\overline{u}$ is differentiable at x_0 and $D\overline{u}(x_0) \in X$. Recalling Hypothesis H1, we select $\overline{r}$ in such a way that

$$0 < \overline{r} < r\left(D\overline{u}(x_0)\right). \tag{27}$$

Taking a positive we apply Lemma 1 with Ω, $\overline{u}$, $\overline{r}$ in place of U, u, r, respectively, and obtain a map $\overline{u}^+$ satisfying the following conditions:

$$\overline{u}^+ - \overline{u} \in W_0^{1,p}(\Omega, \mathbb{R}); \tag{28}$$

$$A \doteq \{x \in \Omega \colon \overline{u}^+(x) > \overline{u}(x)\} \text{ is nonempty and } A \subseteq B(x_0, \rho); \tag{29}$$

$$\begin{cases} |D\overline{u}^+(x) - D\overline{u}(x_0)| = \overline{r} & \text{for a.e. } x \in A; \\ D\overline{u}^+(x) = D\overline{u}(x) & \text{for a.e. } x \in \Omega \setminus A; \end{cases} \tag{30}$$

$$\int_\Omega \overline{u}^+(x)dx > \int_\Omega \overline{u}(x)dx; \tag{31}$$

$$\int_A D\overline{u}^+(x)dx = \int_A D\overline{u}(x)dx. \tag{32}$$

We have suppressed the superscript + on the set A^+ and remark that equality (32) follows easily from (12), (29), and (30).

First, we prove that $\overline{u}^+$ lies in S. By (28) it belongs to $\overline{u} + W_0^{1,p}(\Omega, \mathbb{R}) = u_0 + W_0^{1,p}(\Omega, \mathbb{R})$ and we can compute, using (30),

$$\begin{aligned} \overline{\mathcal{F}}(\overline{u}^+) &= \int_{\Omega\setminus A} f^{**}(D\overline{u}^+(x))dx + \int_A f^{**}(D\overline{u}^+(x))dx \\ &= \int_{\Omega\setminus A} f^{**}(D\overline{u}(x))dx + \int_A f^{**}(D\overline{u}^+(x))dx. \end{aligned} \tag{33}$$

Again by (30) and by the choice (27) of $\overline{r}$, we have that

$$D\overline{u}^+(x) \in B\left(D\overline{u}(x_0), r\left(D\overline{u}(x_0)\right)\right) \qquad \text{a.e. in } A. \tag{34}$$

Setting for convenience $m = m\left(D\overline{u}(x_0)\right)$, $q = q\left(D\overline{u}(x_0)\right)$, and using (24), (25), and (32) we have

$$\begin{aligned} \int_A f^{**}(D\overline{u}^+(x))dx &= \int_A \left(\left\langle m, D\overline{u}^+(x)\right\rangle + q\right) dx \\ &= \left\langle m, \int_A D\overline{u}^+(x)dx \right\rangle + q\mu(A) \\ &= \left\langle m, \int_A D\overline{u}(x)dx \right\rangle + q\mu(A) \\ &= \int_A \left(\langle m, D\overline{u}(x)\rangle + q\right) dx \leq \int_A f^{**}(D\overline{u}(x))dx. \end{aligned} \tag{35}$$

Inserting (35) into (33), we have

$$\overline{\mathcal{F}}(\overline{u}^+) = \int_A f^{**}(D\overline{u}^+(x))dx \leq \int_A f^{**}(D\overline{u}(x))dx = \min \mathcal{F}. \tag{36}$$

Hence $\overline{u}^+$ belongs to S so that (31) contradicts (26), and this ends the proof.

Remark 3. In the proof of Theorem 1, in place of the map $\overline{u}$ we may introduce a map $\underline{u} \in S$ such that $\int_\Omega \underline{u}(x)dx \leq \int_\Omega u(x)dx$ for every $u \in S$. By an analogous indirect argument, using the map $\underline{u}^-$ instead of the map $\overline{u}^+$, it can be shown that $\underline{u}$ minimizes $\mathcal{F}$.

According to Theorem 1, the local affinity of f^{**} appears a sufficient condition for the existence of minimizers of $\mathcal{F}$ for arbitrary data. We see now that it is also necessary. This fact is well known and we recall it for completeness. Consider affine boundary data: given $\eta \in \mathbb{R}^n$, set

$$u_\eta(x) \doteq \langle \eta, x \rangle, \qquad x \in \Omega \tag{37}$$

and write

$$\mathcal{F}_\eta(u) = \int_\Omega f(Du(x))dx, \qquad u \in u_\eta + W_0^{1,p}(\Omega, \mathbb{R});$$

$$\overline{\mathcal{F}}_\eta(u) = \int_\Omega f^{**}(Du(x))dx, \qquad u \in u_\eta + W_0^{1,p}(\Omega, \mathbb{R}).$$

We have the following result, the full proof of which is in [4].

THEOREM 2. *Assume that $\mathcal{F}_\eta$ attains its minimum for every $\eta \in \mathbb{R}^n$. Then f^{**} is locally affine on the set X.*

4. The nonhomogeneous case. We consider a lower continuous function $f : \Omega \times \mathbb{R}^n \to \mathbb{R}$ and its bipolar $f^{**} : \Omega \times \mathbb{R}^n \to \mathbb{R}$. For every $x \in \Omega$ we introduce the set

$$X(x) \doteq \{\xi \in \mathbb{R}^n : f(x, \xi) > f^{**}(x, \xi)\}. \tag{38}$$

Given $p \in [1, \infty]$ and $u_0 \in W^{1,p}(\Omega, \mathbb{R})$, we consider the functionals

$$\mathcal{F}(u) = \int_\Omega f(x, Du(x))dx, \qquad u \in u_0 + W^{1,p}(\Omega, \mathbb{R});$$

$$\overline{\mathcal{F}}(u) = \int_\Omega f^{**}(x, Du(x))dx, \qquad u \in u_0 + W^{1,p}(\Omega, \mathbb{R}).$$

We need a version of the affinity condition (H1) introduced in the previous section which refines the one used in [11], which is recalled in points (a)–(b) in the introduction of this paper. We assume that the function f^{**} is continuous on $\Omega \times \mathbb{R}^n$, hence for every $x \in \Omega$ the set $X(x)$ is the countable union of its connected components $X_j(x)$,

$$X(x) = \bigcup_{j=1}^{\infty} X_j(x). \tag{39}$$

Hypothesis H3. For every $j \in \mathbb{N}$ we set

$$\Omega_{X_j} \doteq \{x \in \Omega : X_j(x) \text{ is nonempty}\} \tag{40}$$

and assume that, for every $j \in \mathbb{N}$, there exists a field $m_j \in W^{1,p'}_{\mathrm{loc}}(\Omega_{X_j}, \mathbb{R}^n) \cap C^0(\Omega_{X_j}, \mathbb{R}^n)$ and a map $q_j \in C^0(\Omega_{X_j}, \mathbb{R})$ such that

$$f^{**}(x, \xi) = \langle m_j(x), \xi \rangle + q_j(x) \quad \forall x \in \Omega_{X_j} \text{ and } \forall \xi \in X_j(x) \tag{41}$$

and

$$f^{**}(x, \xi) \geq \langle m_j(x), \xi \rangle + q_j(x) \quad \forall x \in \Omega_{X_j} \text{ and } \forall \xi \in \mathbb{R}^n. \tag{42}$$

We assume that for every j the function $\operatorname{div} m_j$ admits a measurable extension on the whole Ω, still denoted by $\operatorname{div} m_j$, satisfying the following condition:

(43) For every $j \in \mathbb{N}$ and for almost every $x \in \Omega$, there exists an open neighborhood U of x contained in Ω such that either $\operatorname{div} m_j \geq 0$ almost everywhere on U or $\operatorname{div} m_j \leq 0$ almost everywhere on U.

In particular, if the boundary of Ω_{X_j} has measure zero, we may extend $\operatorname{div} m_j$ by zero on $\Omega \setminus \Omega_{X_j}$. We stress that if m_j is of class C^1, condition (43) is automatically satisfied if hypothesis (b) in the introduction holds.

Remark 4. Since f is lower semicontinuous and f^{**} is continuous, the set $X_j(x)$ is open for every $j \in \mathbb{N}$ and for every $x \in \Omega$. Hence, recalling (39) and (40), the set Ω_{X_j} is also open for every j. In particular, given $x_0 \in \Omega_{X_j}$ and $\xi_0 \in X_j(x_0)$, there exists $r > 0$ such that for every $x \in \Omega$ and for every $\xi \in \mathbb{R}^n$ such that $|x - x_0| < r$ and $|\xi - \xi_0| < r$, we have $x \in \Omega_{X_j}$ and $\xi \in X_j(x)$. To see this fact, take $x_0 \in \Omega_{X_j}$ and $\xi_0 \in X_j(x_0)$ and assume, by contradiction, that there exists two sequences (x_k) and (ξ_k) such that $x_k \to x_0$ and $\xi_k \to \xi_0$ and $\xi_k \notin X_j(x_k)$. By the definition of f^{**} and of X_j this means that $f(x_k, \xi_k) = f^{**}(x_k, \xi_k)$, and from this we obtain

$$f(x_0, \xi_0) \leq \liminf f(x_k, \xi_k) = \lim f^{**}(x_k, \xi_k) = f^{**}(x_0, \xi_0).$$

This contradicts that ξ_0 belongs to $X_j(x_0)$. The fact that the Ω_{X_j}'s are open justifies the assumption that m_j belongs to $W^{1,p'}(\Omega_{X_j}, \mathbb{R}^n)$, which is made in Hypothesis H3.

Definition 1. *Fix $j \in \mathbb{N}$. By condition (43) there exists a null set $N \subset \Omega$ such that for every $x \in \Omega \setminus N$ there exists $\rho_x > 0$ such that $\operatorname{div} m_j$ has definite sign almost everywhere on $B(x, \rho)$ for every $\rho \in]0, \rho_x[$. Call $B_j^+(x, \rho)$ the open balls on which $\operatorname{div} m_j \geq 0$ a.e. and $B_j^-(x, \rho)$ the open balls on which $\operatorname{div} m_j \leq 0$ a.e. The collection*

$$\left\{B_j^+(x, \rho), B_j^-(x, \rho) : x \in \Omega, \rho \in]0, \rho_x[\right\}$$

is a Vitali covering of $\Omega \setminus N$, hence, by the Vitali covering lemma, there exists a sequence of positive numbers (ρ_k) and a sequence (x_k) in $\Omega \setminus N$ such that the elements of the countable family

$$\left\{B_j^+(x_k, \rho_k), B_j^-(x_k, \rho_k), k \in \mathbb{N}\right\}$$

are pairwise disjoint and the set

$$(\Omega \setminus N) \setminus \left(\cup_{k \in \mathbb{N}} \left(B_j^-(x_k, \rho_k) \cup B_j^+(x_k, \rho_k)\right)\right)$$

has measure zero.

Now, for every $j \in \mathbb{N}$, we set

$$V_j^+ \doteq \bigcup_{k \in \mathbb{N}} B_j^+(x_k, \rho_k), \qquad V_j^- \doteq \bigcup_{k \in \mathbb{N}} B_j^-(x_k, \rho_k) \tag{44}$$

and define the map

$$\gamma_j \doteq 1_{V_j^+} - 1_{V_j^-}. \tag{45}$$

Clearly, for every $j \in \mathbb{N}$, γ_j belongs to $L^\infty(\Omega, \mathbb{R})$, $|\gamma_j| \equiv 1$ almost everywhere and, by (43), satisfies the following property.

(46) *For every $j \in \mathbb{N}$ and for almost every $x \in \Omega$, there exists an open ball B contained in Ω such that γ_j is constant almost everywhere on B. If $\gamma_j = +1$ a.e. on B, then $\operatorname{div} m_j \geq 0$ a.e. on B; if $\gamma_j = -1$ a.e. on B, then $\operatorname{div} m_j \leq 0$ a.e. on B.*

Remark 5. We stress that the function γ_j does not coincide with the function $\operatorname{sign}(\operatorname{div} m_j)$ since the value zero for $\operatorname{div} m_j$ is treated differently.

Remark 6. Let $u_0 \in W^{1,p}(\Omega, \mathbb{R})$. By Hypothesis H2' below and by the convexity of f^{**} with respect to the second variable, the set S of minimizers of $\overline{\mathcal{F}}$ is nonempty. Assume that there exists $\tilde{u} \in S$ such that $D\tilde{u}(x) \in \mathbb{R}^n \setminus X(x)$ for almost every $x \in \Omega$. Then $\tilde{u}$ is a minimizer of $\mathcal{F}$. Indeed, since $f(x,\xi) \geq f^{**}(x,\xi)$ for every pair $(x,\xi) \in \Omega \times \mathbb{R}^n$, we have that $\inf \mathcal{F} \geq \min \overline{\mathcal{F}}$. Since, by hypothesis, $f^{**}(x, D\tilde{u}(x)) = f(x, D\tilde{u}(x))$ for almost every $x \in \Omega$, we have

$$\mathcal{F}(\tilde{u}) = \int_\Omega f(x, D\tilde{u}(x))dx = \int_\Omega f^{**}(x, D\tilde{u}(x))dx = \min \overline{\mathcal{F}} \leq \inf \mathcal{F}.$$

As in the previous section, we need differentiability almost everywhere of the minimizers of the functional $\overline{\mathcal{F}}$ and the sequential compactness of the set of its minimizers. Such properties are ensured by the generalization of Hypothesis H2 (see the reference quoted there).

Hypothesis H2'.

$$\begin{array}{lll} 1 < p \leq n, & \exists\, a,b,c,d > 0: & a|\xi|^p - b \leq f(x,\xi) \leq c|\xi|^p + d \ \forall (x,\xi) \in \Omega \times \mathbb{R}^n, \\ n < p < \infty, & \exists\, a,b > 0: & f(x,\xi) \geq a|\xi|^p - b \ \forall (x,\xi) \in \Omega \times \mathbb{R}^n, \\ p = \infty, & \exists\, K > 0: & f(x,\xi) = +\infty \ \forall (x,\xi) \in \Omega \times \mathbb{R}^n: \ |\xi| \geq K. \end{array}$$

Example. An easy example of an integrand satisfying the above hypotheses is the following. Let $a \in C(\overline{\Omega})$ be strictly positive for some x and let $m \in W^{1,\frac{4}{3}}(\Omega, \mathbb{R}^n) \cap C(\Omega)$ be such that $\operatorname{div} m$ satisfies condition (43). Then set

$$f(x,\xi) = \big(\xi^2 - a(x)\big)^2 + \langle m(x), \xi \rangle.$$

Clearly f fulfills the requirements (with $p = 4$) and the set $X(x)$, which is nonempty whenever $a(x) > 0$, coincides with the ball $B(0, \sqrt{a(x)})$. Indeed, we have

$$f^{**}(x,\xi) = \begin{cases} \langle m(x), \xi \rangle, & \xi \in X(x), \\ \big(\xi^2 - a(x)\big)^2 + \langle m(x), \xi \rangle, & \xi \in \mathbb{R}^n \setminus X(x). \end{cases}$$

We are ready to prove the main result of this section.

THEOREM 3. *Let $p \in [1,\infty]$ and $u_0 \in W^{1,p}(\Omega, \mathbb{R})$. Let $f : \Omega \times \mathbb{R}^n \to \mathbb{R}$ be a lower semicontinuous function satisfying* H3*. Assume that the set S of minimizers of $\overline{\mathcal{F}}$ is nonempty and sequentially compact with respect to the strong topology of $L^1(\Omega)$. Assume that all the elements of S are differentiable almost everywhere in Ω. Then the functional $\mathcal{F}$ attains its minimum.*

Proof. Recall the properties of the map γ_1 from Definition 1.

Step 1. The functional

$$S \ni u \mapsto \int_\Omega \gamma_1(x) u(x) dx$$

is continuous with respect to the strong topology of $L^1(\Omega)$; hence the set

$$S_1 \doteq \left\{ u_1 \in S : \int_\Omega \gamma_1(x) u_1(x) dx \geq \int_\Omega \gamma_1(x) u(x) dx \ \forall u \in S \right\} \tag{47}$$

is nonempty and sequentially compact with respect to the same topology.

Claim. The set

$$\{x \in \Omega : Du_1(x) \in X_1(x)\}$$

has measure zero for every $u_1 \in S_1$.

We prove that given any $u_1 \in S_1$, $Du_1(x) \in \mathbb{R}^n \setminus X_1(x)$ for every $x \in \Omega$ at which u_1 is differentiable. Assume, by contradiction, that there exists $x_0 \in \Omega$ at which u_1 is differentiable with $Du_1(x_0) \in X_1(x_0)$. Recalling Remark 4, we can find $r > 0$ and $\rho > 0$ such that γ_1 is constant almost everywhere on $B(x_0, \rho)$ and

$$B(x_0, \rho) \subseteq \Omega_{X_1}, \qquad \overline{B(Du_1(x_0), r)} \subseteq X_1(x). \tag{48}$$

Suppose, to fix ideas, that

$$\gamma_1 = +1 \quad \text{a.e. in } B(x_0, \rho); \tag{49}$$

that is to say, recalling Definition 1 and (45),

$$\operatorname{div} m_1 \geq 0 \quad \text{a.e. in } B(x_0, \rho). \tag{50}$$

We apply Lemma 1 with r and ρ chosen above and with the map u_1 in place of u, obtaining a map u_1^+ satisfying (4), (5), (7), (9), (10), and (12). By (9) and (48) we have

$$Du_1^+(x) \in X_1(x) \quad \text{for a.e. } x \in A^+. \tag{51}$$

First we prove that u_1^+ lies in S. Property (4) implies that u_1^+ belongs to $u_1 + W_0^{1,p}(\Omega, \mathbb{R}) = u_0 + W_0^{1,p}(\Omega, \mathbb{R})$. Then, using (41) and (51), we have that

$$f^{**}(x, Du_1^+(x)) = \big\langle m_1(x), Du_1^+(x)\big\rangle + q_1(x) \quad \text{for a.e. } x \in A^+, \tag{52}$$

while (42) and (48) imply that

$$f^{**}(x, Du_1(x)) \geq \langle m_1(x), Du_1(x)\rangle + q_1(x) \quad \text{for a.e. } x \in B(x_0, \rho). \tag{53}$$

Hence (9) and (52) give

$$\begin{aligned} &\int_{B(x_0,\rho)} f^{**}(x, Du_1^+(x))dx \\ &= \int_{B(x_0,\rho)\setminus A^+} f^{**}(x, Du_1^+(x))dx + \int_{A^+} f^{**}(x, Du_1^+(x))dx \\ &= \int_{B(x_0,\rho)\setminus A^+} f^{**}(x, Du_1(x))dx + \int_{A^+} (\langle m_1(x), Du_1^+(x)\rangle + q_1(x))dx, \end{aligned} \tag{54}$$

while (53) gives

$$\begin{aligned} &\int_{B(x_0,\rho)} f^{**}(x, Du_1(x))dx \\ &\geq \int_{B(x_0,\rho)\setminus A^+} f^{**}(x, Du_1(x))dx + \int_{A^+} (\langle m_1(x), Du_1(x)\rangle + q_1(x))dx. \end{aligned} \tag{55}$$

Subtracting (54) from (55) we obtain

$$
\begin{aligned}
&\int_{B(x_0,\rho)} \left(f^{**}(x, Du_1(x))dx - f^{**}(x, Du_1^+(x))\right) dx \\
&\geq \int_{A^+} \langle m_1(x), Du_1(x) - Du_1^+(x)\rangle dx \\
&= \int_{B(x_0,\rho)} \langle m_1(x), Du_1(x) - Du_1^+(x)\rangle dx,
\end{aligned}
\tag{56}
$$

since Du_1 and Du_1^+ agree on $B(x_0,\rho) \setminus A^+$. Remarking that, by Hypothesis H3, the restriction of the field m_1 to the ball $B(x_0,\rho)$ belongs to $W^{1,p'}(B(x_0,\rho),\mathbb{R})$ and that the restrictions of u_1 and u_1^+ to the same ball lie in $W^{1,p}(B(x_0,\rho),\mathbb{R})$ (recall (4)), we have, by (12), (5), (50), and (56), the following.

$$
\begin{aligned}
&\int_{B(x_0,\rho)} (f^{**}(x, Du_1(x)) - f^{**}(x, Du_1^+(x)))dx \\
&\qquad \geq \int_{B(x_0,\rho)} \operatorname{div} m_1(x)(u_1^+(x) - u_1(x))dx \geq 0.
\end{aligned}
$$

Thus,

$$
\int_{B(x_0,\rho)} f^{**}(x, Du_1^+(x))dx \leq \int_{B(x_0,\rho)} f^{**}(x, Du_1(x))dx. \tag{57}
$$

Hence (9) and (57) imply that

$$
\begin{aligned}
\overline{\mathcal{F}}(u_1^+) &= \int_{\Omega\setminus B(x_0,\rho)} f^{**}(x, Du_1^+(x))dx \\
&\quad + \int_{B(x_0,\rho)} f^{**}(x, Du_1^+(x))dx \\
&\leq \int_{\Omega\setminus B(x_0,\rho)} f^{**}(x, Du_1(x)) \\
&\quad + \int_{B(x_0,\rho)} f^{**}(x, Du_1(x))dx \\
&= \overline{\mathcal{F}}(u_1),
\end{aligned}
$$

and it follows that u_1^+ lies in S. Recalling (5), (7), and (49) we have

$$
\int_\Omega \gamma_1(x)u_1^+(x)dx > \int_\Omega \gamma_1(x)u_1(x)dx, \tag{58}
$$

and this contradicts (47).

If, on the contrary, we have

$$
\gamma_1 = -1 \quad \text{a.e. in } B(x_0,\rho), \tag{59}
$$

we proceed analogously defining a map u_1^- as in Lemma 1 and obtaining the same contradiction inserting (6), (8), and (59) into (57) and into (58).

The claim is so proved.

Step 2. Consider now the set

$$S_2 \doteq \left\{u_2 \in S_1 : \int_\Omega \gamma_2(x)u_2(x)dx \geq \int_\Omega \gamma_2(x)u_1(x)dx \quad \forall u_1 \in S_1\right\},$$

which is nonempty and sequentially compact with respect to the strong topology of $L^1(\Omega)$ by the sequential compactness of S_1, the sequential lower semicontinuity of $\overline{\mathcal{F}}$, and the sequential continuity of the map

$$S_1 \ni u_1 \mapsto \int_\Omega \gamma_2(x)u_1(x)dx.$$

We reproduce the arguments of Step 1. Take $u_2 \in S_2$ and assume, by contradiction, that there exists $x_0 \in \Omega$ at which u_2 is differentiable with $Du_2(x_0) \in X_2(x_0)$. Since both $\operatorname{div} m_1$ and $\operatorname{div} m_2$ have sign in a suitably small ball centered at x_0, by the same argument used previously we get a contradiction and obtain that, for every $u_2 \in S_2$, the set

$$\{x \in \Omega : Du_2(x) \in X_2(x)\}$$

has measure zero for every $u_2 \in S_2$. Then, since $S_2 \subseteq S_1$, we have

$$\mu\left(\{x \in \Omega : Du_2(x) \in X_1(x) \cup X_2(x)\}\right) = 0 \quad \forall u_2 \in S_2. \tag{60}$$

Step 3. Iterating the procedure we obtain a sequence (S_k) of nonempty sequentially compact (with respect to strong topology in $L^1(\Omega)$) subsets of S given by

$$S_k \doteq \left\{u_k \in S_{k-1} : \int_\Omega \gamma_k u_k \, dx \geq \int_\Omega \gamma_k u_{k-1} \, dx \quad \forall u_{k-1} \in S_{k-1}\right\},$$

for which we have, as in (60),

$$\mu\left(\left\{x \in \Omega : Du_k(x) \in \cup_{j=1}^k X_j(x)\right\}\right) = 0 \quad \forall u_k \in S_k. \tag{61}$$

Since, by construction,

$$S \supseteq S_1 \supseteq \cdots \supseteq S_k \supseteq S_{k+1} \supseteq \cdots \quad \forall k, \tag{62}$$

the set

$$S_\infty \doteq \bigcap_{k\in\mathbb{N}} S_k \tag{63}$$

is nonempty.

Take an element $u_\infty \in S_\infty$; by (61), (62), and (63) we have that

$$\mu\left(\left\{x \in \Omega : Du_\infty(x) \in \cup_{j=1}^k X_j(x)\right\}\right) = 0 \quad \forall k \in \mathbb{N}.$$

Hence

$$\mu\left(\{x \in \Omega : Du_\infty(x) \in X_j(x)\}\right) = 0 \quad \forall j \in \mathbb{N},$$

and, consequently,

$$\begin{aligned} &\mu\left(\{x \in \Omega : Du_\infty(x) \in X(x)\}\right) \\ &= \mu\left(\bigcup_{j\in\mathbb{N}} \{x \in \Omega : Du_\infty(x) \in X_j(x)\}\right) \\ &\leq \sum_{j\in\mathbb{N}} \mu\left(\{x \in \Omega : Du_\infty(x) \in X_j(x)\}\right) = 0. \end{aligned} \tag{64}$$

Recalling Remark 6, (64) shows that any element of the nonempty set S_∞ is a minimizer of $\mathcal{F}$. This ends the proof.

Remark 7. It is hard to establish if the condition (43) is necessary for the existence of minimizers of the nonconvex functional since, if it is violated, what fails is a contradictory argument. It is clear from the proof of Theorem 3 that the gradient of any minimizer of the relaxed functional (and not only the one of the maximal function $\overline{u}$) must take values away from the open sets contained in $X_j(x)$ in which $\operatorname{div} m_j$ is strictly positive or negative. If not, indeed, it would be possible to reduce the value of the functional by a variation like the one performed in our argument. On the open sets in which $\operatorname{div} m_j$ vanishes, it is the maximization of the integral $\int_\Omega u\,dx$ which forces the gradient to be extremal. If (43) is not fulfilled, the rapid changes of sign of $\operatorname{div} m_j$ may induce oscillations of minimizing sequences so that the gradient of all the minimizers may coalesce inside the "bad" set $X_j(x)$.

REFERENCES

[1] P. Celada and S. Perrotta, *Minimizing non-convex multiple integrals: A density result*, Proc. Royal Soc. Edinburgh Sect. A, 130 (2000), pp. 721–741.

[2] P. Celada and S. Perrotta, *On the minimum problem for nonconvex, multiple integrals of product type*, Calc. Var. Partial Differential Equations, 12 (2001), pp. 371–398.

[3] P. Celada, G. Cupini, and M. Guidorzi, *A sharp attainment result for nonconvex variational problems*, Calc. Var. Partial Differential Equations, 20 (2004), pp. 301–228.

[4] A. Cellina, *On minima of a functional of the gradient: Necessary conditions*, Nonlinear Anal., 20 (1993), pp. 337–341.

[5] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, 1989.

[6] B. Dacorogna and P. Marcellini, *General existence theorems for Hamilton-Jacobi equations in the scalar and vectorial cases*, Acta Math., 178 (1997), pp. 1–37.

[7] B. Dacorogna and P. Marcellini, *Attainment of minima and implicit partial differential equations*, Ricerche Mat., 48 (1999), pp. 311–346.

[8] B. Dacorogna and P. Marcellini, *Implicit Partial Differential Equations*, Birkhäuser, Basel, 1999.

[9] F. S. De Blasi and G. Pianigiani, *On the Dirichlet Problem for Hamilton-Jacobi Equations. A Baire Category Approach*, preprint, 1997.

[10] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[11] I. Fonseca, N. Fusco, and P. Marcellini, *An existence result for a nonconvex variational problem via regularity*, Control. Optim. Calc. Var., 7 (2002), pp. 69–95.

[12] G. Friesecke, *A necessary and sufficient condition for nonattainment and formation of microstructure almost everywhere in scalar variational problems*, Proc. Royal Soc. Edinburgh Sect. A, 124 (1994), pp. 437–471.

[13] T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.

[14] S. Zagatti, *Minimization of functionals of the gradient by Baire's theorem*, SIAM J. Control Optim., 38 (2000), pp. 384–399.

SIAM J. MATH. ANAL.
Vol. 37, No. 3, pp. 996–1026

GLOBAL SOLUTIONS OF THE HUNTER–SAXTON EQUATION*

ALBERTO BRESSAN† AND ADRIAN CONSTANTIN‡

Abstract. We construct a continuous semigroup of weak, dissipative solutions to a nonlinear partial differential equation modeling nematic liquid crystals. A new distance functional, determined by a problem of optimal transportation, yields sharp estimates on the continuity of solutions with respect to the initial data.

Key words. nonlocal conservation law, global conservative solutions, optimal transportation, dissipative solution

AMS subject classifications. 35D05, 49J20, 76A15

DOI. 10.1137/050623036

1. Introduction.

In this paper we investigate the Cauchy problem

$$u_t + \left(\frac{u^2}{2}\right)_x = \frac{1}{4}\left(\int_{-\infty}^{x} - \int_{x}^{\infty}\right) u_x^2\, dx, \qquad u(0,x) = \bar{u}(x). \tag{1.1}$$

Formally differentiating the above equation with respect to the spatial variable x, we obtain

$$(u_t + uu_x)_x = \frac{1}{2}\, u_x^2, \tag{1.2}$$

whereas yet another differentiation leads to

$$u_{txx} + 2u_x u_{xx} + uu_{xxx} = 0. \tag{1.3}$$

Either of the forms (1.2) and (1.3) of the equation in (1.1) is known as the Hunter–Saxton equation. In this paper we analyze various concepts of solutions for the above equations and construct a semigroup of globally defined solutions. Moreover, we introduce a new distance functional, related to a problem of optimal transportation, which monitors the continuous dependence of solutions on the initial data. The global solutions constructed in this paper are dissipative. It is worthwhile to point out that it is possible (and simpler) to construct a corresponding optimal-transport functional for global energy-conserving solutions—see [BF] for an example of a metric valid for conservative solutions of an equation similar to (1.1).

Physical significance. The Hunter–Saxton equation describes the propagation of waves in a massive director field of a nematic liquid crystal [HS], with the orientation of the molecules described by the field of unit vectors $\mathbf{n}(t,x) = (\cos u(t,x), \sin u(t,x))$, x being the space variable in a reference frame moving with the linearized wave velocity, and t being a slow time variable. The liquid crystal state is a distinct phase

*Received by the editors January 21, 2005; accepted for publication (in revised form) May 9, 2005; published electronically December 6, 2005.
http://www.siam.org/journals/sima/37-3/62303.html

†Department of Mathematics, Pennsylvania State University, University Park, PA 16802 (bressan@math.psu.edu). This author's research was supported by the Italian M.I.U.R., within the research project 2002017219 "Equazioni iperboliche e paraboliche non lineari."

‡Trinity College, Department of Mathematics, Dublin 2, Ireland (adrian@maths.tcd.ie). This author's research was supported by the Science Foundation Ireland within the research project 04/BR/M0042 "Nonlinear Waves."

of matter observed between the solid and liquid states. More specifically, liquids are isotropic (that is, with no directional order) and without a positional order of their molecules, whereas the molecules in solids are constrained to point only in certain directions and to be only in certain positions with respect to each other. The liquid crystal phase exists between the solid and the liquid phases—the molecules in a liquid crystal do not exhibit any positional order, but they do possess a certain degree of orientational order. Not all substances can have a liquid crystal phase, e.g., water molecules melt directly from solid crystalline ice to liquid water. Liquid crystals are fluids made up of long rigid molecules, with an average orientation that specifies the local direction of the medium. Their orientation is described macroscopically by a field of unit vectors $\mathbf{n}(t,\mathbf{x})$). There are many types of liquid crystals, depending upon the amount of order in the material. Nematic liquid crystals are invariant under the transformation $\mathbf{n} \mapsto -\mathbf{n}$, in which case $\mathbf{n}$ is called a director field, so that the rodlike molecules have no positional order but tend to point in the same direction (along the director). The director field does not remain the same but generally fluctuates. Obtaining the equation governing the director field represents the crucial point for the modeling of nematic liquid crystals since it is advantageous to study the dynamics of the director field instead of studying the dynamics of all the molecules. The fluctuations of the director field are mainly due to the thermodynamical force caused by elastic deformations in the form of twisting, bending, and splaying (the last being a fan-shaped spreading out from the original direction, bending being a change of direction, while twisting corresponds to a rotation of the direction in planes orthogonal to the axis of rotation). Consider director fields that lie on a circle and depend on a single spatial variable x so that twisting is not allowed. To describe the dynamics of the director field independently of the coupling with the fluid flow, let $u(t,x)$ be the perturbation about a constant value. The asymptotic equation for weakly nonlinear unidirectional waves is precisely (1.2), obtained as the Euler–Lagrange equation of the variational principle

$$\delta \int_{t_1}^{t_2} \int_{\mathbb{R}} (u_t u_x + u u_x^2)\, dx dt = 0$$

for the internal stored energy of deformation of the director field if dissipative effects are neglected (corresponding to the case when inertia effects dominate viscosity)—see [HS] for the details of the derivation. Unlike other studies, in the Hunter–Saxton model the kinetic energy of the director field is not neglected. In the asymptotic regime in which (1.2) is derived (see [HS]), the nondimensionalized kinetic energy density is u_x^2 so that the condition

$$\int_{\mathbb{R}} u_x^2(t,x)\, dx < \infty \tag{1.4}$$

has to hold at any fixed time t for a physically meaningful solution to the Hunter–Saxton equation.

Equation (1.1) is also relevant in other physical situations; e.g., it is a high-frequency limit of the Camassa–Holm equation [DP], a nonlinear shallow water equation [CH, J] modeling solitons [CH, CS] as well as breaking waves [CE].

Geometric interpretation. An interesting aspect of the Hunter–Saxton equation (see [KM]) is the fact that, for spatially periodic functions, it describes geodesic flow on the homogeneous space $\mathrm{Diff}(\mathcal{S})/\mathrm{Rot}(\mathcal{S})$ of the infinite-dimensional Lie group

$\mathrm{Diff}(\mathcal{S})$ of smooth orientation-preserving diffeomorphisms of the unit circle $\mathcal{S}$ modulo the rotations $\mathrm{Rot}(\mathcal{S})$, with respect to the right-invariant homogeneous metric $\langle f,\, g\rangle = \int_{\mathcal{S}} f_x g_x\, dx$. The geometric interpretation of the Hunter–Saxton equation establishes a natural connection with the Camassa–Holm equation, which describes geodesic flow on $\mathrm{Diff}(\mathcal{S})$ with respect to the right-invariant metric $\langle f,\, g\rangle = \int_{\mathcal{S}} (fg + f_x g_x)\, dx$; see [K, CK]. A similar geometric interpretation of (1.1) on the diffeomorphism group of the line holds also for smooth initial data $\bar{u}$ in certain weighted function spaces, but the involved technicalities are more intricate (see [C] for the case of the Camassa–Holm equation).

Integrable structure. The Hunter–Saxton equation has an integrable structure. The equation has a reduction (see [BSS, HZ1]) to a finite-dimensional completely integrable Hamiltonian system whose phase space consists of piecewise linear solutions of the form

$$u(t,x) = \sum_{i=1}^{n} \alpha_i(t)\, |x - x_i(t)|, \tag{1.5}$$

with the constraint

$$\sum_{i=1}^{n} \alpha_i(t) = 0, \tag{1.6}$$

the Hamiltonian being

$$H(x,\alpha) = \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j\, |x_i - x_j|.$$

Due to their lack of regularity, functions of the form (1.5) are not classical solutions of (1.2). Below we will discuss in what sense they are weak solutions of the Hunter–Saxton equation. Let us point out that the constraint (1.6) is the necessary and sufficient condition to ensure that the distributional derivative $x \mapsto u_x(t,x)$ of a function of the form (1.5) belongs to the space $L^2(\mathbb{R})$. Thus (1.4) holds.

In the family of smooth functions $u : \mathbb{R} \mapsto \mathbb{R}$ all of whose derivatives $\partial_x^n u$ decay rapidly as $x \to \pm\infty$, the Hunter–Saxton equation is bi-Hamiltonian [HZ1]. If D^{-1} is the skew-adjoint antiderivative operator given by

$$(D^{-1}f)(x) = \frac{1}{2}\left(\int_{-\infty}^{x} - \int_{x}^{\infty}\right) f(x)\, dx, \qquad f \in \mathcal{D}(\mathbb{R}),$$

the first Hamiltonian form for the Hunter–Saxton equation is

$$u_t = J_1 \frac{\delta \mathcal{H}_1}{\delta u}, \qquad J_1 = u_x D^{-2} - D^{-2} u_x, \qquad \mathcal{H}_1 = \frac{1}{2}\int_{\mathbb{R}} u_x^2\, dx,$$

whereas the second, compatible Hamiltonian structure is

$$u_t = J_2 \frac{\delta \mathcal{H}_2}{\delta u}, \qquad J_2 = D^{-1}, \qquad \mathcal{H}_2 = \frac{1}{2}\int_{\mathbb{R}} u u_x^2\, dx.$$

Moreover, the Hunter–Saxton equation is formally integrable; e.g., it has an associated Lax pair (see [BSS]). However, the complete integrability of the equation has been established only in the previously mentioned case when it reduces to a finite-dimensional dynamical system.

The notion of solution. Physically relevant solutions of the Hunter–Saxton equation need to be of finite kinetic energy so that (1.4) must hold. This leads naturally to functions $u(t,x)$ with distributional derivative $u_x(t,\cdot)$ square integrable at every instant t. Note that the integrability assumption $u_x(t,\cdot) \in L^2(\mathbb{R})$ already imposes a certain degree of regularity on the function u. This suggests that it might be possible to incorporate a reasonably high degree of regularity in the concept of weak solutions to the Hunter–Saxton equation. Let us first consider the concept of weak solutions introduced by Hunter and Zheng [HZ2].

DEFINITION 1.1. *A function $u(t,x)$ defined on $[0,T] \times \mathbb{R}$ is a solution of* (1.2) *if the following hold:*

(i) $u \in C([0,T] \times \mathbb{R}; \mathbb{R})$ *and* $u(0,x) = \bar{u}(x)$ *pointwise on* $\mathbb{R}$.

(ii) *For each $t \in [0,T]$, the map $x \mapsto u(t,x)$ is absolutely continuous with $u_x(t,\cdot) \in L^2(\mathbb{R})$. Moreover, the map $t \mapsto u_x(t,\cdot)$ belongs to the space $L^\infty([0,T]; L^2(\mathbb{R}))$ and is locally Lipschitz continuous on $[0,T]$ with values in $H^{-1}_{loc}(\mathbb{R})$.*

(iii) *Equation* (1.2) *holds in the sense of distributions.*

Here and below, by a mapping f that is locally Lipschitz or locally bounded on $[0,T]$ with values in $H^{-1}_{loc}(\mathbb{R})$ we understand the following: for every $n \geq 1$ there is a constant $K_n \geq 0$ such that

$$\sup_{\{\psi \in \mathcal{D}(-n,n):\ \|\psi\|_{H^1(\mathbb{R})} \leq 1\}} \big|\langle f(t) - f(s), \psi\rangle\big| \leq K_n\, |t-s|, \qquad t,s \in [0,T],$$

respectively,

$$\sup_{\{\psi \in \mathcal{D}(-n,n):\ \|\psi\|_{H^1(\mathbb{R})} \leq 1\}} \big|\langle f(t), \psi\rangle\big| \leq K_n, \qquad t \in [0,T].$$

Here $\mathcal{D}(a,b)$ is the family of smooth functions $f : \mathbb{R} \to \mathbb{R}$ with compact support within $(a,b) \subset \mathbb{R}$.

To a function $u : [0,T] \times \mathbb{R} \to \mathbb{R}$ with the above properties associate the function $F : [0,T] \times \mathbb{R} \to \mathbb{R}$ defined by

$$F(t,x) = \frac{1}{4}\left(\int_{-\infty}^{x} - \int_{x}^{\infty}\right) u_x^2\, dx. \tag{1.7}$$

Then $F \in L^\infty_{loc}([0,T] \times \mathbb{R}; \mathbb{R}) \subset L^2_{loc}([0,T] \times \mathbb{R}; \mathbb{R})$. Moreover, $F_x = \frac{1}{2}\, u_x^2$ so that (1.2) becomes

$$(u_t + uu_x - F)_x = 0 \tag{1.8}$$

in the sense of distributions. Note that $uu_x \in L^2_{loc}([0,T] \times \mathbb{R}; \mathbb{R})$. From (1.8) we infer the existence of a distribution $h(t)$ so that $u_t + uu_x - F = h(t) \otimes 1(x)$, where $1(x)$ stands for the constant function with value 1 on $\mathbb{R}$. If $H(t)$ is a primitive of the distribution $h(t)$, we deduce that the distribution $U = u - H(t) \otimes 1(x)$ satisfies $U_t = u_t - h(t) \otimes 1(x) = F - uu_x \in L^2_{loc}([0,T] \times \mathbb{R}; \mathbb{R})$ and $U_x = u_x \in L^2_{loc}([0,T] \times \mathbb{R}; \mathbb{R})$. Therefore $U \in H^1_{loc}([0,T] \times \mathbb{R})$. Moreover, since $U_t = F - uu_x \in L^\infty_{loc}([0,T]; H^{-1}_{loc}(\mathbb{R}))$ ensures that U is locally Lipschitz as a function from $[0,T]$ to $H^{-1}_{loc}(\mathbb{R})$ and so is u, we deduce that $h(t) \otimes 1(x) = u - U$ shares this property too. But then $h : [0,T] \to \mathbb{R}$ has to be Lipschitz continuous. We infer that $u = U + H(t) \otimes 1(x)$ belongs to the space $H^1_{loc}([0,T] \times \mathbb{R})$. Since requirement (iii) in Definition 1.1 ensures that the identity

$$\int_0^T \int_{\mathbb{R}} \left(\phi_{xt} u + \frac{1}{2}\, \phi_{xx} u^2 - \frac{1}{2}\, \phi u_x^2\right) dx dt = 0$$

holds for every smooth function $\phi : (0,T)\times\mathbb{R} \to \mathbb{R}$ with compact support in $(0,T)\times\mathbb{R}$, we see that the notion of weak solution in the sense of Definition 1.1 is stronger than the concept of weak solution introduced by Hunter and Saxton [HS]. Another useful conclusion that can be drawn from the previous considerations is that for a function u with regularity properties specified in (i)–(ii) of Definition 1.1, requirement (iii) from Definition 1.1 is equivalent to asking that the equation

$$u_t + uu_x = F + h(t) \otimes 1(x) \tag{1.9}$$

hold in the distribution sense for some Lipschitz continuous function $h : [0,T] \to \mathbb{R}$. Any such function h is admissible. Among all these possibilities the most natural one corresponds to the special choice $h \equiv 0$. This leads us to the form (1.1) of the Hunter–Saxton equation.

In the following, we say that a map $t \mapsto u(t,\cdot)$ from $[0,T]$ into $\mathbf{L}^p_{loc}(\mathbb{R})$ is *absolutely continuous* if, for every bounded interval $[a,b]$, the restriction of u to $[a,b]$ is absolutely continuous as a map with values in $\mathbf{L}^p\big([a,b]\big)$. We can thus adopt the following notion of a weak solution.

Definition 1.2. *A function $u(t,x)$ defined on $[0,T]\times\mathbb{R}$ is a solution of* (1.2) *if the following hold:*

(i) *$u \in C([0,T]\times\mathbb{R};\mathbb{R})$ and $u(0,x) = \bar{u}(x)$ pointwise on $\mathbb{R}$.*

(ii) *For each $t \in [0,T]$, the map $x \mapsto u(t,x)$ is absolutely continuous with $u_x(t,\cdot) \in L^2(\mathbb{R})$. Moreover, the map $t \mapsto u_x(t,\cdot)$ belongs to the space $L^\infty([0,T];L^2(\mathbb{R}))$.*

(iii) *The map $t \mapsto u(t,\cdot) \in L^2_{loc}(\mathbb{R})$ is absolutely continuous and satisfies* (1.1) *for almost every $t \in [0,T]$.*

The concept of solution introduced in Definition 1.2 is stronger than that corresponding to Definition 1.1. Indeed, for a function u satisfying all the requirements of Definition 1.2 we infer by (1.1) that $u_{tx} \in L^\infty_{loc}([0,T];H^{-1}_{loc}(\mathbb{R}))$ since $u_t = -uu_x + F$ and uu_x, $F \in L^\infty_{loc}([0,T];L^2_{loc}(\mathbb{R}))$. This yields that the map $t \mapsto u_x(t,\cdot)$ is locally Lipschitz continuous on $[0,T]$ with values in $H^{-1}_{loc}(\mathbb{R})$. We thus recover the apparently missing part from requirement (ii) in Definition 1.1.

We remark that even with this stronger definition, solutions are far from unique. For example, consider the initial data

$$\bar{u}(x) = 0. \tag{1.10}$$

There are now two ways to prolong the solution for times $t > 0$. On one hand, we can define

$$u(t,x) = 0, \qquad x \in \mathbb{R}, \quad t \geq 0. \tag{1.11}$$

On the other hand, the function

$$u(t,x) \doteq \begin{cases} -2t & \text{if} \quad x \leq -t^2 \\ \frac{2x}{t} & \text{if} \quad |x| < t^2 \\ 2t & \text{if} \quad x \geq t^2 \end{cases} \quad \text{for} \quad t \geq 0, \tag{1.12}$$

provides yet another solution. To distinguish between these two solutions, we need to consider the evolution equation satisfied by the "energy density" u_x^2, namely,

$$(u_x^2)_t + (uu_x^2)_x = 0. \tag{1.13}$$

For smooth solutions, the conservation law (1.13) is satisfied pointwise. Notice that the solution defined by (1.10), (1.12) satisfies the additional conservation law (1.13) in the distributional sense, i.e.,

$$\int\int_{\mathbb{R}_+\times\mathbb{R}} \left\{u_x^2\varphi_t + uu_x^2\,\varphi_x\right\} dxdt = 0 \tag{1.14}$$

for every test function $\varphi \in \mathcal{C}_c^1(\mathbb{R}_+\times\mathbb{R})$ whose compact support is contained in the half plane where $t > 0$. On the contrary, the solution defined by (1.10)–(1.11) dissipates energy. More precisely, for every $t_2 \geq t_1 \geq 0$ we have

(1.15)
$$\int_{\mathbb{R}} u_x^2(t_2,x)\,\varphi(t_2,x)\,dx - \int_{\mathbb{R}} u_x^2(t_2,x)\,\varphi(t_2,x)\,dx \leq \int_{t_1}^{t_2}\int_{\mathbb{R}} \left\{u_x^2\,\varphi_t + uu_x^2\varphi_x\right\} dxdt$$

for every test function $\varphi \in \mathcal{C}_c^1(\mathbb{R}_+\times\mathbb{R})$. In what follows, we say that a solution is *dissipative* if (1.15) holds for every $t_2 > t_1 > 0$, $\varphi \in \mathcal{C}_c^1(\mathbb{R}_+\times\mathbb{R})$. Notice that the solution (1.10), (1.12) does not satisfy (1.15) when $t_1 = 2$, $t_2 > 0$.

At this point in the discussion it is worthwhile to point out that the most important feature in the definition of weak solutions is the requirement (1.4). A continuous function $u : [0,T]\times\mathbb{R} \to \mathbb{R}$ with square integrable distributional derivative $u_x(t,\cdot)$ belonging to the space $L^\infty([0,T];L^2(\mathbb{R}))$ is not necessarily bounded, nor does it have a predetermined asymptotic behavior at infinity, as one can see from the example

$$u(t,x) = \begin{cases} |x|^{\frac{1}{5}}\ \sin(|x|^{\frac{1}{5}}) & \text{if} \quad t\geq 0,\ |x|\geq 1,\\ |x|^{\frac{2}{3}}\ \sin(|x|^{\frac{2}{3}}) & \text{if} \quad t\geq 0,\ |x|\leq 1. \end{cases}$$

Nevertheless, the possibility that some additional structural information about the behavior of such functions at infinity might be inferred from some invariance properties of the Hunter–Saxton equation should be ruled out. To do this, consider solutions of the type (1.5) with the constraint (1.6). This type of solution enters into the framework of Definition 1.2, and for any $N(t) > \max\{|x_1(t)|,\dots,|x_n(t)|\}$ we have

$$u_t(t,x) = F(x) \qquad \text{a.e. on} \qquad |x| \geq N(t),$$

so that for all $j \geq N(t)$,

$$u_t(t,j) - u_t(-j) = F(j) - F(-j) = \frac{1}{2}\int_{-j}^{j} u_x^2(t,x)\,dx = \frac{1}{2}\int_{\mathbb{R}} u_x^2(t,x)\,dx. \tag{1.16}$$

But the quantity $I = \frac{1}{2}\int_{\mathbb{R}} u_x^2(t,x)\,dx$ is an invariant (time-independent); cf. [HZ1, BSS]. Moreover, the special form of the solutions guarantees that at every fixed $t \geq 0$,

$$u_\infty(t) \doteq \lim_{x\to\infty} u(t,x) = -\sum_{i=1}^{n} \alpha_i(t)\,x_i(t) = -\lim_{x\to-\infty} u(t,x)$$

and $u_\infty(t) = u(t,j) = -u(t,-j)$ for all $j \geq N(t)$. Thus (1.16) yields

$$u_\infty(t) = u_\infty(0) + \frac{1}{2}\,It, \qquad t \geq 0. \tag{1.17}$$

Unless $I = 0$, in which case u is constant, we see from (1.17) that the asymptotic behavior of the solutions changes with time. On the basis of this set of examples we conclude that the asymptotic behavior of the solutions at infinity should not be prescribed a priori. However, the previous set of examples indicates that a possible restriction would be to require $u \in L^\infty([0,T] \times \mathbb{R})$ if $\bar{u} \in L^\infty(\mathbb{R})$. In this case the space of functions introduced in Definition 1.2 (that is, bounded functions with all the properties specified in Definition 1.2 except the condition that u satisfies (1.1) in $L^2[-n,n]$ for every $n \geq 1$) is a Banach space when endowed with the norm

$$\|u\|_T = \sup_{(t,x)\in[0,T]\times\mathbb{R}} \{|u(t,x)|\} + \operatorname*{ess-sup}_{t\in[0,T]} \int_{\mathbb{R}} u_x^2(t,x)\,dx. \tag{1.18}$$

It is also worth noticing that a function entering the framework of Definition 1.2 has further regularity properties that are not explicitly stated. For example, we have the Hölder continuity property

$$|u(t,x) - u(t,y)| \leq K(t)\sqrt{|x-y|}, \qquad t \geq 0,\ x, y \in \mathbb{R},$$

with $K(t) = \|u_x(t,\cdot)\|_{L^2(\mathbb{R})}$, since

$$|u(t,x) - u(t,y)|^2 = \left|\int_x^y u_x(t,\zeta)\,d\zeta\right|^2 \leq |x-y| \cdot \left|\int_x^y u_x^2(t,\zeta)\,d\zeta\right| \leq |x-y| \int_{\mathbb{R}} u_x^2(t,\zeta)\,d\zeta.$$

2. Global existence of dissipative solutions. For twice continuously differentiable solutions, the derivative $v \doteq u_x$ of the solution u to (1.1) satisfies the equations

$$v_t + uv_x = -\frac{v^2}{2}, \tag{2.1}$$

$$(v^2)_t + (uv^2)_x = 0. \tag{2.2}$$

Define the characteristic $t \mapsto \xi(t,y)$ as the solution to the ODE

$$\frac{\partial}{\partial t}\xi(t,y) = u\big(t,\, \xi(t,y)\big), \qquad \xi(0,y) = y. \tag{2.3}$$

From (1.2) it follows that the evolution of the gradient u_x along each characteristic is described by

$$\frac{d}{dt}\,u_x\big(t,\xi(t,y)\big) = -\frac{1}{2}u_x^2\big(t,\xi(t,y)\big). \tag{2.4}$$

Observe that the solution of the ODE

$$\dot{z} = -z^2/2, \qquad z(0) = z_0$$

is given by

$$z(t) = \frac{2z_0}{2 + tz_0}. \tag{2.5}$$

If $z_0 \geq 0$, this solution is defined for all $t \geq 0$, whereas if $z_0 < 0$, this solution approaches $-\infty$ at the blow-up time

$$T(z_0) = -2/z_0. \tag{2.6}$$

Note that if $\bar{u}(x) \not\equiv 0$, then there is some $x_0 \in \mathbb{R}$ with $\bar{u}(x_0) < 0$ so that the characteristic curve $t \mapsto \xi(t, \bar{u}(x_0))$ will blow up in finite time. Nevertheless, if $\liminf_{x\in\mathbb{R}}\{\bar{u}_x(x)\} > -\infty$, then $T_0 > 0$, where

$$T_0 = \inf_{\{x\in\mathbb{R}:\ \bar{u}_x(x)<0\}} \left\{ \frac{-2}{\bar{u}_x(x)} \right\} \geq 0, \tag{2.7}$$

and on the time interval $[0, T_0)$ the method of characteristics can be used to construct the unique solution of (1.1). Let us describe the construction in detail. From (2.3) we get

$$\frac{\partial}{\partial t}\, \xi_x = u_x(t,\xi) \cdot \xi_x = \frac{2\,\bar{u}_x}{2 + t\,\bar{u}_x} \cdot \xi_x \tag{2.8}$$

since

$$u_x(t, \xi(t,y)) = \frac{2\,\bar{u}_x(y)}{2 + t\,\bar{u}_x(y)} \tag{2.9}$$

in view of (2.4) and the solution formula (2.5). The unique solution of the linear ODE (2.7) with initial data $\xi_x(0,y) = 1$ is given by

$$\xi_x(t,y) = \left(1 + \frac{t}{2}\,\bar{u}_x(y)\right)^2. \tag{2.10}$$

Since $1 + \frac{t}{2}\,\bar{u}_x(y) > 0$ for $t \in [0, T_0)$, relation (2.10) shows that for each $t \in [0, T_0)$ the map $y \mapsto \xi(t,y)$ is an absolutely continuous increasing diffeomorphism of the line. Define the absolutely continuous function φ by

$$\varphi(y) = \frac{1}{4} \int_{\mathbb{R}} \operatorname{sign}(y - x)\,\bar{u}_x^2(x)\,dx \tag{2.11}$$

so that $\varphi_y(y) = \frac{1}{2}\,\bar{u}_x^2(y)$. Note that by (2.10),

$$\xi_{tx} = \bar{u}_x + t\,\frac{\bar{u}_x^2}{2}. \tag{2.12}$$

Since $\xi_t(0,y) = \bar{u}(y)$ as $\xi(0,y) = y$, integration of (2.12) with respect to the spatial variable x yields

$$\xi_t(t,y) = \bar{u}(y) + \frac{t}{4} \int_{\mathbb{R}} \operatorname{sign}(y - x)\,\bar{u}_x^2(x)\,dx \tag{2.13}$$

and thus

$$\xi(t,y) = y + \int_0^t \xi_t(s,y)\,ds = y + t\,\bar{u}(y) + \frac{t^2}{8} \int_{\mathbb{R}} \operatorname{sign}(y - x)\,\bar{u}_x^2(x)\,dx. \tag{2.14}$$

The value of the solution u along the characteristic curve $t \mapsto \xi(t,y)$ is

$$u(t, \xi(t,y)) = \bar{u}(y) + \frac{t}{4} \int_{\mathbb{R}} \operatorname{sign}(y - x)\,\bar{u}_x^2(x)\,dx. \tag{2.15}$$

This relation is obtained by combining (2.13) with (2.3). The increasing diffeomorphism of the line $y \mapsto \xi(t,y)$ given by (2.14) and formula (2.15) yield the unique solution of the Hunter–Saxton equation on the time interval $[0, T_0)$. The above approach works as long as $2 + t\,\bar{u}_x(x) > 0$ but breaks down at $T = T_0$ with T_0 given by (2.7). The reason for the breakdown is that

$$\liminf_{t\uparrow T_0} \left\{ \inf_{x\in\mathbb{R}} u_x(t,x) \right\} = -\infty \tag{2.16}$$

in view of (2.9) and the definition (2.7) of T_0. Note that at $t = T_0$ we might have $\xi_x(t,x) = 0$ for all $x \in (a,b) \subset \mathbb{R}$ so that the map $y \mapsto \xi(t,y)$ is no longer an increasing diffeomorphism of the line. Nevertheless, the previous considerations suggest the following approach in the general case when $\bar{u}_x \in L^2(\mathbb{R})$, covering situations when possibly $T_0 = 0$ as it is the case for, e.g., $\bar{u}(x) = x^{\frac{2}{3}}(1-x)^{\frac{2}{3}}\,\chi_{[0,1]}$. Here χ_A stands for the characteristic function of the set A, defined by $\chi_A(x) = 1$ if $x \in A$ and $\chi_A(x) = 0$ if $x \notin A$.

Let $\bar{u} \in C(\mathbb{R})$ be such that its distributional derivative $\bar{u}_x$ is square integrable. Define $\varphi : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ by

$$\varphi_y(t,y) = \frac{1}{2}\,\bar{u}_x^2(y)\,\chi_{[\bar{u}_x > -2/t]}(y) \tag{2.17}$$

so that

$$\varphi(t,y) = \frac{1}{4}\int_{[\bar{u}_x > -2/t]} \operatorname{sign}(y-x)\,\bar{u}_x^2(x)\,dx, \qquad t > 0, \tag{2.18}$$

with the understanding that

$$\varphi(0,y) = \frac{1}{4}\int_{\mathbb{R}} \operatorname{sign}(y-x)\,\bar{u}_x^2(x)\,dx.$$

In other words, if $u(t,\xi(t,x_0))$ blows up before $t_0 > 0$, then the point x_0 is not included in the domain of the integral defining $\varphi(t_0,\cdot)$, because

$$T\big(\bar{u}_x(x)\big) > t \qquad \text{if and only if} \qquad \bar{u}_x(x) > -2/t,$$

according to (2.4) and (2.6). Observe that (2.18) and Young's inequality yield

$$\|\varphi(t,\cdot)\|_{L^\infty(\mathbb{R})} \le \frac{1}{4}\int_{\mathbb{R}} \bar{u}_x^2(x)\,dx, \qquad t \ge 0. \tag{2.19}$$

In the (t,x)-plane, the characteristic curve starting at y is obtained as

$$\xi(t,y) = y + t\bar{u}(y) + \int_0^t (t-s)\,\varphi(s,y)\,ds. \tag{2.20}$$

The value of the solution u along this curve is

$$u\big(t,\,\xi(t,y)\big) = \bar{u}(y) + \int_0^t \varphi(s,y)\,ds. \tag{2.21}$$

Observe that for all $t \ge 0$ and $y \in \mathbb{R}$,

$$\xi_t(t,y) = u(t,\xi(t,y)) \tag{2.22}$$

in view of (2.20)–(2.21).

THEOREM 2.1. *Given any absolutely continuous function* $\bar{u} : \mathbb{R} \to \mathbb{R}$ *with derivative* $\bar{u}_x \in L^2(\mathbb{R})$, *the formulas* (2.18)–(2.20) *provide a dissipative solution to* (1.1), *defined for all times* $t \geq 0$.

Proof. We proceed in several steps. First of all, for any fixed $t \geq 0$, the map $y \mapsto \xi(t,y)$ is absolutely continuous since $\varphi_y(t,\cdot) \in L^2(\mathbb{R})$. We claim that for any fixed $t \geq 0$ the map $y \mapsto \xi(t,y)$ is nondecreasing on $\mathbb{R}$ with $\lim_{y\to\pm\infty} \xi(t,y) = \pm\infty$.

Indeed, if $\bar{u}_x(y) > -\frac{2}{t}$, then $\bar{u}_x(y) > -\frac{2}{s}$ for all $s \in [0,t]$ so that $\varphi_y(s,y) = \frac{1}{2}\,\bar{u}_x^2(y)$ for $s \in [0,t]$ by (2.17). Since

$$\xi_y(t,y) = 1 + t\bar{u}_x(y) + \int_0^t (t-s)\,\varphi_y(s,y)\,ds \tag{2.23}$$

we find that in this case

$$\xi_y(t,y) = 1 + t\bar{u}_x(y) + \frac{t^2}{4}\,\bar{u}_x^2(y) = \frac{1}{4}\Big(2 + t\,\bar{u}_x(y)\Big)^2. \tag{2.24}$$

In the remaining cases we have that $\bar{u}_x(y) = -\frac{2}{t_0} \leq -\frac{2}{t}$ for some $t_0 \in (0,t]$. Therefore (2.17) yields $\varphi_y(s,y) = \frac{1}{2}\,\bar{u}_x^2(y)$ for $s \in [0,t_0)$, while $\varphi_y(s,y) = 0$ for $s \in (t_0,t]$. From (2.23) we infer that

$$\xi_y(t,y) = 1 - \frac{2t}{t_0} + \frac{2t - t_0}{t_0} = 0. \tag{2.25}$$

The relations (2.24)–(2.25) confirm the monotonicity of the map $y \mapsto \xi(t,y)$. Since $\xi(0,x) = x$, it remains to prove that $\lim_{y\to\pm\infty} \xi(t,y) = \pm\infty$ for any $t > 0$. Fix $t > 0$. Since $\bar{u}_x \in L^2(\mathbb{R})$, the Lebesgue measure $l(t)$ of the set $\{y \in \mathbb{R} :\ \bar{u}_x(y) \leq -\frac{1}{t}\}$ is finite. On the complement $C(t)$ of this set we obviously have $\bar{u}_x(y) > -\frac{1}{t}$ and thus $\xi_y(t,y) \geq \frac{1}{4}$ by taking into account (2.24). Therefore, given $x_2 > x_1$, we infer that

$$\begin{aligned}\xi(t,x_2) - \xi(t,x_1) &= \int_{x_1}^{x_2} \xi_y(t,y)\,dy \geq \int_{[x_1,x_2]\,\cap\, C(t)} \xi_y(t,y)\,dy \\ &\geq \int_{[x_1,x_2]\,\cap\, C(t)} \frac{1}{4}\,dy \geq \frac{x_2 - x_1 - l(t)}{4}.\end{aligned}$$

This proves the claim about the limiting behavior of $\xi(t,\cdot)$ at $\pm\infty$. While for times t up to the blow-up time T_0, given by (2.7), the map $y \mapsto \xi(t,y)$ is an absolutely continuous diffeomorphism of the real line, for $t \geq T_0$ this map is nondecreasing and onto but is not necessarily a bijection. Nevertheless, we would like to define the solution u by the formula (2.21) for all $t \geq 0$.

To show that u is well-defined via (2.21), due to the monotone and surjective character of the map $y \mapsto \xi(t,y)$, it is sufficient to show that if $\xi(t,y_1) = \xi(t,y_2)$ for some $y_2 > y_1$, then the values of u given by (2.21) are also equal. Indeed, we must have that $\xi(t,y) = \xi(t,y_1)$ for all $y \in [y_1,y_2]$, and a glance at (2.24)–(2.25) confirms that $\bar{u}_x(y) \leq -\frac{2}{t}$ for $y \in [y_1,y_2]$. This means that for every fixed $y \in (y_1,y_2)$ we have $\bar{u}_x(y) = -\frac{2}{t_0(y)}$ for some $t_0(y) \in [0,t]$. Consequently $\varphi_y(s,y) = \frac{1}{2}\,\bar{u}_x^2(y)\,\chi_{[0,t_0(y)]}(s)$ for $s \in [0,t]$, and differentiation of the right-hand side of (2.21) yields

$$\partial_y\left(\bar{u}(y) + \int_0^t \varphi(s,y)\,ds\right) = \bar{u}_x(y) + \int_0^{t_0(y)} \frac{1}{2}\,\bar{u}_x^2(y)\,ds = -\frac{2}{t_0(y)} + \frac{t_0(y)}{2}\,\frac{4}{t_0^2(y)} = 0$$

for $y \in (y_1, y_2)$. In particular, the values of the right-hand side of (2.21) are equal when evaluated at (t, y_1) and at (t, y_2). This proves that u is well-defined.

The next step is to prove that for every $t \geq 0$, the map $y \mapsto u(t,y)$ is continuous on $\mathbb{R}$ with distributional derivative in $L^2(\mathbb{R})$. Given $t \geq 0$ and $y_0 \in \mathbb{R}$, let $I_0 = \{x \in \mathbb{R} : \ \xi(t,x) = y_0\}$. The previously established properties of the map $x \mapsto \xi(t,x)$ ensure that $I_0 = [a,b]$ for some $a \leq b$. For any sequence $y_n \to y_0$, choose $x_n \in \mathbb{R}$ with $\xi(t,x_n) = y_n$. If we show that $\min\{|x_n - a|, |x_n - b|\} \to 0$ as $n \to \infty$, by the continuous dependence on the y-variable of the right-hand side of (2.21), we infer that

$$u(t,y_n) = u(t,\xi(t,n)) \to u(t,\xi(t,a)) = u(t,\xi(t,b)) = u(t,y_0)$$

since $\xi(t,x_n) \to \xi(t,a) = \xi(t,b) = y_0$. Thus $y \mapsto u(t,y)$ would be continuous at y_0. If it would be possible that $\min\{|x_{n_k} - a|, |x_{n_k} - b|\} \geq \varepsilon > 0$ for a sequence $n_k \to \infty$, then

$$|y_{n_k} - y_0| = |\xi(t,x_{n_k}) - \xi(t,a)| = |\xi(t,x_{n_k}) - \xi(t,b)| \geq \min\{y_0 - \xi(t,a-\varepsilon),\, \xi(t,b+\varepsilon)\} > 0$$

must hold by the definition of $[a,b]$ and the monotonicity property of the function $x \mapsto \xi(t,x)$. But this is a contradiction since $y_n \to y_0$ as $n \to \infty$. We therefore proved the continuity of the map $y \mapsto u(t,y)$ for every fixed $y \in \mathbb{R}$. Actually, a glance at the previous considerations confirms the continuity of the map $u : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ since $\xi : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ is continuous in view of (2.14). To show that for each $t \geq 0$ the distributional derivative $u_x(t,\cdot)$ belongs to $L^2(\mathbb{R})$, due to the absolute continuity of the nondecreasing surjective map $\xi(t,\cdot) : \mathbb{R} \to \mathbb{R}$, we first show that at every point $y = \xi(t,x)$ where $\xi_x(t,x) > 0$ exists, $u_x(t,y) \in \mathbb{R}$ exists. Indeed, at such a point y the right-hand side of (2.21), formally equal to $u_x(t,\xi(t,x)) \cdot \xi_x(t,x)$, is differentiable with derivative

$$\bar{u}_x(x) + \int_0^t \varphi_y(t,x)\, ds = \bar{u}_x(x) + \frac{t}{2}\, \bar{u}_x^2(x),$$

since in view of (2.24)–(2.25) we must have $\bar{u}_x(x) > -\frac{2}{t}$ and $\varphi_y(s,x) = \frac{1}{2}\, \bar{u}_x^2(x)$ for all $s \in [0,t]$. Since $\xi_y(t,x) = 1 + t\, \bar{u}_x(x) + \frac{t^2}{4}\, \bar{u}_x^2(x)$, we infer that $u_x(t,y)$ exists, being given by the formula

$$u_x(t,y) = \frac{\bar{u}_x(x) + \frac{t}{2}\, \bar{u}_x^2(x)}{1 + t\bar{u}_x(x) + \frac{t^2}{4}\, \bar{u}_x^2(x)} = \frac{\bar{u}_x(x)}{1 + \frac{t}{2}\, \bar{u}_x(x)}, \tag{2.26}$$

where $y = \xi(t,x)$. From (2.26) we deduce that for any interval $[x_1, x_2]$ where $\xi_x(t,x) > 0$ a.e., we have

$$\begin{aligned} \int_{y_1}^{y_2} u_x^2(t,y)\, dy &= \int_{\xi(t,x_1)}^{\xi(t,x_2)} u_x^2(t,\xi(t,x)) \,\cdot\, \xi_x(t,x)\, dx \\ &= \int_{x_1}^{x_2} \frac{\bar{u}_x^2(x)}{\big(1 + \frac{t}{2}\, \bar{u}_x(x)\big)^2} \left(1 + t\, \bar{u}_x(x) + \frac{t^2}{4}\, \bar{u}_x^2(x)\right) dx = \int_{x_1}^{x_2} \bar{u}_x^2(x)\, dx \end{aligned}$$

if $y_1 = \xi(t,x_1)$, $y_2 = \xi(t,x_2)$ and if we take into account (2.24). Summing up over such intervals, we obtain that

$$\int_{\mathbb{R}} u_x^2(t,x)\, dx = \int_{\{\bar{u}_x(x) > -\frac{2}{t}\}} \bar{u}_x^2(x)\, dx. \tag{2.27}$$

In particular, the map $t \mapsto \|u_x(t,\cdot)\|_{L^2(\mathbb{R})}$ is nonincreasing on $\mathbb{R}_+$. Moreover, we also have that

$$\int_{-\infty}^{\xi(t,y)} u_x^2(t,x)\,dx = \int_{\{x\in(-\infty,y]:\ \bar u_x(x) > -\frac{2}{t}\}} \bar u_x^2(x)\,dx$$

and

$$\int_{\xi(t,y)}^{\infty} u_x^2(t,x)\,dx = \int_{\{x\in[y,\infty):\ \bar u_x(x) > -\frac{2}{t}\}} \bar u_x^2(x)\,dx.$$

A comparison with (2.18) yields

$$\varphi(t,y) = \frac{1}{4}\int_{\mathbb{R}} \operatorname{sign}\left(\xi(t,y)-x\right) u_x^2(t,x)\,dx. \tag{2.28}$$

Furthermore, if $\xi_x(t,x) > 0$ exists, relation (2.20) ensures for $y = \xi(t,x)$ the existence of $u_t(t,y)$, given by the formula

$$u_t(t,y) = -\,u_x(t,y)\,u(t,y) + \varphi(t,x)$$

obtained by differentiation and taking into account (2.22). In combination with (2.28), this yields

$$(u_t + uu_x)(t,\xi(t,x)) = \frac{1}{4}\int_{\mathbb{R}} \operatorname{sign}(\xi(t,x)-\zeta)\,u_x^2(t,\zeta)\,d\zeta,$$

which is precisely (1.1) evaluated at $(t,\xi(t,x))$. In view of the previously established properties of the map $x \mapsto \xi(t,x)$ we deduce that the constructed function u satisfies also condition (iii) of Definition 2.1. Since the other properties required by Definition 2.1 were proved above, we conclude that u qualifies as a solution to (1.1) in the sense of Definition 2.1. This completes the proof of Theorem 2.1. $\square$

3. A distance functional. If $\bar u : \mathbb{R} \to \mathbb{R}$ is also bounded, in addition to being continuous and with distributional derivative $\bar u_x \in L^2(\mathbb{R})$, then the global solution $u(t,\cdot)$ constructed in Theorem 2.1 will be bounded at every fixed time $t \geq 0$. More precisely, in view of (2.19) and (2.21) we have that

$$\sup_{t\geq 0,\,x\in\mathbb{R}} \left|u(t,x)\right| \leq \sup_{x\in\mathbb{R}} \left|\bar u(x)\right| + \frac{t}{4}\int_{\mathbb{R}} \bar u_x^2(x)\,dx.$$

Thus, if $\bar u : \mathbb{R} \to \mathbb{R}$ is a bounded continuous function with distributional derivative $\bar u_x \in L^2(\mathbb{R})$, then at each fixed time $t \geq 0$, the solution $u(t,\cdot)$ to (1.1), constructed in Theorem 2.1, belongs to the Banach space $\mathcal{X}$ of bounded continuous functions $f : \mathbb{R} \to \mathbb{R}$ with distributional derivative $f_x \in L^2(\mathbb{R})$, endowed with the norm

$$\|f\|_{\mathcal{X}} = \sup_{x\in\mathbb{R}}\{|f(x)|\} + \left(\int_{\mathbb{R}} f_x^2(x)\,dx\right)^{\frac{1}{2}}.$$

The Banach space $\mathcal{X}$ seems suitable for (1.1)—see also [BZZ], where a construction similar to the one performed in Theorem 2.1 is presented. However, the map $t \mapsto u(t,\cdot)$ is generally not continuous from $\mathbb{R}_+$ to $\mathcal{X}$. Indeed, if for some $\tau > 0$ we have that the set $\{x \in \mathbb{R} :\ \bar u_x(x) = -\frac{2}{\tau}\}$ is of positive Lebesgue measure, then a discontinuity occurs at time $t=\tau$ for the map $t\mapsto u(t,\cdot)\in\mathcal{X}$ since from (2.27) we infer that for $t<\tau$,

$$\int_{\mathbb{R}} u_x^2(t,x)\,dx - \int_{\mathbb{R}} u_x^2(\tau,x)\,dx \geq \int_{\{x\in x:\ \bar u_x(x)=-\frac{2}{\tau}\}} \bar u_x^2(x)\,dx > 0.$$

Our aim will be to construct a distance functional in the space of solutions to (1.1) with respect to which we will have both continuity with respect to time as well as continuity with respect to the initial data for the solutions to (1.1). More precisely, for nonsmooth solutions the conservation law (2.2) is replaced by

$$(v^2)_t + (uv^2)_x = -\mu, \tag{3.1}$$

where μ is the positive measure on the (t,x)-plane defined as

$$\mu(\Omega) = \int_{\{(T(y),\ \xi(T(y),y))\,\in\,\Omega\}} \bar u_x^2(y)\,dy$$

for every open set $\Omega\subset\mathbb{R}_+\times\mathbb{R}$. Here $T(y)$ is the blow-up time along the characteristic curve starting at y, namely,

$$T(y) \doteq \begin{cases} -2/\bar u_x(y) & \text{if} \quad \bar u_x(y)<0,\\ \infty & \text{otherwise.}\end{cases}$$

For any $\bar u\in\mathcal{X}$, we can use the semigroup notation $S_t\bar u \doteq u(t,\cdot)$ to denote the solution of (1.1) constructed in section 2. Indeed

$$S_0\bar u = \bar u, \qquad S_{t+s}\bar u = S_t\big(S_s\bar u\big). \tag{3.2}$$

To prove (3.2), we first show that

$$\xi_1(t+s,y) = \xi_2(t,\xi_1(s,y)), \qquad t,\ s\geq 0,\ \ y\in\mathbb{R}, \tag{3.3}$$

where ξ_2 is the characteristic built upon the initial data $y\mapsto u(s,\xi_1(s,y))$. To check (3.3), we view both expressions as functions of t. At $t=0$ they are both equal to $\xi_1(s,y)$. For $t>0$, differentiation of (3.3) yields

$$\bar u(y) + \int_0^{t+s} \varphi_1(r,y)\,dr = u(s,\xi_1(s,y)) + \int_0^t \varphi_2(r,\xi_1(s,y))\,dr \tag{3.4}$$

in view of (2.20). We use (2.21) to express the right-hand side of (3.4) as

$$\bar u(y) + \int_0^s \varphi_1(r,y)\,dr + \int_0^t \varphi_2(r,\xi_1(s,y))\,dr.$$

Therefore, to get (3.4), which yields (3.3) by integration, it suffices to show that

$$\int_s^{t+s} \varphi_1(r,y)\,dr = \int_0^t \varphi_2(r,\xi_1(s,y))\,dr. \tag{3.5}$$

To prove (3.5), we note that by (2.18),

$$\begin{aligned}\int_0^t \varphi_2(r,\xi_1(s,y))\,dr &= \frac14\int_0^t\int_{\{x:\ u_x(s,x)>-\frac{2}{r}\}} \operatorname{sign}(\xi_1(s,y)-x)\,u_x^2(s,x)\,dx\,dr\\ &= \frac14\int_0^t\int_{\{x:\ u_x(s,\xi_1(s,x))>-\frac{2}{r}\}} \operatorname{sign}(\xi_1(s,y)-\xi_1(s,x))\,u_x^2(s,\xi_1(s,x))\,\partial_x\xi_1(s,x)\,dx\,dr\end{aligned}$$

if we change variables $x \mapsto \xi_1(s,x)$. Now taking (2.9)–(2.10) into account, we infer that

$$\begin{aligned}\int_0^t \varphi_2(r,\xi_1(s,y))\,dr &= \frac{1}{4}\int_0^t \int_{\{x:\; u_x(s,\xi_1(s,x))>-\frac{2}{r}\}} \operatorname{sign}(\xi_1(s,y)-\xi_1(s,x))\,\bar u_x^2(x)\,dx\,dr\\ &= \frac{1}{4}\int_0^t \int_{\{x:\; u_x(s,\xi_1(s,x))>-\frac{2}{r}\}} \operatorname{sign}(y-x)\,\bar u_x^2(x)\,dx\,dr\end{aligned}$$

since the function $x \mapsto \xi_1(s,x)$ is nondecreasing. But

$$u_x(s,\xi_1(s,x)) = \frac{2\,\bar u_x(x)}{2+s\,\bar u_x(x)} > -\frac{2}{r} \quad \text{if and only if} \quad \bar u_x(x) > -\frac{2}{s+r}$$

since the function $y \mapsto \frac{2y}{2+s\,y}$ is strictly increasing for $y > -\frac{2}{s}$, so that in the end we get

$$\begin{aligned}\int_0^t \varphi_2(r,\xi_1(s,y))\,dr &= \frac{1}{4}\int_0^t \int_{\{x:\; \bar u_x(x)>-\frac{2}{r+s}\}} \operatorname{sign}(y-x)\,\bar u_x^2(x)\,dx\,dr\\ &= \frac{1}{4}\int_s^{t+s} \int_{\{x:\; \bar u_x(x)>-\frac{2}{\tau}\}} \operatorname{sign}(y-x)\,\bar u_x^2(x)\,dx\,d\tau\end{aligned} \tag{3.6}$$

where $\tau = r+s$. On the other hand, by (2.18),

$$\int_s^{t+s} \varphi_1(r,y)\,dr = \frac{1}{4}\int_s^{t+s} \int_{\{x:\; \bar u_x(x)>-\frac{2}{\tau}\}} \operatorname{sign}(y-x)\,\bar u_x^2(x)\,dx\,d\tau$$

so that (3.4) holds and (3.3) is proved. Knowing (3.3), to infer $S_{t+s}\bar u = S_t(S_s\bar u)$, it suffices to show that

$$u(t+s,\xi_1(t+s,y)) = u(t,\xi_2(t,\xi_1(s,y))).$$

But, by (2.21), the left-hand side is precisely

$$\begin{aligned}\bar u(y) + \int_0^{t+s} \varphi(r,y)\,dr &= \bar u(y) + \int_0^s \varphi_1(r,y)\,dr\\ &\quad + \int_s^{t+s} \varphi_1(r,y)\,dr = u(s,\xi_1(s,y)) + \int_s^{t+s} \varphi_1(r,y)\,dr,\end{aligned}$$

which, taking into account (3.5), equals

$$u(s,\xi_1(s,y)) + \int_0^t \varphi_2(r,\xi_1(s,y))\,ds = u(t,\xi_2(t,\xi_1(s,y)))$$

in view of (2.21). This completes the proof of (3.2).

Notice that in general the map $t \mapsto S_t\bar u$ is *not* continuous from $[0,\infty[$ into $\mathcal{X}$. It is thus interesting to identify some distance $J(u,v)$ which is well adapted to the evolution generated by (1.1). More precisely, given an arbitrary constant M, in this section we shall construct a functional $J(u,v)$ with the following property: For any initial data $\bar u, \bar v \in \mathcal{X}$ with

$$\|\bar u_x\|_{\mathbf{L}^2} \le M, \qquad \|\bar v_x\|_{\mathbf{L}^2} \le M,$$

the corresponding dissipative solutions u, v constructed in Theorem 2.1 satisfy

$$J(u(t),\ v(t)) \leq e^{C_M\, t} J(\bar{u},\ \bar{v}).$$

To begin the construction, consider the metric space

$$X \doteq (\mathbb{R}^2 \times]-\pi/2,\ \pi/2]) \cup \{\infty\} \tag{3.7}$$

with distance

$$\begin{aligned} d((x,u,w),\ (\tilde{x},\tilde{u},\tilde{w})) \doteq \min\{&|x-\tilde{x}| + |u-\tilde{u}| + \kappa_0\,|w-\tilde{w}|,\\ &\kappa_0\,|\pi/2+w| + \kappa_0|\pi/2+\tilde{w}|\},\\ d((x,u,w),\ \infty) = \kappa_0\,&|\pi/2+w|. \end{aligned} \tag{3.8}$$

Here κ_0 is a suitably large constant, whose precise value will be specified later. Notice that X is obtained from the metric space $\mathbb{R}^2 \times [-\pi/2,\ \pi/2]$ by identifying all points $(x,\ u,\ -\pi/2)$ into a single point, called "∞".

Let $M(X)$ be the space of all bounded Radon measures on X. To each function $u \in H^1_{loc}(\mathbb{R})$ with $u_x \in L^2(\mathbb{R})$ we now associate the measure $\mu^u \in \mathcal{M}(X)$ defined as

$$\mu^u(\{\infty\}) = 0, \qquad \mu^u(A) = \int_{\{x\in\mathbb{R}:\ (x,\,u(x),\,\arctan\ u_x(x)\,)\in A\}} u_x^2(x)\,dx \tag{3.9}$$

for every Borel set $A \subseteq \mathbb{R}^2 \times]-\pi/2,\ \pi/2]$.

As distance between two functions $u, v \in \mathcal{X}$ we now introduce a kind of Kantorovich distance $J(u,v)$ related to an optimal transportation problem. Call $\mathcal{F}$ the family of all triples (ψ, ϕ_1, ϕ_2), where $\phi_1, \phi_2 : \mathbb{R} \mapsto [0,1]$ are simple Borel measurable maps (that is, their range is a finite number of points and the preimage of each such point is a Borel set) and $\psi : \mathbb{R} \mapsto \mathbb{R}$ is a nondecreasing absolute continuous surjective map. Assuming that

$$\phi_1(x)\,u_x^2(x) = \psi'(x)\cdot\phi_2(\psi(x))\,v_x^2(\psi(x)) \qquad \text{for almost every } x\in\mathbb{R}, \tag{3.10}$$

we define

$$\begin{aligned} J^{(\psi,\phi_1,\phi_2)}(u,v) \doteq &\int d((x,\ u(x),\ \arctan u_x(x)),\ (\psi(x),\ v(\psi(x)),\ \arctan v_x(\psi(x))))\\ &\qquad\cdot\phi_1(x)\,u_x^2(x)\,dx\\ &+\int d((x,\ u(x),\ \arctan u_x(x)),\ \infty)\cdot(1-\phi_1(x))\,u_x^2(x)\,dx\\ &+\int d((\psi(x),\ v(\psi(x)),\ \arctan v_x(\psi(x))),\ \infty)\\ &\qquad\cdot(1-\phi_2(\psi(x)))\,v_x^2(\psi(x))\,\psi'(x)\,dx. \end{aligned} \tag{3.11}$$

Observe that (ψ, ϕ_1, ϕ_2) can be regarded as a *transportation plan*, in order to transport the measure μ^u onto the measure μ^v. Since these two positive measures need not have the same total mass, we allow some of the mass to be transferred to the point ∞. More precisely, the mass transferred is $(1-\phi_1)\cdot\mu^u$ and $(1-\phi_2)\cdot\mu^v$. The last two integrals in (3.11) account for the additional cost of this transportation. Integrating (3.10) over the real line, one finds

$$\int_{\mathbb{R}} \phi_1(x)\,u_x^2(x)\,dx = \int_{\mathbb{R}} \phi_2(y)\,v_x^2(y)\,dy.$$

We can thus transport the measure $\phi_1\mu^u$ onto $\phi_2\,\mu^v$ by a map $\Psi : (x,\, u(x)\,\arctan u_x(x)) \mapsto (y,\, v(y),\, \arctan v_x(y))$, with $y = \psi(x)$. The associated cost is given by the first integral in (3.11). In this case the measure $\phi_2\,\mu^v$ is obtained as the push-forward of the measure $\phi_1\mu^u$. We recall that the *push-forward* of a measure μ by a mapping Ψ is defined as $(\Psi\#\mu)(A) \doteq \mu(\Psi^{-1}(A))$ for every measurable set A. Here $\Psi^{-1}(A) \doteq \{z :\ \Psi(z) \in A\}$.

We now define our distance functional by optimizing over all transportation plans, namely,

$$J(u,v) \doteq \inf_{(\psi,\phi_1,\phi_2)} \{J^{(\psi,\phi_1,\phi_2)}(u,v)\}, \tag{3.12}$$

where the infimum is taken over all triples $(\psi,\phi_1,\phi_2) \in \mathcal{F}$ such that (3.10) holds.

To check that (3.12) actually defines a distance, let $u,\ v,\ w \in \mathcal{X}$ be given functions.

1. Let us show that $J(u,v) = J(v,u)$. In order to do this, it is enough to prove that for every triple $(\psi,\ \phi_1,\ \phi_2) \in \mathcal{F}$ satisfying (3.10) and every $\varepsilon > 0$, there is a triple $(\eta,\varphi_1,\varphi_2) \in \mathcal{F}$ satisfying (3.10) such that $\eta : \mathbb{R} \to \mathbb{R}$ is a strictly increasing absolutely continuous bijection and

$$|J^{(\eta,\varphi_1,\varphi_2)}(u,v) - J^{(\psi,\phi_1,\phi_2)}(u,v)| \le \varepsilon. \tag{3.13}$$

Indeed, given $(\psi,\phi_1,\phi_2) \in \mathcal{F}$ satisfying (3.10), define $\tilde{\psi} = \eta^{-1}$, $\tilde{\phi_1} = \varphi_2$, $\tilde{\phi_2} = \varphi_1$. The properties of η ensure the absolute continuity of $\tilde{\psi}$ (see [N]) so that we obtain $J^{(\tilde{\psi},\tilde{\phi_1},\tilde{\phi_2})}(v,u) = J^{(\eta,\varphi_1,\varphi_2)}(u,v)$ by performing the change of variables $x \mapsto \eta(x)$. Since $\varepsilon > 0$ was arbitrary, we infer that $J(v,u) \le J(u,v)$. Interchanging the roles of u and v we get $J(u,v) = J(v,u)$.

To prove (3.13), it is convenient to view $\psi : \mathbb{R} \to \mathbb{R}$ as a maximal monotone multifunction $\psi : \mathbb{R} \mapsto \mathcal{P}(\mathbb{R})$ with domain and range $\mathbb{R}$. Here $\mathcal{P}(\mathbb{R})$ is the family of all subsets of $\mathbb{R}$. The conditions for a multifunction $F : \mathbb{R} \mapsto \mathcal{P}(\mathbb{R})$ to be maximal monotone with domain and range $\mathbb{R}$ may be explicitly written as follows [Z]:

- for every $x \in \mathbb{R}$, the set $F(x) \subset \mathbb{R}$ is nonempty (i.e., the domain of F is $\mathbb{R}$);
- for every $y \in \mathbb{R}$ there is at least some $x \in \mathbb{R}$ with $y \in F(x)$, expressing the fact that the range of F is $\mathbb{R}$;
- there are no couples (x_1,y_1) and (x_2,y_2) with $y_1 \in F(x_1)$ and $y_2 \in F(x_2)$ such that $x_1 < x_2$ and $y_2 < y_1$, meaning that F is monotone;
- if we associate to F its graph $\{(x,y) \in \mathbb{R}^2 :\ y \in F(x)\}$, then this graph has no proper extension satisfying the first three properties (condition defining the maximal monotonicity property).

We recall some important features presented by such maps [AA, Z]:

- the set $F(x)$ is an interval of the form $[a_x,b_x]$ with $a_x \le b_x$ for all $x \in \mathbb{R}$ and $a_x = b_x$ for all $x \in \mathbb{R}$, except perhaps an at most countable set (so F is single-valued with the exception of at most countably many points);
- F is a.e. differentiable, that is, for almost all $x_0 \in \mathbb{R}$ there exists $F'(x_0) \in \mathbb{R}$ such that

$$\lim_{x\to x_0,\, y\in F(x)} \frac{y - F(x_0) - (x-x_0)\,F'(x_0)}{x - x_0} = 0;$$

- we can define the inverse $F^{-1} : \mathbb{R} \to \mathcal{P}(\mathbb{R})$ of F by asking $y \in F^{-1}(x)$ if and only if $x \in F(x)$ and F^{-1} is again a maximal monotone multifunction with domain and range $\mathbb{R}$.

Since the multifunction ψ^{-1} is maximal monotone, let $\{y_n\}$ be the (at most countable) set of points where it is multivalued, that is, $\psi^{-1}(y) = [a_n, b_n]$ with $b_n > a_n$. Then $\psi(x) = y_n$ for $x \in [a_n, b_n]$ and, ψ being absolutely continuous, $\psi_x > 0$ a.e. on $\mathbb{R} - \bigcup_n [a_n, b_n]$ since ψ is strictly increasing on this set. Given $\gamma > 0$, the absolute continuity of $\psi : \mathbb{R} \to \mathbb{R}$ allows us to choose some $\delta > 0$ such that the total variation of ψ over the union of disjoint closed intervals with the sum of their lengths less than δ is less than γ; cf. [BGH]. On each interval $[a_n - \frac{\delta}{2^n}, b_n + \frac{\delta}{2^n}]$ we replace ψ with the linear function η which takes the values $\psi(a_n - \frac{\delta}{2^n})$, respectively, $\psi(b_n + \frac{\delta}{2^n})$, at the endpoints. By the way a_n and b_n were defined, we know that $\psi(b_n + \frac{\delta}{2^n}) > \psi(a_n - \frac{\delta}{2^n})$ so that $\eta'(x)$ is a positive constant on $[a_n - \frac{\delta}{2^n}, b_n + \frac{\delta}{2^n}]$ with

$$\sum_n \int_{a_n - \frac{\delta}{2^n}}^{b_n + \frac{\delta}{2^n}} \eta'(x)\, dx \le \sum_n \left(\psi\left(b_n + \frac{\delta}{2^n}\right) - \psi\left(a_n - \frac{\delta}{2^n}\right)\right) \le \gamma.$$

Setting $\eta(x) = \psi(x)$ for $x \notin [a_n - \frac{\delta}{2^n}, b_n + \frac{\delta}{2^n}]$, we obtain a strictly increasing absolutely continuous bijection $\eta : \mathbb{R} \to \mathbb{R}$. Let us now show that the triple $(\eta, \varphi_1, \varphi_2) \in \mathcal{F}$ satisfies both (3.10) and (3.13), where φ_1, φ_2 are defined by setting $\varphi_1(x) = 0$ for $x \in [a_n - \frac{\delta}{2^n}, b_n + \frac{\delta}{2^n}]$ and $\varphi_1(x) = \phi_1(x)$ for $x \notin [a_n - \frac{\delta}{2^n}, b_n + \frac{\delta}{2^n}]$, while $\varphi_2(\eta(x)) = \phi_2(\psi(x))$ for $x \notin [a_n - \frac{\delta}{2^n}, b_n + \frac{\delta}{2^n}]$ and $\varphi_2(\eta(x)) = 0$ for $x \in [a_n - \frac{\delta}{2^n}, b_n + \frac{\delta}{2^n}]$. On the complement of the set $\bigcup_n [a_n - \frac{\delta}{2^n}, b_n + \frac{\delta}{2^n}]$ relation (3.10) clearly holds a.e. for $(\eta, \varphi_1, \varphi_2)$, being unmodified from (3.10) for (ψ, ϕ_1, ϕ_2). If $x \in [a_n - \frac{\delta}{2^n}, b_n + \frac{\delta}{2^n}]$, then (3.10) for $(\eta, \varphi_1, \varphi_2)$ holds again since both sides are zero as $\varphi_1(x) = \varphi_2(\eta(x)) = 0$ in this case. Finally, to check (3.13), notice that if we denote

$$E_\delta = \bigcup_n \left\{ \left[a_n - \frac{\delta}{2^n}, a_n\right] \cup \left[b_n, b_n + \frac{\delta}{2^n}\right]\right\}, \qquad A = \bigcup_n [a_n, b_n],$$

then

(3.14)
$$|J^{(\eta,\varphi_1,\varphi_2)}(u,v) - J^{(\psi,\phi_1,\phi_2)}(u,v)| \le 2\kappa_0\pi \int_{E_\delta} u_x^2\, dx + 2\kappa_0\pi \int_{E_\delta \cup A} v_x^2(\eta(x))\, \eta'(x)\, dx.$$

Indeed, the distance d is less than $2\kappa_0\pi$ and the integrands in $J^{(\eta,\varphi_1,\varphi_2)}(u,v)$ and $J^{(\psi,\phi_1,\phi_2)}(u,v)$ agree on the complement of the set $\bigcup_n [a_n - \frac{\delta}{2^n}, b_n + \frac{\delta}{2^n}]$ by definition. Also, for almost every $x \in [a_n, b_n]$ we have $\phi_1(x)\, u_x^2 = 0$ by (3.10) as $\psi'(x) = 0$, and $\varphi_1(x) = 0$ by its definition. We obtain (3.14). Since the absolutely continuous map η maps $E_\delta \cup A$ into a set of Lebesgue measure less than γ, and u_x^2, $v_x^2 \in L^1(\mathbb{R})$, from (3.14) we infer (3.13) by choosing $\delta > 0$ and $\gamma > 0$ small enough. This completes the argumentation needed to show that $J(u,v) = J(v,u)$.

2. Choosing $\psi(x) = x$, $\phi_1(x) = \phi_2(x) = 1$, we immediately see that $J(u,u) = 0$. Moreover, we have $J(u,v) > 0$ if $u \ne v$. To check this, note that $J(u,v) = 0$ implies that there is a sequence $(\psi^n, \phi_1^n, \phi_2^n)$ along which $J^{(\psi^n,\phi_1^n,\phi_2^n)}(u,v) \to 0$. The second term in (3.11) yields

$$\left(\frac{\pi}{2} + \arctan u_x(x)\right) (1 - \phi_1^n(x))\, u_x^2(x) \to 0 \quad \text{in} \quad L^1(\mathbb{R}),$$

so that along a subsequence $(1 - \phi_1^{n_k})\, u_x^2 \to 0$ a.e. on $\mathbb{R}$ since $u_x > -\infty$ a.e. On the set $S = \{x \in \mathbb{R} :\ u_x(x) \ne 0\}$ we therefore have $\phi_1^{n_k} \to 1$ a.e. Moreover, the first term in (3.11) forces

$$\begin{aligned} &\phi_1^{n_k}(x)u_x^2(x)\cdot \min\Big\{|x-\psi^{n_k}(x)|+|u(x)-v(\psi^{n_k}(x))| \\ &\qquad\qquad +\kappa_0|\arctan u_x(x)-\arctan v_x(\psi^{n_k}(x))|, \\ &\kappa_0\left[\frac{\pi}{2}+\arctan u_x(x)+\frac{\pi}{2}+\arctan v_x(\psi^{n_k}(x))\right]\Big\}\to 0 \quad \text{in} \quad L^1(\mathbb{R}). \end{aligned} \tag{3.15}$$

Since $u_x>-\infty$ a.e. ensures

$$\frac{\pi}{2}+\arctan u_x(x)+\frac{\pi}{2}+\arctan v_x(\psi^{n_k}(x))\ge \frac{\pi}{2}+\arctan u_x(x)>0 \quad \text{a.e. on} \quad \mathbb{R},$$

we infer from (3.15), by passing to another subsequence, that

$$|x-\psi^{n_k}(x)|+|u(x)-v(\psi^{n_k}(x))|\to 0 \quad \text{a.e. on} \quad S.$$

In view of the continuity of v, $\psi^{n_k}(x)\to x$ a.e. on S guarantees $v(\psi^{n_k}(x))\to v(x)$ a.e. on S so that $u=v$ a.e. on S since also $v(\psi^{n_k}(x))\to u(x)$ a.e. on S. Repeating this argument with the roles of u and v reversed, we find that $u=v$ a.e. on the set $\{x\in\mathbb{R}:\ v_x\neq 0\}$. Combining this with the previous conclusion, we have $u=v$ a.e. on the complement of the set $\{x\in\mathbb{R}: u_x=v_x=0\}$. Since $u_x,\, v_x\in L^2(\mathbb{R})$, this is possible only if $u=v$ on $\mathbb{R}$. Thus $J(u,v)=0$ if and only if $u=v$.

3. Finally, to prove the triangle inequality, it suffices to show that for every choice of $(\psi^\flat,\phi_1^\flat,\phi_2^\flat)$ satisfying (3.10), and of $(\psi^\sharp,\phi_1^\sharp,\phi_2^\sharp)$ satisfying (3.10) for (v,w), the triplet (ψ,ϕ_1,ϕ_2) defined by

$$\psi(x)=\psi^\sharp(\psi^\flat(x)),\quad \phi_1(x)=\phi_1^\flat(x)\cdot\phi_1^\sharp(\psi^\flat(x)),\quad \phi_2(y)=\phi_2^\sharp(y)\cdot\phi_2^\flat(\psi^\flat(x))$$

satisfies (3.10) for (u,w) and

$$J^{(\psi,\phi_1,\phi_2)}(u,w)\le J^{(\psi^\flat,\phi_1^\flat,\phi_2^\flat)}(u,v)+J^{(\psi^\sharp,\phi_1^\sharp,\phi_2^\sharp)}(v,w). \tag{3.16}$$

Notice that composing the relation (3.10) for (v,w) a.e. to the right with $\psi^\flat$, and multiplying the outcome by $\phi_2^\flat\circ\psi^\flat\cdot(\psi^\flat)'$, we infer that (3.10) holds a.e. on $\mathbb{R}$ for (u,w) with our choice of (ψ,ϕ_1,ϕ_2), and we can now concentrate on proving (3.16).

To simplify matters, we introduce the following notation:

$$\begin{aligned} &P_1=(x,\,u,\,\arctan u_x),\quad P_2=(\psi^\flat,\,v\circ\psi^\flat,\,\arctan v_x\circ\psi^\flat),\\ &P_3=(\psi,\,w\circ\psi,\,\arctan w_x\circ\psi),\\ &\qquad m_1=u_x^2,\qquad m_2=v_x^2\circ\psi^\flat\cdot(\psi^\flat)',\qquad m_3=w_x^2\circ\psi\cdot\psi'. \end{aligned}$$

The relations of type (3.10) then yield that a.e. on $\mathbb{R}$,

$$\phi_1^\flat\cdot m_1=\phi_2^\flat\circ\psi^\flat\cdot m_2,\qquad \phi_1^\sharp\circ\psi^\flat\cdot m_2=\phi_2^\sharp\circ\psi\cdot m_3,\qquad \phi_1\cdot m_1=\phi_2\circ\psi\cdot m_3. \tag{3.17}$$

Also,

$$\begin{aligned} J^{(\psi,\phi_1,\phi_2)}(u,w) &= \int_{\mathbb{R}} \{d(P_1,P_3)\cdot\phi_1 m_1 + d(P_1,\infty)\cdot(1-\phi_1)m_1 + d(P_3,\infty) \\ &\qquad \cdot(1-\phi_2\circ\psi)m_3\}\,dx, \\ J^{(\psi^\flat,\phi_1^\flat,\phi_2^\flat)}(u,v) &= \int_{\mathbb{R}} \{d(P_1,P_2)\cdot\phi_1^\flat m_1 + d(P_1,\infty)\cdot(1-\phi_1^\flat)m_1 + d(P_2,\infty) \\ &\qquad \cdot(1-\phi_2^\flat\circ\psi^\flat)m_2\}\,dx, \\ J^{(\psi^\sharp,\phi_1^\sharp,\phi_2^\sharp)}(v,w) &= \int_{\mathbb{R}} \{d(P_2,P_3)\cdot\phi_1^\sharp\circ\psi^b m_1 + d(P_2,\infty)\cdot(1-\phi_1^\sharp\circ\psi^\flat)m_2 + d(P_3,\infty) \\ &\qquad \cdot(1-\phi_2^\sharp\circ\psi)m_3\}\,dx, \end{aligned}$$

the last relation being obtained after the change of variables $x\mapsto\psi^\flat(x)$ in the integral. We will prove (3.16) by deriving an appropriate inequality valid a.e. pointwise between the integrands in the previous expressions. Since

$$(1-\phi_2^\flat\circ\psi^\flat)(1-\phi_1^\sharp\circ\psi^\flat)\geq 0,$$

we have

$$1-\phi_2^\flat\circ\psi^\flat + 1-\phi_1^\sharp\circ\psi^\flat \geq \phi_2^\flat\circ\psi^\flat(1-\phi_1^\sharp\circ\psi^\flat)+\phi_1^\sharp\circ\psi^\flat(1-\phi_2^\flat\circ\psi^\flat).$$

Multiplication of both sides by $d(P_2,\infty)\cdot m_2$ leads to

$$\begin{aligned} &d(P_2,\infty)\cdot(1-\phi_2^\flat\circ\psi^\flat)m_2 + d(P_2,\infty)\cdot(1-\phi_1^\sharp\circ\psi^\flat)m_2 \\ &\quad\geq d(P_2,\infty)\cdot\phi_1^\flat(1-\phi_1^\sharp\circ\psi^\flat)m_1 + d(P_2,\infty)\cdot\phi_1^\sharp\circ\psi^\flat(1-\phi_2^\flat\circ\psi^\flat)m_2 \end{aligned}\tag{3.18}$$

in view of (3.17). Multiply now the inequalities

$$d(P_1,P_2)-d(P_1,\infty)+d(P_2,\infty)\geq 0,\qquad d(P_2,P_3)-d(P_3,\infty)+d(P_2,\infty)\geq 0$$

by $\phi_1^\flat(1-\phi_1^\sharp\circ\psi^\flat)m_1$, respectively, $\phi_1^\sharp\circ\psi^\flat(1-\phi_2^\flat\circ\psi^\flat)m_2$, and add them up. The outcome yields in combination with (3.18) that

$$\begin{aligned} &d(P_1,P_2)\cdot\phi_1^\flat(1-\phi_1^\sharp\circ\psi^\flat)m_1 - d(P_1,\infty)\cdot\phi_1^\flat(1-\phi_1^\sharp\circ\psi^\flat)m_1 \\ &\quad + d(P_2,P_3)\cdot\phi_1^\sharp\circ\psi^\flat(1-\phi_2^\flat\circ\psi^\flat)m_2 + d(P_2,\infty)\cdot(1-\phi_2^\flat\circ\psi^\flat)m_2 \\ &\quad + d(P_2,\infty)\cdot(1-\phi_1^\sharp\circ\psi^\flat)m_2 \geq d(P_3,\infty)\cdot\phi_1^\sharp\circ\psi^\flat(1-\phi_2^\flat\circ\psi^\flat)m_2. \end{aligned}$$

Adding to both sides the quantity

$$\begin{aligned} &d(P_1,\infty)\cdot m_1 + d(P_3,\infty)\cdot m_3 + d(P_1,P_2)\cdot\phi_1^\flat\cdot\phi_1^\sharp\circ\psi^\flat\cdot m_1 - d(P_1,\infty)\cdot\phi_1^\flat \\ &\quad\cdot\phi_1^\sharp\circ\psi^\flat\cdot m_1 + d(P_2,P_3)\cdot\phi_1^\sharp\circ\psi^\flat\cdot\phi_2^\flat\circ\psi^\flat\cdot m_2 - d(P_3,\infty)\cdot\phi_2^\sharp\circ\psi\cdot m_3 \end{aligned}$$

we deduce by (3.17) that the integrand of $J^{(\psi^\flat,\phi_1^\flat,\phi_2^\flat)}(u,v)+J^{(\psi^\sharp,\phi_1^\sharp,\phi_2^\sharp)}(v,w)$, equal a.e. precisely to the left-hand side of the new inequality, is a.e. pointwise larger than

$$\begin{aligned} &d(P_3,\infty)\cdot\phi_1^\sharp\circ\psi^\flat(1-\phi_2^\flat\circ\psi^\flat)m_2 + d(P_1,\infty)\cdot m_1 + d(P_3,\infty)\cdot m_3 \\ &\quad + d(P_1,P_2)\cdot\phi_1^\flat\cdot\phi_1^\sharp\circ\psi^\flat\cdot m_1 - d(P_1,\infty)\cdot\phi_1^\flat\cdot\phi_1^\sharp\circ\psi^\flat\cdot m_1 \\ &\quad + d(P_2,P_3)\cdot\phi_1^\sharp\circ\psi^\flat\cdot\phi_2^\flat\circ\psi^\flat\cdot m_2 - d(P_3,\infty)\cdot\phi_2^\sharp\circ\psi\cdot m_3. \end{aligned}$$

Taking into account (3.17) and the definition $\phi_1 = \phi_1^\flat \cdot \phi_1^\sharp \circ \psi^\flat$, we see that the above expression equals

$$\begin{aligned} &d(P_1,\infty)\cdot(1-\phi_1)m_1 + d(P_3,\infty)\cdot(1-\phi_2\circ\psi)m_3 + (d(P_1,P_2)+d(P_2,P_3)\cdot\phi_1 m_1)\\ &\quad \geq d(P_1,\infty)\cdot(1-\phi_1)m_1 + d(P_3,\infty)\cdot(1-\phi_2\circ\psi)m_3 + d(P_1,P_3)\cdot\phi_1 m_1. \end{aligned}$$

The lower estimate is a.e. precisely the integrand in $J^{(\psi,\phi_1,\phi_2)}(u,w)$, and (3.16) holds. The proof that J satisfies the triangle inequality is therefore complete.

In the remainder of this section we examine how the distance $J(\cdot,\cdot)$ behaves in connection with solutions of (1.1).

Continuity with respect to time. Let $t \mapsto u(t)$ be the solution of (1.1) constructed in section 2. For any fixed $t > 0$, we define a transportation plan of $\mu^{\bar u}$ to $\mu^{u(t)}$ by setting

$$\psi(x) \doteq \xi(t,x), \qquad \phi_1(x) \doteq \begin{cases} 1 & \text{if} \quad T(x) > t, \\ 0 & \text{if} \quad T(x) \leq t, \end{cases} \quad \phi_2(x) \equiv 1. \tag{3.19}$$

Relation (3.6) follows from (2.9)–(2.10) on $\{T(x) > t\}$ and from (2.25) on $\{T(x) \leq t\}$. The cost of this plan is estimated by

$$\begin{aligned} J^{(\psi,\phi_1,\phi_2)}(\bar u,\ u(t)) \leq &\int_{\{T(x)>t\}} \{|x-\xi(t,x)| + |\bar u(x) - u(t,\xi(t,x))| \\ &\qquad + \kappa_0|\arctan \bar u_x(x) - \arctan u_x(t,\xi(t,x))|\}\, \bar u_x^2(x)\,dx \\ &+ \int_{\{T(x)\leq t\}} |\pi/2 + \arctan \bar u_x(x)|\, \bar u_x^2(x)\,dx. \end{aligned} \tag{3.20}$$

By (2.4) we have that a.e.

$$\left|\frac{d}{dt}\arctan u_x\big(t,\xi(t,x)\big)\right| = \left|\frac{\frac{d}{dt}u_x\big(t,\xi(t,x)\big)}{1+u_x^2\big(t,\xi(t,x)\big)}\right| \leq \frac{1}{2}. \tag{3.21}$$

An integration on $[0,t]$ yields

$$|\arctan \bar u_x(x) - \arctan u_x(t,\xi(t,x))| \leq \frac{t}{2}, \qquad t \geq 0. \tag{3.22}$$

On the other hand, using (2.20), we get

$$|x-\xi(t,x)| \leq t\,|\bar u(x)| + \int_0^t (t-s)\,|\varphi(s,x)|\,ds \leq t\,|\bar u(x)| + \frac{t^2}{8}\int_{\mathbb{R}} \bar u_x^2(x)\,dx, \quad t\geq 0,\ x\in\mathbb{R}, \tag{3.23}$$

if we take into account (2.19). From (2.21) and (2.19), we also infer

$$|\bar u(x) - u(t,\xi(t,x))| \leq \int_0^t |\varphi(s,y)|\,dy \leq \frac{t}{4}\int_{\mathbb{R}} \bar u_x^2(x)\,dx, \qquad t\geq 0,\ x\in\mathbb{R}. \tag{3.24}$$

To estimate the last term in (3.20), notice that

$$\{x\in\mathbb{R}:\ T(x)\leq t\} = \left\{x\in\mathbb{R}:\ -\frac{2}{\bar u_x(x)} \leq t\right\} = \left\{x\in\mathbb{R}:\ \bar u_x(x) \leq -\frac{2}{t}\right\}, \qquad t>0. \tag{3.25}$$

Furthermore, since $\lim_{x\to-\infty} x(\frac{\pi}{2}+\arctan x) = -1$, there is a constant $c > 0$ such that

$$0 \le \frac{\pi}{2} + \arctan y \le \frac{c}{|y|}, \qquad y \le -1,$$

whereas

$$\left|\frac{\pi}{2} + \arctan y\right| y^2 \le \pi \qquad \text{if} \quad -1 \le y \le 0,$$

so that

$$\left|\frac{\pi}{2} + \arctan \bar u_x(x)\right| \bar u_x^2(x) \le \pi + c\,|\bar u_x(x)| \qquad \text{if} \quad \bar u_x(x) \le -\frac{2}{t}. \tag{3.26}$$

On the other hand, if $\bar u_x(x) \le -\frac{2}{t}$, then $t^2\bar u_x^2(x) \ge 4$ so that

$$\int_{\{T(x)\le t\}} 1\,dx \le \frac{t^2}{4}\int_{\{T(x)\le t\}} \bar u_x^2(x)\,dx. \tag{3.27}$$

From (3.25)–(3.27) we infer that

$$\int_{\{T(x)\le t\}} \left|\frac{\pi}{2} + \arctan \bar u_x(x)\right| \bar u_x^2(x)\,dx \le \frac{\pi t^2}{4}\|\bar u_x\|_{L^2}^2 + c\int_{\{T(x)\le t\}} |\bar u_x(x)|\,dx$$
$$\le \frac{\pi t^2}{4}\|\bar u_x\|_{L^2}^2 + c\left(\int_{\{T(x)\le t\}} 1\,dx\right)^{\frac12}\left(\int_{\{T(x)\le t\}} \bar u_x^2(x)\,dx\right)^{\frac12} \le \frac{\pi t^2}{4}\|\bar u_x\|_{L^2}^2 + \frac{ct}{2}\,\|\bar u_x\|_{L^2}^2.$$

By (3.20), (3.22)–(3.25), and the previous inequality we conclude

$$J^{(\psi,\phi_1,\phi_2)}(\bar u,\ u(t)) \le \left(\frac{\pi t}{4} + \frac{c+\kappa_0}{2} + \|\bar u\|_{L^\infty} + \frac{t+2}{8}\,\|\bar u_x\|_{L^2}^2\right) t\,\|\bar u_x\|_{L^2}^2, \qquad t \ge 0. \tag{3.28}$$

It is now clear that each semigroup trajectory $t \mapsto S_t\bar u$ is Lipschitz continuous as a map from $[0,\infty[$ into the metric space X equipped with our distance functional J. The Lipschitz constant remains uniformly bounded as $\bar u$ ranges over bounded subsets of X.

Continuity with respect to the initial data. We now consider two distinct solutions and study how the distance $J\big(u(t),\ \tilde u(t)\big)$ varies in time. Recall that the solution $u = u(t,x)$ is computed by (2.20)–(2.22), also in the case where the gradient blows up. The same formula of course holds for $\tilde u$. Let $(\psi_0, \phi_{1,0}, \phi_{2,0})$ be an optimal transportation plan of the measure $\mu^{u(0)}$ to the measure $\mu^{\tilde u(0)}$. In view of the approximation property established in (3.13), we can restrict our attention to the case when ψ_0 is strictly increasing on $\mathbb{R}$. For any $t > 0$, we define a transportation plan $(\psi^t, \phi_1^t, \phi_2^t)$ of the measure $\mu^{u(t)}$ to $\mu^{\tilde u(t)}$ as follows:

$$\psi^t\big(\xi(t,y)\big) \doteq \tilde\xi(t,\ \tilde y) \quad \text{for} \quad \tilde y = \psi_0(y),$$
$$\phi_1^t\big(\xi(t,y)\big) \doteq \begin{cases} \phi_{1,0}(y) & \text{if} \quad T(y) > t \quad \text{and} \quad \widetilde T(\tilde y) > t \quad \text{for } \tilde y = \psi_0(y), \\ 0 & \text{otherwise}, \end{cases}$$
$$\phi_2^t\big(\tilde\xi(t,\ \tilde y)\big) \doteq \begin{cases} \phi_{2,0}(\tilde y) & \text{if} \quad T(y) > t \quad \text{and} \quad \widetilde T(\tilde y) > t \quad \text{for } y = \psi_0^{-1}(\tilde y), \\ 0 & \text{otherwise}. \end{cases}$$

If initially the point y is mapped to $\tilde{y} = \psi_0(y)$, then at any later time $t > 0$ the point $\xi(t, y)$ along the u-characteristic starting from y is sent to the point $\tilde{\xi}(t, \tilde{y})$ along the $\tilde{u}$-characteristic starting from $\tilde{y} = \psi_0(y)$. We thus transport the mass from the point $(\xi(t, y),\ u(t, \xi(t, y)),\ \arctan u_x(t, \xi(t, y)))$ to the corresponding point $(\tilde{\xi}(t, \tilde{y}), \tilde{u}(t, \tilde{\xi}(t, \tilde{y})), \arctan \tilde{u}_x(t, \tilde{\xi}(t, \tilde{y})))$ with $\tilde{y} = \psi_0(y)$, except in the case where blow-up has occurred within time t along one (or both) of the characteristics $\xi(\cdot, y)$, $\tilde{\xi}(\cdot, \tilde{y})$. In this later case, the mass is transported to the point ∞.

To check (3.10), it suffices to show that a.e.

$$\begin{aligned}\phi_1^t(\xi(t, y)) \cdot u_x^2(t, \xi(t, y)) \cdot \xi_x(t, y) = \phi_2^t(\psi^t(\xi(t, y))) \cdot (\psi^t)'(\xi(t, y))\\ \cdot \xi_x(t, y) \cdot u_x^2(t, \psi^t(\xi(t, y))).\end{aligned}$$

Since the relations $\tilde{y} = \psi_0(y)$, $\psi^t(\xi(t, y)) = \tilde{\xi}(t, \tilde{y})$, and $(\psi^t)'(\xi(t, y)) \cdot \xi_x(t, y) = \tilde{\xi}_x(t, \psi_0(y)) \cdot \psi_0'(y)$ all hold a.e., the desired identity holds a.e. on the complement of the set $\{y :\ \tilde{y} = \psi_0(y),\ T(y) > t,\ \tilde{T}(\tilde{y}) > t\}$ where both sides equal zero since $\phi_1^t(\xi(t, y)) = \phi_2^t(\tilde{\xi}(t, \tilde{y})) = 0$. The identity holds also a.e. on the set $\{y :\ \tilde{y} = \psi_0(y), T(y) > t,\ \tilde{T}(\tilde{y}) > t\}$ since there, in view of (2.9)–(2.10), it practically amounts to relation (3.10) for $(\phi_{1,0},\ \phi_{2,0},\ \psi_0)$.

In the following, our main goal is to provide an estimate on the time derivative of the function

$$J(t) = J^{(\psi^t, \phi_1^t, \phi_2^t)}\big(u(t), \tilde{u}(t)\big).$$

Throughout the remainder of this section, by $\{\tilde{T}(\tilde{y}) \gtrless t\}$ we understand the set of all $y \in \mathbb{R}$ such that $\psi_0(y) = \{\tilde{y}\}$ and $\tilde{T}(\tilde{y}) \gtrless t$. Since $u_x^2(t, \xi(t, y)) \cdot \xi_x(t, y) = \bar{u}_x^2(y)$ on $\{T(y) > t\}$ by (2.9)–(2.10) and $\tilde{u}_x^2(t, \tilde{\xi}(t, \tilde{y})) \cdot \tilde{\xi}_x(t, \tilde{y}) = \tilde{u}_x^2(0, \tilde{y})$ on $\{\tilde{T}(\tilde{y}) > t\}$, while $\psi^t(\xi(t, y)) = \tilde{\xi}(t, \tilde{y})$ for $\psi_0(y) = \{\tilde{y}\}$, performing the change of variables $y \mapsto \xi(t, y)$, we see that

$$\begin{aligned}J(t) = \int_{\{T(y)>t,\, \tilde{T}(\tilde{y})>t\}} &\min\{|\xi(t, y) - \tilde{\xi}(t, \tilde{y})| + |u(t, \xi(t, y)) - \tilde{u}(t, \tilde{\xi}(t, \tilde{y}))|\\ &\quad + \kappa_0 |\arctan u_x(t, \xi(t, y)) - \arctan \tilde{u}_x(t, \tilde{\xi}(t, \tilde{y}))|,\\ &\quad \kappa_0(\pi + \arctan u_x(t, \xi(t, y)) + \arctan \tilde{u}_x(t, \tilde{\xi}(t, \tilde{y})))\}\, \phi_{1,0}(y)\, \bar{u}_x^2(y)\, dy\\ &+ \kappa_0 \int_{\{T(y)>t,\, \tilde{T}(\tilde{y})>t\}} \Big(\frac{\pi}{2} + \arctan u_x(t, \xi(t, y))\Big)\, (1 - \phi_{1,0}(y))\, \bar{u}_x^2(y)\, dy\\ &+ \kappa_0 \int_{\{T(y)\le t \text{ or } \tilde{T}(\tilde{y})\le t\}} \Big(\frac{\pi}{2} + \arctan u_x(t, \xi(t, y))\Big)\, \bar{u}_x^2(y)\, dy\\ &+ \kappa_0 \int_{\{T(y)>t,\, \tilde{T}(\tilde{y})>t\}} \Big(\frac{\pi}{2} + \arctan \tilde{u}_x(t, \tilde{\xi}(t,\, \tilde{y}))\Big)\, (1 - \phi_{2,0}(\tilde{y}))\, \tilde{u}_x^2(0, \tilde{y})\, \psi_0'(y)\, dy\\ &+ \kappa_0 \int_{\{T(y)\le t \text{ or } \tilde{T}(\tilde{y})\le t\}} \Big(\frac{\pi}{2} + \arctan \tilde{u}_x(t,\, \tilde{\xi}(t,\, \tilde{y}))\Big)\, \tilde{u}_x^2(0, \tilde{y})\, \psi_0'(y)\, dy.\end{aligned}$$

To simplify notation, let

$$S(t) = \{T(y) > t,\, \tilde{T}(\tilde{y}) > t\}, \qquad S^c(t) = \mathbb{R} - S(t), \tag{3.29}$$

$$\begin{aligned}E(t, y) = \min\{&|u(t, \xi(t, y)) - \tilde{u}(t, \tilde{\xi}(t, \tilde{y}))| + \kappa_0 |\arctan u_x(t, \xi(t, y))\\ &- \arctan \tilde{u}_x(t, \tilde{\xi}(t, \tilde{y}))| + |\xi(t, y) - \tilde{\xi}(t, \tilde{y})|,\\ &\kappa_0(\pi + \arctan u_x(t, \xi(t, y)) + \arctan \tilde{u}_x(t, \tilde{\xi}(t, \tilde{y})))\}.\end{aligned} \tag{3.30}$$

Since for $h \geq 0$ we have

$$S(t+h) \subset S(t), \qquad S^c(t) \subset S^c(t+h), \tag{3.31}$$

we deduce that

(3.32)

$$\begin{aligned}
J(t+h) - J(t) = &\int_{S(t)} (E(t+h,y) - E(t,y))\, \phi_{1,0}(y)\, \bar{u}_x^2(y)\, dy \\
&- \int_{S(t)\setminus S(t+h)} E(t+h,y)\, \phi_{1,0}(y)\, \bar{u}_x^2(y)\, dy \\
&+ \kappa_0 \int_{S(t)} (\arctan u_x(t+h,\xi(t+h,y)) - \arctan u_x(t,\xi(t,y)))(1-\phi_{1,0}(y))\, \bar{u}_x^2(y)\, dy \\
&+ \kappa_0 \int_{S(t)} (\arctan \tilde{u}_x(t+h,\tilde{\xi}(t+h,\tilde{y})) \\
&\qquad - \arctan \tilde{u}_x(t,\tilde{\xi}(t,\tilde{y})))(1-\phi_{2,0}(\tilde{y}))\, \tilde{u}_x^2(0,\tilde{y})\, \psi_0'(y)\, dy \\
&- \kappa_0 \int_{S(t)\setminus S(t+h)} \Big(\frac{\pi}{2} + \arctan u_x(t+h,\xi(t+h,y))\Big)(1-\phi_{1,0}(y))\, \bar{u}_x^2(y)\, dy \\
&- \kappa_0 \int_{S(t)\setminus S(t+h)} \Big(\frac{\pi}{2} + \arctan \tilde{u}_x(t+h,\tilde{\xi}(t+h,\tilde{y}))\Big)(1-\phi_{2,0}(\tilde{y}))\, \tilde{u}_x^2(0,\tilde{y})\, \psi_0'(y)\, dy \\
&+ \kappa_0 \int_{S^c(t)} (\arctan u_x(t+h,\xi(t+h,y)) - \arctan u_x(t,\xi(t,y)))\bar{u}_x^2(y)\, dy \\
&+ \kappa_0 \int_{S^c(t)} (\arctan \tilde{u}_x(t+h,\tilde{\xi}(t+h,\tilde{y})) - \arctan \tilde{u}_x(t,\tilde{\xi}(t,\tilde{y})))\tilde{u}_x^2(0,\tilde{y})\, \psi_0'(y)\, dy \\
&+ \kappa_0 \int_{S^c(t+h)\setminus S^c(t)} \Big(\frac{\pi}{2} + \arctan u_x(t+h,\xi(t+h,y))\Big)\bar{u}_x^2(y)\, dy \\
&+ \kappa_0 \int_{S^c(t+h)\setminus S^c(t)} \Big(\frac{\pi}{2} + \arctan \tilde{u}_x(t+h,\tilde{\xi}(t+h,\tilde{y}))\Big)\, \tilde{u}_x^2(0,\tilde{y})\, \psi_0'(y)\, dy.
\end{aligned}$$

Noticing that $S(t) \setminus S(t+h) = S(t) \cap S^c(t+h) = S^c(t+h) \setminus S^c(t)$, we see that the combination of the fifth and ninth terms above, with that of the sixth and tenth, added to the second term, amount to

(3.33)

$$\begin{aligned}
&\kappa_0 \int_{S(t)\setminus S(t+h)} \Big(\frac{\pi}{2} + \arctan u_x(t+h,\xi(t+h,y))\Big)\, \phi_{1,0}(y)\, \bar{u}_x^2(y)\, dy \\
&\quad + \kappa_0 \int_{S(t)\setminus S(t+h)} \Big(\frac{\pi}{2} + \arctan \tilde{u}_x(t+h,\tilde{\xi}(t+h,\tilde{y}))\Big)\, \phi_{2,0}(\tilde{y})\, \tilde{u}_x^2(0,\tilde{y})\, \psi_0'(y)\, dy \\
&\quad - \int_{S(t)\setminus S(t+h)} E(t+h,y)\, \phi_{1,0}(y)\, \bar{u}_x^2(y)\, dy \\
&= \kappa_0 \int_{S(t)\setminus S(t+h)} (\pi + \arctan u_x(t+h,\xi(t+h,y)) \\
&\qquad\qquad + \arctan \tilde{u}_x(t+h,\tilde{\xi}(t+h,\tilde{y})))\phi_{1,0}(y)\, \bar{u}_x^2(y)\, dy \\
&\quad - \int_{S(t)\setminus S(t+h)} E(t+h,y)\, \phi_{1,0}(y)\, \bar{u}_x^2(y)\, dy
\end{aligned}$$

by (3.10) for $(\phi_{1,0},\, \phi_{2,0},\, \psi_0)$. In view of (3.29)–(3.30), on $S(t)\setminus S(t+h)$ we have that

$$E(t+h,y)=\pi+\arctan u_x(t+h,\xi(t+h,y))+\arctan \tilde u_x(t+h,\tilde\xi(t+h,\tilde y))$$

since at least one of the expressions $u_x(t+h,\xi(t+h,y))$ and $\tilde u_x(t+h,\tilde\xi(t+h,\tilde y))$ is precisely $-\infty$ on this set. Thus the whole expression (3.33) is identically zero. Therefore (3.32) yields

(3.34)

$$\begin{aligned}
J(t+h)-J(t) &= \int_{S(t)} (E(t+h,y)-E(t,y))\,\phi_{1,0}(y)\,\bar u_x^2(y)\,dy\\
&+\kappa_0\int_{S(t)}(\arctan u_x(t+h,\xi(t+h,y))-\arctan u_x(t,\xi(t,y)))(1-\phi_{1,0}(y))\,\bar u_x^2(y)\,dy\\
&+\kappa_0\int_{S(t)}(\arctan \tilde u_x(t+h,\tilde\xi(t+h,\tilde y))\\
&\qquad -\arctan \tilde u_x(t,\tilde\xi(t,\tilde y)))(1-\phi_{2,0}(\tilde y))\tilde u_x^2(0,\tilde y)\psi_0'(y)\,dy\\
&+\kappa_0\int_{S^c(t)}(\arctan u_x(t+h,\xi(t+h,y))-\arctan u_x(t,\xi(t,y)))\bar u_x^2(y)\,dy\\
&+\kappa_0\int_{S^c(t)}(\arctan \tilde u_x(t+h,\tilde\xi(t+h,\tilde y))-\arctan \tilde u_x(t,\tilde\xi(t,\tilde y)))\tilde u_x^2(0,\tilde y)\,\psi_0'(y)\,dy.
\end{aligned}$$

In view of (3.29), we have $S^c(t)=\{T(y)\le t\}\cup\{\tilde T(\tilde y)\le t\}$. On the set $\{T(y)\le t\}$ we have $\arctan u_x(t+h,\xi(t+h,y))=\arctan u_x(t,\xi(t,y))=-\infty$ so that in the fourth term in (3.34) only the integral over $\{T(y)>t,\,\tilde T(\tilde y)\le t\}$ might have a nonzero contribution. Thus the second and fourth terms in (3.34) combine to

$$\begin{aligned}
(3.35)\qquad &\kappa_0\int_{\{T(y)>t\}}(\arctan u_x(t+h,\xi(t+h,y))-\arctan u_x(t,\xi(t,y)))\\
&\qquad\times(1-\phi_{1,0}(y))\,\bar u_x^2(y)\,dy\\
&+\kappa_0\int_{\{\tilde T(\tilde y)\le t<T(y)\}}(\arctan u_x(t+h,\xi(t+h,y))-\arctan u_x(t,\xi(t,y)))\\
&\qquad\times\phi_{1,0}(y)\,\bar u_x^2(y)\,dy.
\end{aligned}$$

Similarly, the third and fifth terms combine to

$$\begin{aligned}
(3.36)\quad &\kappa_0\int_{\{\tilde T(\tilde y)>t\}}(\arctan \tilde u_x(t+h,\xi(t+h,\tilde y))\\
&\qquad-\arctan \tilde u_x(t,\xi(t,\tilde y)))(1-\phi_{2,0}(\tilde y))\,\tilde u_x^2(0,\tilde y)\,\psi_0'(y)\,dy\\
&+\kappa_0\int_{\{T(y)\le t<\tilde T(\tilde y)\}}(\arctan \tilde u_x(t+h,\xi(t+h,\tilde y))\\
&\qquad-\arctan \tilde u_x(t,\xi(t,\tilde y)))\,\phi_{2,0}(\tilde y)\,\tilde u_x^2(0,\tilde y)\,\psi_0'(y)\,dy.
\end{aligned}$$

To transform suitably the first term in (3.34), let us denote by $E^1(t,y)$ the first expression in the minimum (3.30), and by $E^2(t,y)$ the second. If $E(t,y)=E^2(t,y)$, then

(3.37)

$$\begin{aligned}
E(t+h,y)-E(t,y)\le \kappa_0\,&(\arctan u_x(t+h,\xi(t+h,y))-\arctan u_x(t,\xi(t,y))\\
&+\arctan \tilde u_x(t+h,\tilde\xi(t+h,\tilde y))-\arctan \tilde u_x(t,\tilde\xi(t,\tilde y))),
\end{aligned}$$

since $E(t+h,y) \le E^2(t+h,y)$. On the other hand, if $E(t,y) = E^1(t,y)$, then the triangle inequality and the relation $E(t+h,y) \le E^1(t+h,y)$ ensure that

$$\begin{aligned}(3.38)\quad E(t+h,y) - E(t,y) \le& \left|\xi(t+h,y) - \xi(t,y) + \tilde{\xi}(t+h,\tilde{y}) - \tilde{\xi}(t,\tilde{y})\right| \\ &+ \left|u(t+h,\xi(t+h,y)) - \tilde{u}(t+h,\tilde{\xi}(t+h,\tilde{y})) - u(t,\xi(t,y)) + \tilde{u}(t,\xi(t,\tilde{y}))\right| \\ &+ \kappa_0 \left| \arctan u_x(t+h,\xi(t+h,y)) - \arctan u_x(t,\xi(t,y)) \right. \\ &\qquad \left. + \arctan \tilde{u}_x(t+h,\tilde{\xi}(t+h,\tilde{y})) - \arctan \tilde{u}_x(t,\tilde{\xi}(t,\tilde{y}))\right|.\end{aligned}$$

Letting $h \downarrow 0$ in (3.34), and taking into account (3.38) and the considerations preceding it, we deduce that

(3.39)

$$\begin{aligned}\limsup_{h\downarrow 0} \frac{J(t+h)-J(t)}{h} \le& \kappa_0 J_0(t) + \int_{S(t)} \phi_{1,0}(y)\, \bar{u}_x^2(y) \left|\frac{d}{dt}\,\xi(t,y) - \frac{d}{dt}\,\tilde{\xi}(t,\psi_0(y))\right| dy \\ &+ \int_{S(t)} \phi_{1,0}(y)\, \bar{u}_x^2(y) \left|\frac{d}{dt}\,u(t,\xi(t,y)) - \frac{d}{dt}\,\tilde{u}(\tilde{\xi}(t,\psi_0(y)))\right| dy \\ &+ \kappa_0 \int_{S(t)} \phi_{1,0}(y)\, \bar{u}_x^2(y) \left|\frac{d}{dt} \text{ arctan } u_x(t,\xi(t,y)) - \frac{d}{dt} \text{ arctan } \tilde{u}_x(\tilde{\xi}(t,\psi_0(y)))\right| dy,\end{aligned}$$

where

$$\begin{aligned}J_0(t) =& \int_{[T(y)>t]} (1-\phi_{1,0}(y)) \left[\frac{d}{dt} \text{ arctan } u_x(t,\xi(t,y))\right] \bar{u}_x^2(y)\, dy \\ &+ \int_{[\tilde{T}(\tilde{y})>t]} (1-\phi_{2,0}(\tilde{y})) \left[\frac{d}{dt} \text{ arctan } \tilde{u}_x(t,\xi(t,\tilde{y}))\right] \tilde{u}_x^2(0,\tilde{y})\, d\tilde{y} \\ &+ \int_{[\tilde{T}(\tilde{y})\le t<T(y)]} \phi_{1,0}(y) \left[\frac{d}{dt} \text{ arctan } u_x(t,\xi(t,y))\right] \bar{u}_x^2(y)\, dy \\ &+ \int_{[T(y)\le t<\tilde{T}(\tilde{y})]} \phi_{2,0}(\tilde{y}) \left[\frac{d}{dt} \text{ arctan } \tilde{u}_x(t,\xi(t,\tilde{y}))\right] \tilde{u}_x^2(0,\tilde{y})\, d\tilde{y} \le 0,\end{aligned}$$

the last inequality being true by (2.4).

Before proceeding with further analysis of (3.39), we establish a few a priori bounds. From (2.22) we get

$$(3.40)\qquad \left|\frac{d}{dt}\,\xi(t,y) - \frac{d}{dt}\,\tilde{\xi}(t,\tilde{y})\right| = |u(t,\xi(t,y)) - \tilde{u}(t,\tilde{\xi}(t,\tilde{y}))|.$$

Also, note that if $v = \arctan z(t)$ and $\dot{z} = -\frac{z^2}{2}$, then

$$\dot{v} = -\frac{z^2}{2+2z^2} = -\frac{1}{2}\sin^2 v.$$

Since $|\sin^2\alpha - \sin^2\beta| \le |\alpha-\beta|$ by the mean-value theorem as $|(\sin^2 z)'| = 2|\sin z\, \cos z| = |\sin(2z)| \le 1$, we infer three useful facts if we set $z = u_x(t,\xi(t,x))$. First of all,

$$(3.41)\qquad \frac{d}{dt} \arctan u_x(t,\xi(t,y)),\ \frac{d}{dt} \arctan \tilde{u}_x(t,\tilde{u}(t,\tilde{\xi}(t,\tilde{y}))) \le 0.$$

Second,

$$\begin{aligned}(3.42)\qquad &\left|\frac{d}{dt}\ \text{arctan}\ u(t,\xi(t,y)) - \frac{d}{dt}\ \text{arctan}\ \tilde{u}(t,\tilde{\xi}(t,\tilde{y}))\right|\\ &\le \frac{1}{2}\,|\arctan u_x(t,\xi(t,y)) - \arctan \tilde{u}_x(t,\tilde{\xi}(t,\tilde{y}))|.\end{aligned}$$

Furthermore, if $\arctan z \le -\frac{\pi}{4}$, then $\sin(\arctan z) \in [-1, -\frac{1}{\sqrt{2}}]$, so that

$$(3.43)\qquad \frac{d}{dt}\arctan u_x(t,\xi(t,y)) \le -\frac{1}{4} \quad \text{if} \quad \arctan u_x(t,\xi(t,y)) \le -\frac{\pi}{4}.$$

On the other hand, using first (2.21) and then (2.18), we have

$$\begin{aligned}
&\left|\frac{d}{dt}\,u(t,\xi(t,y_0)) - \frac{d}{dt}\,\tilde{u}(t,\tilde{\xi}(t,\tilde{y}_0))\right| = |\varphi(t,y_0) - \tilde{\varphi}(t,\tilde{y}_0)|\\
&\quad = \frac{1}{4}\Big|\int_{\{T(y)>t\}} \text{sign}(y_0-y)\,\bar{u}_x^2(y)\,dy - \int_{\{\tilde{T}(\tilde{y})>t\}} \text{sign}(\tilde{y}_0-\tilde{y})\,\tilde{u}_x^2(0,\tilde{y})\,d\tilde{y}\Big|\\
&\quad = \frac{1}{4}\Big|\int_{-\infty}^{y_0} \bar{u}_x^2(y)\chi_{[T(y)>t]}dy - \int_{-\infty}^{\tilde{y}_0} \tilde{u}_x^2(0,\tilde{y})\,\chi_{[\tilde{T}(\tilde{y})>t]}d\tilde{y}\\
&\qquad - \int_{y_0}^{\infty} \bar{u}_x^2(y)\chi_{[T(y)>t]}dy - \int_{\tilde{y}_0}^{\infty} \tilde{u}_x^2(0,\tilde{y})\,\chi_{[\tilde{T}(\tilde{y})>t]}d\tilde{y}\Big|\\
&\quad = \frac{1}{4}\Big|\int_{-\infty}^{y_0} \bar{u}_x^2(y)\chi_{[T(y)>t]}dy - \int_{-\infty}^{y_0} \tilde{u}_x^2(0,\psi_0(y))\,\chi_{[\tilde{T}(\tilde{y})>t]}\,\psi_0'(y)\,dy\\
&\qquad - \int_{y_0}^{\infty} \bar{u}_x^2(y)\chi_{[T(y)>t]}dy - \int_{y_0}^{\infty} \tilde{u}_x^2(0,\psi_0(y))\,\chi_{[\tilde{T}(\tilde{y})>t]}\,\psi_0'(y)\,dy\Big|\\
&\quad = \frac{1}{4}\Big|\int_{-\infty}^{y_0}(1-\phi_{1,0}(y)+\phi_{1,0}(y))\bar{u}_x^2(y)\chi_{[T(y)>t]}dy\\
&\qquad - \int_{-\infty}^{y_0}(1-\phi_{2,0}(\psi_0(y))+\phi_{2,0}(\psi_0(y)))\tilde{u}_x^2(0,\psi_0(y))\,\chi_{[\tilde{T}(\tilde{y})>t]}\,\psi_0'(y)\,dy\\
&\qquad - \int_{y_0}^{\infty}(1-\phi_{1,0}(y)+\phi_{1,0}(y))\bar{u}_x^2(y)\chi_{[T(y)>t]}dy\\
&\qquad - \int_{y_0}^{\infty}(1-\phi_{2,0}(\psi_0(y))+\phi_{2,0}(\psi_0(y)))\tilde{u}_x^2(0,\psi_0(y))\,\chi_{[\tilde{T}(\tilde{y})>t]}\,\psi_0'(y)\,dy\Big|
\end{aligned}$$

after performing in the next to last step in two of the integrals the change of variables $\tilde{y} = \psi_0(y)$. Since equation (3.10) for $(\psi_0, \phi_{1,0}, \phi_{2,0})$ ensures that $\phi_{1,0}(y)\,\bar{u}_x^2(y) = \phi_{2,0}(\psi_0(y))\,\tilde{u}_x^2(0,\psi_0(y))\,\psi_0'(y)$ a.e. on $S(t)$, we deduce that

$$\begin{aligned}(3.44)\qquad &\Big|\frac{d}{dt}\,u(t,\xi(t,y_0)) - \frac{d}{dt}\,\tilde{u}(t,\tilde{\xi}(t,\tilde{y}_0))\Big|\\
&\le \frac{1}{4}\int_{[T(y)>t]}(1-\phi_{1,0}(y))\,\bar{u}_x^2(y)\,dy + \frac{1}{4}\int_{[\tilde{T}(\tilde{y})>t]}(1-\phi_{2,0}(\tilde{y}))\tilde{u}_x^2(0,\tilde{y})\,d\tilde{y}\\
&\quad + \frac{1}{4}\int_{[\tilde{T}(\tilde{y})\le t<T(y)]} \phi_{1,0}(y)\,\bar{u}_x^2(y)\,dy + \frac{1}{4}\int_{[T(y)\le t<\tilde{T}(\tilde{y})]} \phi_{2,0}(\tilde{y})\,\tilde{u}_x^2(0,\tilde{y})\,d\tilde{y}.\end{aligned}$$

Let us now introduce the following sets:

$$
\begin{aligned}
S_1 &= \Big\{y \in \mathbb{R}:\ \arctan\, u_x(t,\xi(t,y)) \le -\frac{\pi}{4} \quad \text{and} \quad \arctan\, \tilde{u}_x(t,\tilde{\xi}(t,\tilde{y})) \le -\frac{\pi}{4}\Big\},\\
S_2 &= \Big\{y \in \mathbb{R}:\ \arctan\, u_x(t,\xi(t,y)) > -\frac{\pi}{4} \quad \text{and} \quad \arctan\, \tilde{u}_x(t,\tilde{\xi}(t,\tilde{y})) > -\frac{\pi}{4}\Big\},\\
S_3 &= \Big\{y \in \mathbb{R}:\ \arctan\, u_x(t,\xi(t,y)) > -\frac{\pi}{4} \ge \arctan\, \tilde{u}_x(t,\tilde{\xi}(t,\tilde{y}))\Big\},\\
S_4 &= \Big\{y \in \mathbb{R}:\ \arctan\, \tilde{u}_x(t,\tilde{\xi}(t,\tilde{y})) > -\frac{\pi}{4} \ge \arctan\, u_x(t,\xi(t,y))\Big\}.
\end{aligned}
$$

The integral on the right-hand side of (3.44) over S_1 is, in view of (3.43), bounded from above by $|J_0(t)| = -\,J_0(t)$. The integral over S_2 is, in view of the formula for $J(t)$ preceding relation (2.29), bounded from above by $\frac{J(t)}{\pi\kappa_0}$. To evaluate the contribution over the integral over S_3, notice that the same formula for $J(t)$ yields

$$
\begin{aligned}
J(t) \ge{}& \kappa_0 \int_{S(t)\,\cap\, S_3} (\arctan u_x(t,\xi(t,y)) - \arctan \tilde{u}_x(t,\tilde{\xi}(t,\tilde{y})))\, \phi_{1,0}(y)\, \bar{u}_x^2(y)\, dy\\
&+ \kappa_0 \int_{S(t)\,\cap\, S_3} \Big(\frac{\pi}{2} + \arctan u_x(t,\xi(t,y))\Big)\, (1-\phi_{1,0}(y))\, \bar{u}_x^2(y)\, dy\\
&+ \kappa_0 \int_{S^c(t)\,\cap\, S_3} \Big(\frac{\pi}{2} + \arctan u_x(t,\xi(t,y))\Big)\, \bar{u}_x^2(y)\, dy\\
&+ \kappa_0 \int_{S(t)\,\cap\, S_3} \Big(\frac{\pi}{2} + \arctan \tilde{u}_x(t,\tilde{\xi}(t,\ \tilde{y}))\Big)\, (1-\phi_{2,0}(\tilde{y}))\, \tilde{u}_x^2(0,\tilde{y})\, \psi_0'(y)\, dy\\
&+ \kappa_0 \int_{S^c(t)\,\cap\, S_3} \Big(\frac{\pi}{2} + \arctan \tilde{u}_x(t,\ \tilde{\xi}(t,\ \tilde{y}))\Big)\, \tilde{u}_x^2(0,\tilde{y})\, \psi_0'(y)\, dy.
\end{aligned}
$$

Using (3.10), the sum of the first and fourth terms is larger than

$$
\begin{aligned}
&\kappa_0 \int_{S(t)\,\cap\, S_3} \Big(\frac{\pi}{2} + \arctan u_x(t,\xi(t,y))\Big)\, \tilde{u}_x^2(0,\tilde{y})\, \psi_0'(y)\, dy\\
&\quad\ge \frac{\kappa_0\pi}{4} \int_{S(t)\,\cap\, S_3} \tilde{u}_x^2(0,\tilde{y})\, \psi_0'(y)\, dy = \frac{\kappa_0\pi}{4} \int_{S(t)\,\cap\, S_3} \tilde{u}_x^2(0,\tilde{y})\, d\tilde{y}\\
&\quad\ge \frac{\kappa_0\pi}{4} \int_{S(t)\,\cap\, S_3} (1-\phi_{2,0}(\tilde{y}))\, \tilde{u}_x^2(0,\tilde{y})\, d\tilde{y} = \frac{\kappa_0\pi}{4} \int_{[\tilde{T}(\tilde{y})>t]\,\cap\, S_3} (1-\phi_{2,0}(\tilde{y}))\, \tilde{u}_x^2(0,\tilde{y})\, d\tilde{y},
\end{aligned}
$$

with the last equality enforced by $S_3 \subset [T(y) > t]$. The second term is bounded from below by

$$
\begin{aligned}
&\frac{\kappa_0\pi}{4} \int_{S(t)\cap\, S_3} (1-\phi_{1,0}(y))\, \bar{u}_x^2(y)\, dy = \frac{\kappa_0\pi}{4} \int_{[T(y)>t]\cap\, S_3} (1-\phi_{1,0}(y))\, \bar{u}_x^2(y)\, dy\\
&-\frac{\kappa_0\pi}{4} \int_{[T(y)>t\ge\tilde{T}(\tilde{y})]\cap\, S_3} (1-\phi_{1,0}(y))\bar{u}_x^2(y)dy = \frac{\kappa_0\pi}{4} \int_{[T(y)>t]\cap S_3} (1-\phi_{1,0}(y))\bar{u}_x^2(y)\, dy\\
&-\frac{\kappa_0\pi}{4} \int_{[T(y)>t\ge\tilde{T}(\tilde{y})]\cap S_3} \bar{u}_x^2(y)dy + \frac{\kappa_0\pi}{4} \int_{[T(y)>t\ge\tilde{T}(\tilde{y})]\cap S_3} \phi_{1,0}(y)\bar{u}_x^2(y)dy
\end{aligned}
$$

if we recall (3.10). The third term is bounded from below by

$$
\frac{\kappa_0\pi}{4} \int_{S^c(t)\,\cap\, S_3} \bar{u}_x^2(y)\, dy \ge \frac{\kappa_0\pi}{4} \int_{[T(y)>t\ge\tilde{T}(\tilde{y})]\,\cap\, S_3} \bar{u}_x^2(y)\, dy.
$$

Summing up, we get

$$J(t) \geq \frac{\kappa\pi}{4}\Bigg\{ \int_{[T(y)>t]\cap S_3} (1-\phi_{1,0}(y))\bar{u}_x^2(y)dy + \int_{[\tilde{T}(\tilde{y})>t]\cap S_3} (1-\phi_{2,0}(\tilde{y}))\tilde{u}_x^2(0,\tilde{y})\,d\tilde{y}$$

$$+ \int_{[\tilde{T}(\tilde{y})\leq t<T(y)]\cap S_3} \phi_{1,0}(y)\bar{u}_x^2(y)dy + \int_{[T(y)\leq t<\tilde{T}(\tilde{y})]\cap S_3} \phi_{2,0}(\tilde{y})\tilde{u}_x^2(0,\tilde{y})d\tilde{y}\Bigg\},$$

since the last term on the right is zero as $S_3 \subset [T(y) > t]$. A similar relation holds with S_4 instead of S_3. Consequently, putting together all this information about the various inequalities valid on the disjoint sets S_1, S_2, S_3, and S_4, we conclude by (3.44) that

$$\left|\frac{d}{dt}\,u(t,\xi(t,y_0)) - \frac{d}{dt}\,\tilde{u}(t,\tilde{\xi}(t,\tilde{y}_0))\right| \leq -J_0(t) + \frac{3J(t)}{\kappa_0\pi}. \tag{3.45}$$

To obtain now a suitable estimate on

$$\limsup_{h\downarrow 0}\frac{J(t+h)-J(t)}{h} = \limsup_{h\downarrow 0}\int_{S(t)} \frac{E(t+h,y)-E(t,y)}{h}\phi_{1,0}(y)\,\bar{u}_x^2(y)\,dy + \kappa_0\,J_0(t) \tag{3.46}$$

we distinguish two cases. If $E(t,y)$ is the second component $E^2(t,y)$ of the minimum in (3.30), then by (3.37) and (3.41) we can estimate the contribution of the first term in (3.46) by zero from above. On the other hand, if the minimum is $E^1(t,y)$, then the first integral term in (3.46) is not larger (pointwise) than the nonnegative expression

$$\left(E(t,y) + \frac{3J(t)}{\kappa_0\pi} - J_0(t)\right)\,\phi_{1,0}(y)\,\bar{u}_x^2(y)$$

in view of the estimates (3.40), (3.42), and (3.45). We conclude that

$$\limsup_{h\downarrow 0}\frac{J(t+h)-J(t)}{h} \leq J(t) + \left(\frac{3J(t)}{\kappa_0\pi} - J_0(t)\right)\,\|\bar{u}_x^2\|_{L^1(\mathbb{R})} + \kappa_0\,J_0(t).$$

Since $J_0(t) \leq 0$, choosing the constant $\kappa_0 \doteq \|\bar{u}_x^2\|_{\mathbf{L}^1(\mathbb{R})}$ we now have

$$\frac{d}{dt}J^{(\psi^t,\phi_1^t,\phi_2^t)}(u(t),\,v(t)) \leq 2\,J^{(\psi^t,\phi_1^t,\phi_2^t)}(u(t),\,v(t)).$$

Optimizing over all triples $(\psi^0,\phi_1^0,\phi_2^0)$ we conclude

$$J(u(t),\,v(t)) \leq J(u(0),\,v(0))\,e^{2t}, \quad t \geq 0. \tag{3.47}$$

Summing up the considerations made above, we proved the following result.

THEOREM 3.1. *The trajectories $t \mapsto u(t)$ of* (1.1) *constructed in Theorem* 2.1 *are locally Lipschitz continuous as maps from $[0,\infty)$ into the metric space $\mathcal{X}$ equipped with the distance functional J. Moreover, the distance between two trajectories is also locally Lipschitz continuous as a map from $[0,\infty)$ into $\mathcal{X}$.*

4. Concluding remarks. The following example shows that, in some sense, our distance functional J in (3.11) is "sharp." Indeed, the convergence of the initial data in $\mathbf{L}^\infty(\mathbb{R}) \cap \mathbf{L}^1(\mathbb{R})$ together with the weak convergence of the derivatives $\bar u_x$ and $\bar u_x^2$ in $\mathbf{L}^2(\mathbb{R})$ does not guarantee the convergence of the corresponding solutions at later times $t > 0$.

Example 1. Consider the functions $f, g : [0,1] \mapsto [0,1]$ defined as

$$f(x) \doteq \begin{cases} 1-2x & \text{if } x \in [0,\ 1/2], \\ 0 & \text{if } x \in [1/2,\ 1], \end{cases} \qquad g(x) \doteq \begin{cases} 1-3x & \text{if } x \in [0,\ 1/6], \\ 1/2 & \text{if } x \in [1/6,\ 1/2], \\ 1-x & \text{if } x \in [1/2,\ 1]. \end{cases}$$

Observe that

$$\int_0^1 f'(x)\,dx = \int_0^1 g'(x)\,dx = -1, \qquad \int_0^1 \big[f'(x)\big]^2\,dx = \int_0^1 \big[g'(x)\big]^2\,dx = 2.$$

Next, consider the function

$$h(x) \doteq \begin{cases} 1-|x| & \text{if } |x| \le 1, \\ 0 & \text{if } |x| \ge 1, \end{cases}$$

and define the sequences of initial values

$$\bar u_n(x) = \begin{cases} h(x) & \text{if } x \notin [0,1], \\ h(i/n) + \frac{1}{n} f(nx-i+1) & \text{if } x \in \left[\frac{i-1}{n},\ \frac{i}{n}\right], \quad i = 1,\dots,n, \end{cases}$$
$$\bar v_n(x) = \begin{cases} h(x) & \text{if } x \notin [0,1], \\ h(i/n) + \frac{1}{n} g(nx-i+1) & \text{if } x \in \left[\frac{i-1}{n},\ \frac{i}{n}\right], \quad i = 1,\dots,n. \end{cases}$$

Letting $n \to \infty$ we now have the strong convergence $\big\|\bar u_n - \bar v_n\big\|_{\mathbf{L}^\infty(\mathbb{R})} \to 0$. Moreover, by construction it is easy to see that at each point $x \in [0,1]$,

$$\lim_{n\to\infty} \int_0^x \big((\bar u_n)_x(y) - (\bar v_n)_x(y)\big)\,dy = \lim_{n\to\infty} \int_0^x \big((\bar u_n)_x^2(y) - (\bar v_n)_x^2(y)\big)\,dy = 0$$

so that in $\mathbf{L}^2[0,1]$ one has the weak convergence

$$(\bar u_n)_x - (\bar v_n)_x \rightharpoonup 0, \qquad (\bar u_n)_x^2 - (\bar v_n)_x^2 \rightharpoonup 0,$$

since both sequences are bounded in $\mathbf{L}^2[0,1]$ and the previous observation identifies the zero function as the only possible weak limit. However,

$$u(t) \doteq \lim_{n\to\infty} u_n(t) \;\neq\; \lim_{n\to\infty} v_n(t) \doteq v(t)$$

for every $t \in (2/3, 1)$, where $T = 2/3$ is the time at which the gradients of the functions v_n blow up. The last assertion follows at once from (2.27).

We also would like to highlight the importance of requiring that the transport map ψ in (3.10) be monotone nondecreasing. If in (3.11) we were to take the minimization over all maps ψ, not necessarily monotone, we would obtain the classical Kantorovich–Rubinstein distance between measures, which generates the weak topology on the space of bounded, positive measures [V]. By restricting ourselves to monotone nondecreasing maps ψ, the corresponding distance functional generates a much stronger topology.

Example 2. Consider the sequence of Lipschitz functions

$$u^m(x) \doteq \begin{cases} 0 & \text{if } x \notin [0,1], \\ x-(i-1)/m & \text{if } (i-1)/m \le x \le (2i-1)/2m, \\ i/m - x & \text{if } (2i-1)/2m \le x \le i/m, \end{cases} \qquad i=1,\dots,m.$$

In this case, $u_x = \pm 1$ and $\arctan u_x = \pm\pi/4$. The corresponding measures μ^{u^m} defined at (3.9) converge weakly to the measure μ on $\mathbb{R}^2 \times [-\pi/2,\ \pi/2]$ defined as

$$\mu(A) \doteq \frac{1}{2}\,\text{meas}\Big\{x\in[0,1];(x,\,0,\,\pi/4)\in A\Big\} + \frac{1}{2}\,\text{meas}\Big\{x\in[0,1]\,;\ \ (x,\,0,\,-\pi/4)\in A\Big\}.$$

In particular, these measures form a Cauchy sequence in the Kantorovich metric. However, these same functions u^m do not form a Cauchy sequence with respect to the distance J. Indeed, let $m<n$. Notice that in our case $\kappa_0 \doteq \|\bar u_x^2\|_{\mathbf{L}^1(\mathbb{R})} = 1$. Consider the open intervals

$$I_i^{m+} = \left]\frac{i-1}{m},\frac{2i-1}{2m}\right[, \qquad\qquad I_i^{m-} = \left]\frac{2i-1}{2m},\frac{i}{m}\right[,$$

where u_x^m takes the values $+1$ and -1, respectively. Define the intervals I_j^{n+}, I_j^{n-} similarly. Now consider any transportation plan (ψ,ϕ_1,ϕ_2) relating u^m to u^n via (3.10), with ψ nondecreasing. For each $i=1,\dots,m$, call ν_i the number of distinct intervals I_j^{n+} which intersect the image $\psi(I_i^{m+})$. Since ψ is monotone, if $\nu_i\ge 2$, this implies that the image $\psi(I_i^{m+})$ entirely covers at least ν_i-1 distinct intervals I_j^{n-}. Because $u_x^m=1$ on I_i^{m+} and $u_x^n=-1$ on each I_j^{n-}, on the union of these intervals I_j^{n-} we have $\arctan u_x^m(\psi(x)) = -\arctan u_x^n(x) = \frac{\pi}{4}$ so that the integrand contribution from the first two parts of (3.11) is pointwise larger than $\frac{\pi}{2}\phi_1(x)+\frac{\pi}{4}\,(1-\phi_1(x))\ge\frac{\pi}{4}$, which therefore accounts for a cost $\ge \pi(\nu_i-1)/8n$. Next, if $\nu_1+\cdots+\nu_m = n^* < n$, there must be $n-n^*$ intervals $I^{n+}_{j(1)},\dots,I^{n+}_{j(n^*-n)}$ which do not intersect any of the sets $\psi(I_i^{m+})$ for $i=1,\dots,m$. These intervals must be contained in the image of some I_i^{m-}, or in the image of the set $\psi\big(\mathbb{R}\setminus[0,1]\big)$, where $u^m\equiv 0$. This accounts for a cost $\ge \pi(n-n^*)/4n$.

The above argument shows that, for any $m<n$, the cost of any transportation plan relating u^m to u^n is bounded below by

$$J^{(\psi,\phi_1,\phi_2)}(u^m,u^n) \ge \frac{\pi}{8n}\cdot\max\left\{\sum_{i=1}^m(\nu_i-1),\ n-\sum_{i=1}^m\nu_i\right\} \ge \frac{\pi}{8n}\cdot\frac{n-m}{2}.$$

For any fixed m, the right-hand side approaches $\pi/16$ as $n\to\infty$. Therefore, the above is not a Cauchy sequence in our transportation metric.

Acknowledgment. The authors are grateful to both referees for useful comments.

REFERENCES

[AA] G. Alberti and L. Ambrosio, *A geometrical approach to monotone functions in* $\mathbb{R}^n$, Math. Z., 230 (1999), pp. 259–316.

[BSS] R. Beals, D. Sattinger, and J. Szmigielski, *Inverse scattering solutions of the Hunter–Saxton equation*, Appl. Anal., 78 (2001), pp. 255–269.

[B] A. Bressan, *Unique solutions for a class of discontinuous differential equations*, Proc. Amer. Math. Soc., 104 (1988), pp. 772–778.

[BC] A. Bressan and G. Colombo, *Existence and continuous dependence for discontinuous O.D.E.'s*, Boll. Un. Mat. Ital. B (7), 4 (1990), pp. 295–311.

[BF] A. Bressan and M. Fonte, *An optimal transportation metric for solutions of the Camassa-Holm equation*, Math. Meth. Anal., to appear.

[BZZ] A. Bressan, P. Zhang, and Y. Zheng, *On asymptotic variational wave equations*, Arch. Rational Mech. Anal., to appear.

[BGH] G. Buttazzo, M. Giaquinta, and S. Hildebrandt, *One-dimensional Variational Problems*, Clarendon Press, Oxford, 1998.

[CH] R. Camassa and D. D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.

[C] A. Constantin, *Existence of permanent and breaking waves for a shallow water equation: A geometric approach*, Ann. Inst. Fourier (Grenoble), 50 (2000), pp. 321–362.

[CE] A. Constantin and J. Escher, *Wave breaking for nonlinear nonlocal shallow water equations*, Acta Math., 181 (1998), pp. 229–243.

[CK] A. Constantin and B. Kolev, *Geodesic flow on the diffeomorphism group of the circle*, Comment. Math. Helv., 78 (2003), pp. 787–804.

[CS] A. Constantin and W. Strauss, *Stability of peakons*, Comm. Pure Appl. Math., 53 (2000), pp. 603–610.

[DP] H. H. Dai and M. Pavolv, *Transformations for the Camassa-Holm equation, its high-frequency limit and the Sinh–Gordon equation*, J. Phys. Soc. Japan, 67 (1998), pp. 3655–3657.

[EG] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[HS] J. K. Hunter and R. Saxton, *Dynamics of director fields*, SIAM J. Appl. Math., 51 (1991), pp. 1498–1521.

[HZ1] J. K. Hunter and Y. Zheng, *On a completely integrable nonlinear hyperbolic variational equation*, Phys. D, 79 (1994), pp. 361–386.

[HZ2] J. K. Hunter and Y. Zheng, *On a nonlinear hyperbolic variational equation* I, Arch. Rational Mech. Anal., 129 (1995), pp. 305–353.

[HZ3] J. K. Hunter and Y. Zheng, *On a nonlinear hyperbolic variational equation* II, Arch. Rational Mech. Anal., 129 (1995), pp. 355–383.

[J] R. S. Johnson, *Camassa–Holm, Korteweg–de Vries and related models for water waves*, J. Fluid Mech., 455 (2002), pp. 63–82.

[KM] B. Khesin and G. Misiolek, *Euler equations on homogeneous spaces and Virasoro orbits*, Adv. Math., 176 (2003), pp. 116–144.

[K] S. Kouranbaeva, *The Camassa–Holm equation as a geodesic flow on the diffeomorphism group*, J. Math. Phys., 40 (1999), pp. 857–868.

[N] J. Natanson, *Theory of Functions of a Real Variable*, F. Ungar, New York, 1964.

[V] C. Villani, *Topics in Optimal Transportation*, AMS, Providence, RI, 2003.

[Z] E. Zeidler, *Nonlinear Functional Analysis and its Applications*, Springer-Verlag, New York, 1990.

[ZZ1] P. Zhang and Y. Zheng, *On oscillations of an asymptotic equation of a nonlinear variational wave equation*, Asymptot. Anal., 18 (1998), pp. 307–327.

[ZZ2] P. Zhang and Y. Zheng, *On the existence and uniqueness of solutions to an asymptotic equation of a variational wave equation*, Acta Math. Sinica, 15 (1999), pp. 115–130.

[ZZ3] P. Zhang and Y. Zheng, *Existence and uniqueness of solutions of an asymptotic equation arising from a variational wave equation with general data*, Arch. Ration. Mech. Anal., 155 (2000), pp. 49–83.

SIAM J. MATH. ANAL.
Vol. 37, No. 4, pp. 1027–1043

AN INVERSE PROBLEM FOR MAXWELL'S EQUATIONS IN BI-ISOTROPIC MEDIA*

SHUMIN LI†

Abstract. In this paper, we consider Maxwell's equations in a bi-isotropic and inhomogeneous medium. We discuss an inverse problem of determining the coefficients ε, ζ, μ in the constitutive relations from a finite number of interior measurements. We prove a Lipschitz stability estimate for the inverse problem by applying the argument on the basis of Carleman estimate.

Key words. inverse problem, Maxwell's equations, bi-isotropic, magnetoelectric, Carleman estimate, stability

AMS subject classifications. 35R25, 35R30, 35Q60

DOI. 10.1137/S003614100444366X

1. Introduction and main results.

We consider Maxwell's equations in a bi-isotropic and inhomogeneous medium:

$$\begin{cases} \partial_t D(x,t) - \nabla \times H(x,t) = 0, & (x,t) \in Q, \\ \partial_t B(x,t) + \nabla \times E(x,t) = 0, & (x,t) \in Q, \\ \nabla \cdot D(x,t) = \nabla \cdot B(x,t) = 0, & (x,t) \in Q, \\ D(x,0) = d(x), \quad B(x,0) = b(x), & x \in \Omega, \\ \nu(x) \times E(x,t) = q(x,t), & (x,t) \in \Sigma, \end{cases} \tag{1.1}$$

with the constitutive relations

$$\begin{cases} D(x,t) = \varepsilon(x)E(x,t) + \zeta(x)H(x,t), & (x,t) \in Q, \\ B(x,t) = \zeta(x)E(x,t) + \mu(x)H(x,t), & (x,t) \in Q. \end{cases} \tag{1.2}$$

Here and henceforth $x = (x_1, x_2, x_3) \in \mathbb{R}^3$, $\partial_t = \frac{\partial}{\partial t}$, $\partial_k = \frac{\partial}{\partial x_k}$ for $k = 1, 2, 3$, $\nabla = (\partial_1, \partial_2, \partial_3)^{\mathrm{T}}$, Δ is the Laplacian in x, $Q = \Omega \times (-T, T)$, Ω is a bounded domain in $\mathbb{R}^3$ with C^2-boundary $\partial\Omega$, $\Sigma = \partial\Omega \times (-T, T)$ and $\nu(x) = (\nu_1(x), \nu_2(x), \nu_3(x))^{\mathrm{T}}$ is the outward unit normal vector to $\partial\Omega$ at x. In (1.1),

$$\begin{aligned} D(x,t) &= (D_1(x,t), D_2(x,t), D_3(x,t))^{\mathrm{T}} : && \text{the electric flux density,} \\ B(x,t) &= (B_1(x,t), B_2(x,t), B_3(x,t))^{\mathrm{T}} : && \text{the magnetic flux density,} \\ E(x,t) &= (E_1(x,t), E_2(x,t), E_3(x,t))^{\mathrm{T}} : && \text{the electric field,} \\ H(x,t) &= (H_1(x,t), H_2(x,t), H_3(x,t))^{\mathrm{T}} : && \text{the magnetic field.} \end{aligned}$$

$d(x)$, $b(x)$, $q(x,t)$ are given vector-valued functions and $\varepsilon(x)$, $\zeta(x)$, $\mu(x)$ are real-valued functions. Here and henceforth $\cdot^{\mathrm{T}}$ denotes the transposes of vectors or matrices under the consideration.

Our consideration is based on some physical background. In fact, there exist materials which can exhibit the magnetoelectric effect. For example, we can consider

*Received by the editors May 13, 2004; accepted for publication (in revised form) April 19, 2005; published electronically December 30, 2005. The author is supported by the Fujyu-kai (Tokyo, Japan) and the 21st century COE program at the Graduate School of Mathematical Sciences of the University of Tokyo.

http://www.siam.org/journals/sima/37-4/44366.html

†Graduate School of Mathematical Sciences, The University of Tokyo, 3-8-1 Komaba Meguro Tokyo 153, Japan (lism@ms.u-tokyo.ac.jp).

some magnetic crystals such as antiferromagnetic Cr_2O_3 and ferromagnetic $GaFeO_3$ (cf. [24, 21]). For details, we refer to [24, 5, 21, 22]. The constitutive relations for magnetoelectric media can be written in the following form (cf. [24, 5]):

$$\begin{cases} D = \overline{\varepsilon}E + \overline{\zeta}H, \\ B = \overline{\zeta}^{\mathrm{T}}E + \overline{\mu}H, \end{cases}$$

where the three 3×3 matrices $\overline{\varepsilon}$, $\overline{\mu}$, and $\overline{\zeta}$ are the permittivity tensor, the permeability tensor, and the Dzyaloshinskii magnetoelectric tensor, respectively. This paper is concerned with the bi-isotropic case, that is, $\overline{\varepsilon} = \varepsilon\mathbf{I}_3$, $\overline{\zeta} = \zeta\mathbf{I}_3$, and $\overline{\mu} = \mu\mathbf{I}_3$, where ε, ζ and μ are real-valued functions of x and $\mathbf{I}_3$ denotes the 3×3 unit matrix.

In this paper, we first consider the following.

Inverse problem: Let $\omega \subset \Omega$ satisfy $\partial\Omega \subset \partial\omega$ and $T > 0$ be suitably given. Then determine $\varepsilon(x)$, $\zeta(x)$, $\mu(x)$ for $x \in \Omega$ from the observation data

$$D(x,t), \quad B(x,t), \qquad (x,t) \in Q_\omega \equiv \omega \times (-T,T).$$

To the author's best knowledge, in the existing papers, inverse problems of determining all the coefficients for Maxwell's equations are mainly considered for the classical isotropic case. The purpose of this paper is to prove the Lipschitz stability for the inverse problem in a more general media. For other inverse problems for Maxwell's equations, we refer to Romanov [26], Romanov and Kabanikhin [27], Sun and Uhlmann [28], Yamamoto [29, 30], and Li and Yamamoto [23].

To state our main results, we introduce the notations. For the coefficients ε, ζ, μ in the constitutive relations (1.2), we set $\epsilon = (\varepsilon, \zeta, \mu)$. By $D[\epsilon;\Phi](\cdot)$, $B[\epsilon;\Phi](\cdot)$, $E[\epsilon;\Phi](\cdot)$, and $H[\epsilon;\Phi](\cdot)$, we denote the solution of (1.1) and (1.2) with the initial and boundary conditions $\Psi \equiv (d, b, q)$.

To guarantee the uniqueness in the inverse problem, we will use two sets of initial and boundary data: $\Phi(j) = \left(d^j, b^j, q^j\right)$, $j = 1, 2$. For the sake of convenience, we assume that d^j, b^j, and q^j $(j = 1, 2)$ are sufficiently smooth and satisfy sufficient compatibility conditions so that $D[\epsilon;\Phi]$, $B[\epsilon;\Phi] \in \left(W^{2,\infty}(Q)\right)^3$. Denote by $\mathbb{G}$ the 12×9 matrix

$$\begin{pmatrix} 0 & e_1\times d^1 & e_1\times b^1 & 0 & e_2\times d^1 & e_2\times b^1 & 0 & e_3\times d^1 & e_3\times b^1 \\ e_1\times d^1 & e_1\times b^1 & 0 & e_2\times d^1 & e_2\times b^1 & 0 & e_3\times d^1 & e_3\times b^1 & 0 \\ 0 & e_1\times d^2 & e_1\times b^2 & 0 & e_2\times d^2 & e_2\times b^2 & 0 & e_3\times d^2 & e_3\times b^2 \\ e_1\times d^2 & e_1\times b^2 & 0 & e_2\times d^2 & e_2\times b^2 & 0 & e_3\times d^2 & e_3\times b^2 & 0 \end{pmatrix},$$

where $e_1 = (1,0,0)^{\mathrm{T}}$, $e_2 = (0,1,0)^{\mathrm{T}}$, and $e_3 = (0,0,1)^{\mathrm{T}}$.

For $W = (w_1, w_2, \ldots, w_n)^{\mathrm{T}}$, we set $|W|^2 = \sum_{k=1}^n |w_k|^2$. Furthermore, $L^2(\Omega)$, $H^1(Q_\omega)$, etc. denote usual Sobolev spaces.

Let $\lambda = \inf_{x\in\overline{\Omega}} |x|$ and $\Lambda = \sup_{x\in\overline{\Omega}} |x|$. We assume that

$$0 < \Lambda^2 < 2\lambda^2. \tag{1.3}$$

We introduce an admissible set of unknown coefficients $\epsilon = (\varepsilon, \zeta, \mu)$, $\widetilde{\epsilon} = \left(\widetilde{\varepsilon}, \widetilde{\zeta}, \widetilde{\mu}\right)$:

$$\mathcal{U} = \mathcal{U}_{M,\theta_0,\theta_1,\epsilon_0} = \big\{(\varepsilon,\zeta,\mu) \in (C^2(\overline{\Omega}))^3 : \quad (\varepsilon,\zeta,\mu) = \epsilon_0 \text{ on } \partial\Omega;$$

$$\|\varepsilon\|_{C^2(\overline{\Omega})}, \|\zeta\|_{C^2(\overline{\Omega})}, \|\mu\|_{C^2(\overline{\Omega})}, \|\nabla(\varepsilon\mu-\zeta^2)\|_{(C(\overline{\Omega}))^3} < M; \varepsilon, \mu, \varepsilon\mu-\zeta^2 > \theta_1 \text{ on } \overline{\Omega};$$

$$\tfrac{x\cdot\nabla(\varepsilon\mu-\zeta^2)}{2(\epsilon\mu-\zeta^2)} > -\theta_0 \text{ on } \overline{\Omega}; \|D[\epsilon;\Phi(j)]\|_{(W^{2,\infty}(Q))^3}, \|B[\epsilon;\Phi(j)]\|_{(W^{2,\infty}(Q))^3} < M \text{ for } j=1,2\big\},$$

where the constants $M > 0$, $0 < \theta_0 < 1$, $\theta_1 > 0$ and a smooth vector-valued function $\epsilon_0 = (\varepsilon_0, \zeta_0, \mu_0)$ on $\partial\Omega$ are given. Furthermore, we take a positive constant β such that

$$0 < \beta < \frac{-\lambda\sqrt{\theta_1} + \sqrt{\lambda^2\theta_1 + 4\theta_1^2(1-\theta_0)}}{2M\theta_1}. \tag{1.4}$$

The following is our main result.

THEOREM 1.1 (stability). *We assume* (1.3), (1.4), *and*

$$\frac{\sqrt{\Lambda^2 - \lambda^2}}{\beta} < T. \tag{1.5}$$

Furthermore, we assume that there exists a constant $\theta_2 > 0$ *such that*

$$\text{the absolute value of one of the } 9 \times 9 \text{ minors of } \mathbb{G}(x) \ge \theta_2, \quad \text{for all } x \in \overline{\Omega}. \tag{1.6}$$

Then there exists a constant $K = K(\Omega, T, \Phi(1), \Phi(2), M, \theta_0, \theta_1, \theta_2, \epsilon_0) > 0$ *such that*

$$\|\widetilde{\varepsilon} - \varepsilon\|_{L^2(\Omega)} + \|\widetilde{\zeta} - \zeta\|_{L^2(\Omega)} + \|\widetilde{\mu} - \mu\|_{L^2(\Omega)} \le K\Theta \tag{1.7}$$

for all $\epsilon = (\varepsilon, \zeta, \mu) \in \mathcal{U}$ *and* $\widetilde{\epsilon} = (\widetilde{\varepsilon}, \widetilde{\zeta}, \widetilde{\mu}) \in \mathcal{U}$, *where*

$$\begin{aligned}\Theta = \sum_{j=1}^{2}\{&\|\partial_t(D[\epsilon;\Phi(j)] - D[\widetilde{\epsilon};\Phi(j)])\|_{(H^1(Q_\omega))^3} \\ &+ \|\partial_t(B[\epsilon;\Phi(j)] - B[\widetilde{\epsilon};\Phi(j)])\|_{(H^1(Q_\omega))^3}\}.\end{aligned} \tag{1.8}$$

We will provide a proof of Theorem 1.1 in section 3 by applying the argument on the basis of Carleman estimate and the energy conservation law. The method of applying Carleman estimate (i.e., a weighted L^2-estimate) to inverse problems was invented by Bukhgeim and Klibanov [3]. For developments of this method, we refer to Bukhgeim [2]; Imanuvilov, Isakov, and Yamamoto [9]; Imanuvilov and Yammamoto [10, 11, 12]; Isakov [14, 15]; Khaĭdarov [17, 18]; Klibanov [19]; and Yamamoto [30, 31]. In section 2, we will state a key Carleman estimate.

For similar inverse problems, we refer to [9] for a Lamé system and [12] for an acoustic equation. In [9, 12], the stability estimate is of Hölder's type. Due to the energy conservation law (3.2) in Lemma 3.1 (ii), we gain here Lipschitz stability estimate (1.7), which is one main achievement.

Furthermore, in Theorem 2.1 in [9], the H^5-norms of observation data are needed in the stability estimate. Here we need at most H^2-norms. There are two reasons. First, in the proof of Theorem 2.1 in [9], higher-order (the second, third, fourth) derivatives with respect to t are used as in [12]. Instead of using higher-order derivatives with respect to t, we use energy estimate (3.1) similarly to [10]. Second, the solution to the corresponding initial-value/boundary-value problem, has only three components in [9]. In this paper, there are in total six components of the vector-valued functions D and B. On the other hand, there are three unknown coefficients both in [9] and in this paper so that we can reduce the order of t-derivatives.

Remark 1.1. In Theorem 1.1, we have to assume a monotonicity condition about the wave speed $(\varepsilon\mu - \zeta^2)^{-(1/2)}$:

$$\frac{x \cdot \nabla(\varepsilon\mu - \zeta^2)}{2(\epsilon\mu - \zeta^2)} > -\theta_0 \text{ on } \overline{\Omega}. \tag{1.9}$$

We may be able to replace (1.9) by a weaker condition, but it is extremely difficult to search for sharpest conditions for deriving a Carleman estimate which is the main tool for our proof. Condition (1.9) is a sufficient condition for the pseudoconvexity, which is a sharp sufficient condition for a Carleman estimate (e.g., [6, 15]), and a similar condition is assumed in the existing references (cf. [4, 9, 12, 13, 15, 16, 17, 20]). In this paper, further searches will not be made for a more general condition than (1.9). In particular, if $\varepsilon\mu - \zeta^2$ is close to a constant function, then (1.9) holds true. Furthermore, we note that similar conditions to (1.3)–(1.5) are adopted in [9].

Remark 1.2. Condition (1.6) enables us to obtain sufficient information from the observation data in order to determine the unknown coefficients and we should a posteriori choose initial data meeting (1.6). It is a common feature in inverse problems with a finite number of measurements that such conditions should be satisfied by initial data (e.g., [9, 10, 12]). There exist initial data satisfying (1.6). For example, we take $d^1(x) = e_3$, $b^1(x) = d^2(x) = e_2$, $b^2(x) = e_1$ for $x \in \overline{\Omega}$. In fact, the 9×9 minor formed by rows 1, 2, 3, 4, 5, 9, 10, 11, and 12 satisfies (1.6) if we take $0 < \theta_2 < 1$.

Remark 1.3. By setting

$$\mathbb{A}_0 = \begin{pmatrix} \varepsilon\mathbf{I}_3 & \zeta\mathbf{I}_3 \\ \zeta\mathbf{I}_3 & \mu\mathbf{I}_3 \end{pmatrix}, \quad \mathbb{A}_k = \begin{pmatrix} 0 & A_k \\ -A_k & 0 \end{pmatrix}, \quad k = 1, 2, 3,$$

$$A_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and $U = (E_1, E_2, E_3, H_1, H_2, H_3)^{\mathrm{T}}$, Maxwell's equations with the constitutive relations (1.2) can be written as

$$\mathbb{A}_0 \partial_t U + \sum_{k=1}^{3} \mathbb{A}_k \partial_k U = 0.$$

It is obvious that $\mathbb{A}_0^{\mathrm{T}} = \mathbb{A}_0$ and $\mathbb{A}_k^{\mathrm{T}} = \mathbb{A}_k$ $(k = 1, 2, 3)$. Moreover, $\mathbb{A}_0$ is a 6×6 positive definite matrix if there exists a constant θ_1 such that ϵ, μ, $\epsilon\mu - \zeta^2 \geq \theta_1$. Therefore, for initial-value/boundary-value problem (1.1) and (1.2), we can refer to the results on symmetric hyperbolic equations.

This paper consists of two more sections. In section 2, we will show a Carleman estimate for Maxwell's equations in bi-isotropic media. In section 3, we will complete the proof of Theorem 1.1.

2. Carleman estimates for Maxwell's equations. Reducing Maxwell's equations to a weakly coupled system of hyperbolic equations and applying a Carleman estimate in the H^{-1}-norms, we will establish a Carleman estimate for Maxwell's equations in bi-isotropic media. For such methods, we can refer to [9, 30].

First, we show a Carleman estimate in the H^{-1}-norms for a second order hyperbolic equation which is proved in [9].

For β and λ, we define the functions $\varphi = \varphi(x, t)$ by

$$\varphi(x, t) = \mathrm{e}^{\varrho(|x|^2 - \beta^2 t^2 - \lambda^2)} \tag{2.1}$$

with some large $\varrho > 0$. For $(\varepsilon, \zeta, \mu) \in \mathcal{U}$, we set

$$\xi = \varepsilon\mu - \zeta^2 \quad \text{on } \overline{\Omega}. \tag{2.2}$$

By $(\varepsilon, \zeta, \mu) \in \mathcal{U}$ and (1.4), we can verify that $a = \xi^{1/2}$ satisfies the conditions $-\theta_0 < \frac{x\cdot\nabla a}{a}$ and $2|\nabla a|\lambda\beta + a^2\beta^2 < 1 - \theta_0$ on $\overline{\Omega}$ (cf. (2.2) and (2.3) in [9]). Furthermore, we note that our weight function φ coincides with that in [9] (cf. (4.3) in [9]).

Proposition 2.1 (see [9]). *Let $\varphi(x,t)$ be given by* (2.1) *and the domain Ω satisfy* (1.3). *We assume that $(\varepsilon, \zeta, \mu) \in \mathcal{U}$ and $0 < T < \frac{\lambda}{\beta}$, where β satisfies* (1.4). *Let $u \in H_0^2(Q)$ satisfy*

$$\xi(x)\left[\partial_t^2 u(x,t)\right] - \Delta u(x,t) = \widetilde{g} + \partial_t g_0 + \sum_{k=1}^{3} \partial_k g_k,\ (x,t) \in Q,$$

where ξ is defined by (2.2). *Then there is $K_1 > 0$ such that for all $s > K_1$,*

$$\int_Q s|u|^2 \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t \le K_1 \int_Q \left(\frac{1}{s^2}|\widetilde{g}|^2 + \sum_{k=0}^{3} |g_k|^2\right) \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t.$$

For the proof of Proposition 2.1, we refer to Theorem 3.2 and section 4 in [9] and section 3.4 in [15]. For similar results, we can refer to Proposition 2.1 in [12] which is derived from Theorem 1.1 in Imanuvilov [8]. Furthermore, we can refer to Ruiz [25]. For Carleman estimate in the H^{-1}-norms for parabolic equations, we refer to [11]. As for Carleman estimate in the L^2-norms, we refer to Hörmander [6, 7], Isakov [15, 16].

Now we state a Carleman estimate for Maxwell's equations, which is the main ingredient for the proof of Theorem 1.1. We will prove it by applying Proposition 2.1.

Theorem 2.2. *Let $\varphi(x,t)$ be given by* (2.1) *and the domain Ω satisfy* (1.3). *We assume that $\epsilon = (\varepsilon, \zeta, \mu) \in \mathcal{U}$ and $0 < T < \frac{\lambda}{\beta}$, where β satisfies* (1.4). *Let D, $B \in \left(H_0^1(Q)\right)^3$ satisfy*

$$\partial_t D - \nabla \times H = R_1, \quad \partial_t B + \nabla \times E = R_2, \tag{2.3}$$

$$D = \varepsilon E + \zeta H, \quad B = \zeta E + \mu H, \tag{2.4}$$

$$\nabla \cdot D = r_1, \quad \nabla \cdot B = r_2 \tag{2.5}$$

in Q. Then there exist $s_0 = s_0(\varrho) > 0$ and $K_2 = K_2(s_0, \beta, \varrho, M, \theta_0, \theta_1, \Omega, T) > 0$ such that for $s > s_0$

$$\begin{aligned} &s\int_Q \left(|D|^2 + |E|^2 + |B|^2 + |H|^2\right) \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t \\ &\quad \le K_2 \sum_{k=1}^{2} \int_Q \left(|r_k|^2 + |R_k|^2\right) \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t. \end{aligned} \tag{2.6}$$

Proof. By the density argument, it suffices to prove (2.6) for D, $B \in \left(C_0^\infty(Q)\right)^3$ satisfying (2.3)–(2.5).

By $\epsilon \in \mathcal{U}$, we have $\xi > \theta_1$ on $\overline{\Omega}$ where ξ is defined by (2.2). By (2.4) and direct calculations, we have

$$E = \gamma_1 D + \gamma_2 B, \quad H = \gamma_2 D + \gamma_3 B, \qquad \text{in } Q, \tag{2.7}$$

where

$$\gamma_1 = \frac{\mu}{\xi}, \quad \gamma_2 = -\frac{\zeta}{\xi}, \quad \gamma_3 = \frac{\varepsilon}{\xi} \qquad \text{on } \overline{\Omega}. \tag{2.8}$$

Moreover, by $\epsilon \in \mathcal{U}$, we have $\gamma_k \in C^2\left(\overline{\Omega}\right)$ for $k = 1, 2, 3$ and

$$\gamma_1\gamma_3 - \gamma_2^2 = \frac{1}{\xi} > 0 \qquad \text{on } \overline{\Omega}. \tag{2.9}$$

Differentiating the second equality in (2.7) with respect to t, in terms of (2.3), (2.7), and (2.9), we have

$$\begin{aligned}
\partial_t H &= \gamma_2 \left(\partial_t D\right) + \gamma_3 \left(\partial_t B\right) \\
&= \gamma_2 \left(\nabla \times H + R_1\right) + \gamma_3 \left(-\nabla \times E + R_2\right) \\
&= \gamma_2 \left[\nabla \times \left(\gamma_2 D + \gamma_3 B\right) + R_1\right] + \gamma_3 \left[-\nabla \times \left(\gamma_1 D + \gamma_2 B\right) + R_2\right] \\
&= \left(\gamma_2^2 - \gamma_1\gamma_3\right)\left(\nabla \times D\right) + \Upsilon(D, B) + \gamma_2 R_1 + \gamma_3 R_2 \\
&= -\frac{1}{\xi}\left(\nabla \times D\right) + \Upsilon(D, B) + \gamma_2 R_1 + \gamma_3 R_2 \qquad \text{in } Q,
\end{aligned}$$

where $\Upsilon(D, B) = \gamma_2 \left[(\nabla\gamma_2) \times D + (\nabla\gamma_3) \times B\right] + \gamma_3 \left[-\left(\nabla\gamma_1\right) \times D - (\nabla\gamma_2) \times B\right]$. For the fourth equality, we have used

$$\nabla \times (\alpha A) = \alpha\left(\nabla \times A\right) + (\nabla\alpha) \times A \qquad \text{on } \overline{\Omega} \tag{2.10}$$

for $\alpha \in C^1\left(\overline{\Omega}\right)$ and $A \in \left(C^1\left(\overline{\Omega}\right)\right)^3$. Therefore, differentiating the first equation of (2.3) with respect to t, we obtain

$$\begin{aligned}
\partial_t R_1 &= \partial_t^2 D - \nabla \times \left(\partial_t H\right) = \partial_t^2 D - \nabla \times \left[-\frac{1}{\xi}(\nabla \times D) + \Upsilon(D, B) + \gamma_2 R_1 + \gamma_3 R_2\right] \\
&= \partial_t^2 D + \frac{1}{\xi}[\nabla \times (\nabla \times D)] + \left[\nabla\left(\frac{1}{\xi}\right)\right] \times (\nabla \times D) - \nabla \times \left[\Upsilon(D, B) + \gamma_2 R_1 + \gamma_3 R_2\right]
\end{aligned}$$

in Q. In the third equality, we have used (2.10). Then by $\nabla \times (\nabla \times D) = -\Delta D + \nabla\left(\nabla \cdot D\right)$ and the first equation of (2.5), we have

$$\partial_t^2 D - \frac{1}{\xi}\Delta D = -\left[\nabla\left(\frac{1}{\xi}\right)\right] \times (\nabla \times D) - \frac{1}{\xi}\nabla r_1 + \nabla \times \left[\Upsilon(D, B) + \gamma_2 R_1 + \gamma_3 R_2\right] + \partial_t R_1$$

in Q. Hence we can apply Proposition 2.1 and obtain that there exist $s_1 = s_1(\varrho) > 0$ and $C_1 > 0$ such that for $s > s_1$

$$s\int_Q |D|^2 \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t \le C_1 \int_Q \left(|r_1|^2 + |D|^2 + |B|^2 + |R_1|^2 + |R_2|^2\right) \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t. \tag{2.11}$$

Here and henceforth, $C_k (k = 1, 2, \dots)$ denote generic positive constants which may depend on $s_l (l = 0, 1, 2, \dots)$, ϱ, M, θ_0, θ_1, ϵ_0, λ, Λ, β, Ω, T, Σ, χ_1, χ_2, δ, η, t_1, t_2, $\left\|d^1\right\|_{\left(C^1(\overline{\Omega})\right)^3}$, $\left\|d^2\right\|_{\left(C^1(\overline{\Omega})\right)^3}$, $\left\|b^1\right\|_{\left(C^1(\overline{\Omega})\right)^3}$, $\left\|b^2\right\|_{\left(C^1(\overline{\Omega})\right)^3}$, but are independent of s and ϑ.

Similarly to (2.11), we can obtain that there exist $s_2 = s_2(\varrho) > 0$ and $C_2 > 0$ such that for $s > s_2$

$$s\int_Q |B|^2 \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t \le C_2 \int_Q \left(|r_2|^2 + |D|^2 + |B|^2 + |R_1|^2 + |R_2|^2\right) \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t. \tag{2.12}$$

By (2.7), (2.11), and (2.12), we see that there exist $s_0 = s_0(\varrho) > 0$ and $C_3, C_4 > 0$ such that for $s > s_0$

$$s\int_Q \left(|D|^2+|E|^2+|B|^2+|H|^2\right) \mathrm{e}^{2s\varphi}\mathrm{d}x\mathrm{d}t \le C_3 s\int_Q \left(|D|^2+|B|^2\right)\mathrm{e}^{2s\varphi}\mathrm{d}x\mathrm{d}t$$
$$\le C_4\int_Q \left(|r_1|^2+|r_2|^2+|R_1|^2+|R_2|^2\right)\mathrm{e}^{2s\varphi}\mathrm{d}x\mathrm{d}t.$$

The proof of Theorem 2.2 is finished. □

The following is a Carleman estimate for a first-order differential equation.

PROPOSITION 2.3 (see [9]). *Let $\varphi(x,t)$ be given by* (2.1). *Then there exists $K_3 > 0$ such that for $s > K_3$*

$$\int_\Omega s|w|^2\mathrm{e}^{2s\varphi(x,0)}\mathrm{d}x \le K_3\sum_{j=1}^{3}\int_\Omega |\partial_j w|^2\mathrm{e}^{2s\varphi(x,0)}\mathrm{d}x \quad \textit{for all } w\in C_0^1\left(\overline{\Omega}\right). \tag{2.13}$$

We can prove Proposition 2.3 directly by integral by parts (e.g., Lemma 3.6 in [9]).

3. Proof of Theorem 1.1. We divide the proof into five steps.

Step 1. First we show a lemma, the first part of which is proved by the argument in [10] and Carleman estimate (2.6) for Maxwell's equations in bi-isotropic media.

LEMMA 3.1. *Let $\varphi(x,t)$ be given by* (2.1) *and the domain Ω satisfy* (1.3). *We assume that $\epsilon = (\varepsilon,\zeta,\mu)\in\mathcal{U}$ and $0 < T < \frac{\lambda}{\beta}$, where β satisfies* (1.4). *Let D, $B\in\left(H^1(Q)\right)^3$ satisfy* (2.3)–(2.5) *and $D = B = 0$ on Σ. Then we have the following.*

(i) *There exist $s_3 = s_3(\varrho) > 0$ and $K_4 = K_4(s_3,\beta,\varrho,M,\theta_0,\theta_1,\Omega,T) > 0$ such that*

$$\int_\Omega \left[|D(\cdot,0)|^2+|B(\cdot,0)|^2\right]\mathrm{e}^{2s\varphi(\cdot,0)}\mathrm{d}x \le K_4\sum_{k=1}^{2}\int_Q\left(|r_k|^2+|R_k|^2\right)\mathrm{e}^{2s\varphi}\mathrm{d}x\mathrm{d}t \tag{3.1}$$

for $s > s_3$, provided that D, $B\in\left(H_0^1(Q)\right)^3$.

(ii) *There exists $K_5 = K_5(M,\theta_0,\theta_1,\Omega,T) > 0$ such that*

$$\begin{aligned}&\int_\Omega \left[|D(\cdot,t_2)|^2+|B(\cdot,t_2)|^2\right]\mathrm{d}x\\ &\quad\le K_5\left\{\int_\Omega\left[|D(\cdot,t_1)|^2+|B(\cdot,t_1)|^2\right]\mathrm{d}x+\int_{Q_\star}\left(|R_1|^2+|R_2|^2\right)\mathrm{d}x\mathrm{d}t\right\}\end{aligned} \tag{3.2}$$

for $-T\le t_1,t_2\le T$, where $Q_\star = \Omega\times(\min\{t_1,t_2\},\max\{t_1,t_2\})\subseteq Q$.

We note that in (ii) D, B need not vanish on $\Omega\times\{\pm T\}$.

Proof. It is obvious that (3.2) holds for $-T\le t_1 = t_2\le T$. In the following, we assume that $t_1\ne t_2$ and $t_1,t_2\in[-T,T]$. Let

$$\iota = \begin{cases}1, & t_2 > t_1,\\ -1, & t_2 < t_1,\end{cases} \tag{3.3}$$

and ξ, γ_1, γ_2, γ_3 be defined by (2.2) and (2.8). As in the proof of Theorem 2.2, by $\epsilon\in\mathcal{U}$ and (2.4), we have $\gamma_k\in C^2\left(\overline{\Omega}\right)$ for $k = 1,2,3$, (2.7) and (2.9). By D, $B\in\left(H^1(Q)\right)^3$ and (2.7), we have E, $H\in\left(H^1(Q)\right)^3$. Furthermore, by $D = B = 0$ on Σ and (2.7), we have $E = H = 0$ on Σ.

By (2.3) and (2.7), it is easy to see that

$$(\partial_t D)\cdot(\gamma_1 D+\gamma_2 B)-(\nabla\times H)\cdot E=R_1\cdot(\gamma_1 D+\gamma_2 B)\qquad \text{in } Q, \tag{3.4}$$

$$(\partial_t B)\cdot(\gamma_2 D+\gamma_3 B)+(\nabla\times E)\cdot H=R_2\cdot(\gamma_2 D+\gamma_3 B)\qquad \text{in } Q. \tag{3.5}$$

Adding (3.4) and (3.5), we have

$$\partial_t\mathcal{B}+\nabla\cdot(E\times H)=R_1\cdot(\gamma_1 D+\gamma_2 B)+R_2\cdot(\gamma_2 D+\gamma_3 B) \tag{3.6}$$

in Q, where we set

$$\mathcal{B}\equiv\frac{1}{2}\left[\gamma_1|D|^2+2\gamma_2\,(D\cdot B)+\gamma_3|B|^2\right]. \tag{3.7}$$

Here we have used $(\nabla\times E)\cdot H-(\nabla\times H)\cdot E=\nabla\cdot(E\times H)$ in Q. Therefore, for $s\ge 0$, we have

$$I_l=I_r, \tag{3.8}$$

where

$$I_l\equiv\int_{t_1}^{t_2}\int_\Omega\left[\partial_t\mathcal{B}+\nabla\cdot(E\times H)\right]\mathrm{e}^{2s\varphi-\iota t}\mathrm{d}x\mathrm{d}t,$$

$$I_r\equiv\int_{t_1}^{t_2}\int_\Omega\left[R_1\cdot(\gamma_1 D+\gamma_2 B)+R_2\cdot(\gamma_2 D+\gamma_3 B)\right]\mathrm{e}^{2s\varphi-\iota t}\mathrm{d}x\mathrm{d}t.$$

Integrating by parts and using $E=H=0$ on Σ, we have

$$\begin{aligned}
I_l&=\int_\Omega\mathcal{B}\left(\cdot,t_2\right)\mathrm{e}^{2s\varphi(\cdot,t_2)-\iota t_2}\mathrm{d}x-\int_\Omega\mathcal{B}\left(\cdot,t_1\right)\mathrm{e}^{2s\varphi(\cdot,t_1)-\iota t_1}\mathrm{d}x\\
&\quad-\int_{t_1}^{t_2}\int_\Omega\left\{-\iota\mathcal{B}+2s\left[(\partial_t\varphi)\,\mathcal{B}+(E\times H)\cdot(\nabla\varphi)\right]\right\}\mathrm{e}^{2s\varphi-\iota t}\mathrm{d}x\mathrm{d}t\\
&\ge\int_\Omega\mathcal{B}\left(\cdot,t_2\right)\mathrm{e}^{2s\varphi(\cdot,t_2)-\iota t_2}\mathrm{d}x-\int_\Omega\mathcal{B}\left(\cdot,t_1\right)\mathrm{e}^{2s\varphi(\cdot,t_1)-\iota t_1}\mathrm{d}x\\
&\quad+\int_{Q_\star}\mathcal{B}\mathrm{e}^{2s\varphi-\iota t}\mathrm{d}x\mathrm{d}t-C_5\mathrm{e}^T s\int_{Q_\star}\left(|\mathcal{B}|+|E|^2+|H|^2\right)\mathrm{e}^{2s\varphi}\mathrm{d}x\mathrm{d}t\quad\text{for } s\ge 0.
\end{aligned} \tag{3.9}$$

For the inequality in (3.9), we have used (3.3) and definition (2.1) of φ. By (2.9) and (3.7), there exist C_6, $C_7>0$ such that

$$C_6\left(|D|^2+|B|^2\right)\le\mathcal{B}\le C_7\left(|D|^2+|B|^2\right)\quad\text{in } Q. \tag{3.10}$$

Therefore, by (3.9) and (3.10), we have

$$\begin{aligned}
I_l&\ge C_6\mathrm{e}^{-T}\int_\Omega\left[|D(\cdot,t_2)|^2+|B(\cdot,t_2)|^2\right]\mathrm{e}^{2s\varphi(\cdot,t_2)}\mathrm{d}x\\
&\quad-C_7\mathrm{e}^T\int_\Omega\left[|D(\cdot,t_1)|^2+|B(\cdot,t_1)|^2\right]\mathrm{e}^{2s\varphi(\cdot,t_1)}\mathrm{d}x\\
&\quad+C_6\mathrm{e}^{-T}\int_{Q_\star}\left(|D|^2+|B|^2\right)\mathrm{e}^{2s\varphi}\mathrm{d}x\mathrm{d}t\\
&\quad-C_8\mathrm{e}^T s\int_{Q_\star}\left(|D|^2+|E|^2+|B|^2+|H|^2\right)\mathrm{e}^{2s\varphi}\mathrm{d}x\mathrm{d}t\quad\text{for } s\ge 0.
\end{aligned} \tag{3.11}$$

By the inequality $2|ab| \le \vartheta a^2 + b^2/\vartheta$ for $\vartheta > 0$, we obtain

$$I_r \le C_9 \mathrm{e}^T \int_{Q_\star} \left[\vartheta \left(|D|^2 + |B|^2 \right) + \frac{1}{\vartheta} \left(|R_1|^2 + |R_2|^2 \right) \right] \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t \tag{3.12}$$

for $\vartheta > 0$ and $s \ge 0$. Hence, by (3.8), (3.11), and (3.12), choosing $\vartheta > 0$ sufficiently small and fixing it, we have

$$\begin{aligned} &\int_\Omega \left[|D(\cdot, t_2)|^2 + |B(\cdot, t_2)|^2 \right] \mathrm{e}^{2s\varphi(\cdot, t_2)} \mathrm{d}x \\ &\quad \le C_{10} \left\{ \int_\Omega \left[|D(\cdot, t_1)|^2 + |B(\cdot, t_1)|^2 \right] \mathrm{e}^{2s\varphi(\cdot, t_1)} \mathrm{d}x \right. \\ &\qquad + \int_{Q_\star} \left(|R_1|^2 + |R_2|^2 \right) \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t \\ &\qquad \left. + s \int_{Q_\star} \left(|D|^2 + |E|^2 + |B|^2 + |H|^2 \right) \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t \right\} \end{aligned} \tag{3.13}$$

for $s \ge 0$. Taking $s = 0$ in (3.13), we obtain (3.2).

Next, we will complete the proof of (3.1). Since $D, B \in \left(H_0^1(Q) \right)^3$, we can apply Theorem 2.2, so that (2.6) holds for sufficiently large $s > 0$. Therefore, taking $t_2 = 0$, $t_1 = T$ in (3.13), and using $Q_\star \subseteq Q$ and (2.6), we see that there exist $s_3 = s_3(\varrho) > 0$ and $K_4 = K_4(s_3, \beta, \varrho, M, \theta_0, \theta_1, \Omega, T) > 0$ such that (3.1) holds for $s > s_3$.

The proof of Lemma 3.1 is complete. $\square$

Step 2. As in [9], by (1.3) and (1.5), we can assume that $0 < T < \frac{\lambda}{\beta}$, where β satisfies (1.4).

By $\epsilon, \widetilde{\epsilon} \in \mathcal{U}$, we have $\xi \equiv \varepsilon\mu - \zeta^2 > \theta_1$, $\widetilde{\xi} \equiv \widetilde{\varepsilon}\widetilde{\mu} - \widetilde{\zeta}^2 > \theta_1$ on $\overline{\Omega}$. By the constitutive relations (1.2), we have

$$\begin{aligned} E[\epsilon; \Phi(j)] &= \gamma_1 D[\epsilon; \Phi(j)] + \gamma_2 B[\epsilon; \Phi(j)], \\ H[\epsilon; \Phi(j)] &= \gamma_2 D[\epsilon; \Phi(j)] + \gamma_3 B[\epsilon; \Phi(j)], \\ E[\widetilde{\epsilon}; \Phi(j)] &= \widetilde{\gamma}_1 D[\widetilde{\epsilon}; \Phi(j)] + \widetilde{\gamma}_2 B[\widetilde{\epsilon}; \Phi(j)], \\ H[\widetilde{\epsilon}; \Phi(j)] &= \widetilde{\gamma}_2 D[\widetilde{\epsilon}; \Phi(j)] + \widetilde{\gamma}_3 B[\widetilde{\epsilon}; \Phi(j)] \quad \text{in } Q, \end{aligned} \tag{3.14}$$

where $j = 1, 2$ and

$$\gamma_1 = \frac{\mu}{\xi}, \quad \gamma_2 = -\frac{\zeta}{\xi}, \quad \gamma_3 = \frac{\varepsilon}{\xi} \qquad \text{on } \overline{\Omega}, \tag{3.15}$$

$$\widetilde{\gamma}_1 = \frac{\widetilde{\mu}}{\widetilde{\xi}}, \quad \widetilde{\gamma}_2 = -\frac{\widetilde{\zeta}}{\widetilde{\xi}}, \quad \widetilde{\gamma}_3 = \frac{\widetilde{\varepsilon}}{\widetilde{\xi}} \qquad \text{on } \overline{\Omega}. \tag{3.16}$$

Furthermore, by $\epsilon, \widetilde{\epsilon} \in \mathcal{U}$, we have γ_k, $\widetilde{\gamma}_k \in C^2(\overline{\Omega})$ for $k = 1, 2, 3$ and

$$\gamma_1\gamma_3 - \gamma_2^2 = \frac{1}{\xi} > 0, \quad \widetilde{\gamma}_1\widetilde{\gamma}_3 - \widetilde{\gamma}_2^2 = \frac{1}{\widetilde{\xi}} > 0, \quad \text{on } \overline{\Omega}. \tag{3.17}$$

We set

$$Y(\cdot; j) = \left\{ \partial_t D[\epsilon; \Phi(j)] - \partial_t D[\widetilde{\epsilon}; \Phi(j)] \right\}(\cdot) \quad \text{in } Q, \tag{3.18}$$

$$Z(\cdot; j) = \left\{ \partial_t B[\epsilon; \Phi(j)] - \partial_t B[\widetilde{\epsilon}; \Phi(j)] \right\}(\cdot) \quad \text{in } Q, \tag{3.19}$$

$$U(\cdot;j) = \gamma_1 Y(\cdot;j) + \gamma_2 Z(\cdot;j), \quad V(\cdot;j) = \gamma_2 Y(\cdot;j) + \gamma_3 Z(\cdot;j) \quad \text{in } Q, \tag{3.20}$$

$$R_3(\cdot;j) = \{\partial_t D[\widetilde{\epsilon};\Phi(j)]\}(\cdot), \quad R_4(\cdot;j) = \{\partial_t B[\widetilde{\epsilon};\Phi(j)]\}(\cdot) \quad \text{in } Q, \tag{3.21}$$

$$f_k = \widetilde{\gamma}_k - \gamma_k \quad \text{on } \overline{\Omega}, \quad k = 1,2,3, \tag{3.22}$$

$$\Psi_1(\cdot;j) = \nabla \times [f_1 R_3(\cdot;j) + f_2 R_4(\cdot;j)] \quad \text{in } Q, \tag{3.23}$$

$$\Psi_2(\cdot;j) = \nabla \times [f_2 R_3(\cdot;j) + f_3 R_4(\cdot;j)] \quad \text{in } Q, \quad j = 1,2,3. \tag{3.24}$$

Then we have $Y(\cdot;j)$, $Z(\cdot;j)$, $U(\cdot;j)$, $V(\cdot;j)$, $R_3(\cdot;j)$, $R_4(\cdot;j) \in \left(W^{1,\infty}(Q)\right)^3$,

$$\partial_t Y(\cdot;j) - \nabla \times V(\cdot;j) = -\Psi_2(\cdot;j), \quad \partial_t Z(\cdot;j) + \nabla \times U(\cdot;j) = \Psi_1(\cdot;j), \tag{3.25}$$

$$Y(\cdot;j) = \varepsilon U(\cdot;j) + \zeta V(\cdot;j), \quad Z(\cdot;j) = \zeta U(\cdot;j) + \mu V(\cdot;j), \tag{3.26}$$

$$\nabla \cdot Y(\cdot;j) = \nabla \cdot Z(\cdot;j) = 0 \quad \text{in } Q. \tag{3.27}$$

In fact, by (3.14) and (3.22), we have

$$\begin{aligned} E[\epsilon;\Phi(j)] - E[\widetilde{\epsilon};\Phi(j)] &= \gamma_1 \{D[\epsilon;\Phi(j)] - D[\widetilde{\epsilon};\Phi(j)]\} \\ &+ \gamma_2 \{B[\epsilon;\Phi(j)] - B[\widetilde{\epsilon};\Phi(j)]\} - f_1 D[\widetilde{\epsilon};\Phi(j)] - f_2 B[\widetilde{\epsilon};\Phi(j)] \quad \text{in } Q, \end{aligned} \tag{3.28}$$

$$\begin{aligned} H[\epsilon;\Phi(j)] - H[\widetilde{\epsilon};\Phi(j)] &= \gamma_2 \{D[\epsilon;\Phi(j)] - D[\widetilde{\epsilon};\Phi(j)]\} \\ &+ \gamma_3 \{B[\epsilon;\Phi(j)] - B[\widetilde{\epsilon};\Phi(j)]\} - f_2 D[\widetilde{\epsilon};\Phi(j)] - f_3 B[\widetilde{\epsilon};\Phi(j)] \quad \text{in } Q. \end{aligned} \tag{3.29}$$

Differentiating (3.28) and (3.29) with respect to t, and using (3.18)–(3.21) and t-independence of γ_1, γ_2, γ_3, we have

$$\{\partial_t E[\epsilon;\Phi(j)] - \partial_t E[\widetilde{\epsilon};\Phi(j)]\}(\cdot) = U(\cdot;j) - f_1 R_3(\cdot;j) - f_2 R_4(\cdot;j) \quad \text{in } Q, \tag{3.30}$$

$$\{\partial_t H[\epsilon;\Phi(j)] - \partial_t H[\widetilde{\epsilon};\Phi(j)]\}(\cdot) = V(\cdot;j) - f_2 R_3(\cdot;j) - f_3 R_4(\cdot;j) \quad \text{in } Q. \tag{3.31}$$

By (1.1), (3.18), and (3.19), we have

$$Y(\cdot;j) = \{\nabla \times H[\epsilon;\Phi(j)] - \nabla \times H[\widetilde{\epsilon};\Phi(j)]\}(\cdot) \quad \text{in } Q, \tag{3.32}$$

$$Z(\cdot;j) = -\{\nabla \times E[\epsilon;\Phi(j)] - \nabla \times E[\widetilde{\epsilon};\Phi(j)]\}(\cdot) \quad \text{in } Q. \tag{3.33}$$

Differentiating (3.32) and (3.33) with respect to t and using (3.23) and (3.24), and (3.30) and (3.31), we can obtain (3.25). By definition (3.15) of γ_1, γ_2, γ_3 and (3.20), direct calculations yield (3.26). Moreover, by (1.1), (3.18), and (3.19), we obtain (3.27). Furthermore, we note that $\|R_3\|_{(W^{1,\infty}(Q))^3}$, $\|R_4\|_{(W^{1,\infty}(Q))^3} \le M$ by $\widetilde{\epsilon} \in \mathcal{U}$.

By $\epsilon, \widetilde{\epsilon} \in \mathcal{U}$, and definitions (3.15), (3.16), and (3.22), we have $f_k \in C_0^1(\overline{\Omega})$ $(k = 1,2,3)$. Therefore, we can apply Proposition 2.3 to f_k $(k = 1,2,3)$. As a result, we have

$$\sum_{k=1}^{3} \int_\Omega |f_k|^2 \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x \le \frac{C_{11}}{s} \sum_{k=1}^{3} \int_\Omega |\nabla f_k|^2 \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x \tag{3.34}$$

for all sufficiently large $s > 0$.

By (1.5) and definition (2.1) of φ, we have

$$\varphi(x,0) \geq 1 \quad \text{and} \quad 0 < \varphi(x,-T) = \varphi(x,T) < 1, \quad x \in \overline{\Omega}. \tag{3.35}$$

Therefore, for any given small $\eta \in (0, 1 - \sup_{x\in\overline{\Omega}} \varphi(x,T))$, we can choose a sufficiently small $\delta = \delta(\eta) > 0$, such that

$$\varphi(x,t) \leq 1 - \eta, \quad (x,t) \in \overline{\Omega} \times ([-T, -T+2\delta] \cup [T-2\delta, T]). \tag{3.36}$$

In order to apply Lemma 3.1, we introduce two cut-off functions χ_1 and χ_2 satisfying $\chi_1 \in C_0^\infty(\Omega)$, $\chi_2 \in C^\infty(\mathbb{R})$, $0 \leq \chi_1(x) \leq 1$ for $x \in \overline{\Omega}$, $0 \leq \chi_2(t) \leq 1$ for $t \in \mathbb{R}$, $\chi_1(x) = 1$ for $x \in \overline{\Omega \backslash \omega}$ and

$$\chi_2(t) = \begin{cases} 0, & t \in [-T, -T+\delta] \cup [T-\delta, T], \\ 1, & t \in [-T+2\delta, T-2\delta]. \end{cases}$$

Step 3. We set $Y_1(\cdot;j) = \chi_1 Y(\cdot;j) \in \left(W^{1,\infty}(Q)\right)^3$, $Z_1(\cdot;j) = \chi_1 Z(\cdot;j) \in \left(W^{1,\infty}(Q)\right)^3$, $U_1(\cdot;j) = \chi_1 U(\cdot;j) \in \left(W^{1,\infty}(Q)\right)^3$, $V_1(\cdot;j) = \chi_1 V(\cdot;j) \in \left(W^{1,\infty}(Q)\right)^3$. Then, by definition of χ_1, (2.10) and (3.25)–(3.27), we have

$$\partial_t Y_1(\cdot;j) - \nabla \times V_1(\cdot;j) = -\chi_1 \Psi_2(\cdot;j) - (\nabla\chi_1) \times V(\cdot;j) \quad \text{in } Q, \tag{3.37}$$

$$\partial_t Z_1(\cdot;j) + \nabla \times U_1(\cdot;j) = \chi_1 \Psi_1(\cdot;j) + (\nabla\chi_1) \times U(\cdot;j) \quad \text{in } Q, \tag{3.38}$$

$$Y_1(\cdot;j) = \varepsilon U_1(\cdot;j) + \zeta V_1(\cdot;j), \quad Z_1(\cdot;j) = \zeta U_1(\cdot;j) + \mu V_1(\cdot;j) \quad \text{in } Q, \tag{3.39}$$

$$\nabla \cdot Y_1(\cdot;j) = (\nabla\chi_1) \cdot Y(\cdot;j), \quad \nabla \cdot Z_1(\cdot;j) = (\nabla\chi_1) \cdot Z(\cdot;j) \quad \text{in } Q, \tag{3.40}$$

$$Y_1(\cdot;j) = Z_1(\cdot;j) = 0 \quad \text{on } \Sigma. \tag{3.41}$$

Therefore, by Lemma 3.1 (ii) and $(\nabla\chi_1)(x) = 0$ for $x \in \Omega \setminus \omega$, we have

$$\begin{aligned} w(t_2;j) \leq c_{12} \Big\{ & w(t_1;j) + \int_Q \Big[\left| -\chi_1 \Psi_2(\cdot;j) - (\nabla\chi_1) \times V(\cdot;j) \right|^2 \\ & + \left| \chi_1 \Psi_1(\cdot;j) + (\nabla\chi_1) \times U(\cdot;j) \right|^2 \Big] \mathrm{d}x\mathrm{d}t \Big\} \\ \leq c_{13} \Big\{ & w(t_1;j) + \int_Q \Big[|\Psi_1(\cdot;j)|^2 + |\Psi_2(\cdot;j)|^2 \Big] \mathrm{d}x\mathrm{d}t \\ & + \int_{Q_\omega} |\nabla\chi_1|^2 \Big[|U(\cdot;j)|^2 + |V(\cdot;j)|^2 \Big] \mathrm{d}x\mathrm{d}t \Big\} \end{aligned} \tag{3.42}$$

for $-T \leq t_1, t_2 \leq T$, where

$$w(t;j) = \int_\Omega \Big[|Y_1(\cdot,t;j)|^2 + |Z_1(\cdot,t;j)|^2 \Big] \mathrm{d}x, \quad t \in [-T,T]. \tag{3.43}$$

By (3.23) and (3.24), $\|R_3(\cdot;j)\|_{(W^{1,\infty}(Q))^3}, \|R_4(\cdot;j)\|_{(W^{1,\infty}(Q))^3} \leq M$, and t-independency of f_k, we have

$$\begin{aligned} \int_Q \Big[|\Psi_1(\cdot;j)|^2 + |\Psi_2(\cdot;j)|^2 \Big] \mathrm{d}x\mathrm{d}t &\leq C_{14} \sum_{k=1}^3 \int_Q \left(|f_k|^2 + |\nabla f_k|^2 \right) \mathrm{d}x\mathrm{d}t \\ &\leq C_{15} \sum_{k=1}^3 \int_\Omega \left(|f_k|^2 + |\nabla f_k|^2 \right) \mathrm{d}x. \end{aligned} \tag{3.44}$$

By definition (1.8) of Θ and (3.18)–(3.20), we have

$$\begin{aligned}&\int_{Q_\omega} |\nabla\chi_1|^2 \left[|U(\cdot;j)|^2 + |V(\cdot;j)|^2\right] \mathrm{d}x\mathrm{d}t \\ &\quad\le C_{16} \int_{Q_\omega} \left[|Y(\cdot;j)|^2 + |Z(\cdot;j)|^2\right] \mathrm{d}x\mathrm{d}t \le C_{17}\Theta^2.\end{aligned} \tag{3.45}$$

Hence, it follows from (3.42), (3.44), and (3.45) that

$$w(t_2;j) \le c_{18}\left[w(t_1;j) + \sum_{k=1}^{3}\int_\Omega \left(|f_k|^2 + |\nabla f_k|^2\right)\mathrm{d}x + \Theta^2\right] \tag{3.46}$$

for $-T \le t_1, t_2 \le T$.

Step 4. We set $Y_2(\cdot;j) = \chi_2 Y_1(\cdot;j)$, $Z_2(\cdot;j) = \chi_2 Z_1(\cdot;j)$, $U_2(\cdot;j) = \chi_2 U_1(\cdot;j)$, $V_2(\cdot;j) = \chi_2 V_1(\cdot;j)$. Then, by definitions of χ_1, χ_2, and (3.37)–(3.41), we have $Y_2(\cdot;j)$, $Z_2(\cdot;j)$, $U_2(\cdot;j)$, $V_2(\cdot;j) \in \left(H_0^1(Q)\right)^3$,

$$\begin{gathered}\partial_t Y_2(\cdot;j) - \nabla\times V_2(\cdot;j) = -\chi_1\chi_2\Psi_2(\cdot;j) - \chi_2\left(\nabla\chi_1\right)\times V(\cdot;j) + \left(\partial_t\chi_2\right)Y_1(\cdot;j),\\ \partial_t Z_2(\cdot;j) + \nabla\times U_2(\cdot;j) = \chi_1\chi_2\Psi_1(\cdot;j) + \chi_2\left(\nabla\chi_1\right)\times U(\cdot;j) + \left(\partial_t\chi_2\right)Z_1(\cdot;j),\\ Y_2(\cdot;j) = \varepsilon U_2(\cdot;j) + \zeta V_2(\cdot;j), \quad Z_2(\cdot;j) = \zeta U_2(\cdot;j) + \mu V_2(\cdot;j),\\ \nabla\cdot Y_2(\cdot;j) = \chi_2\left(\nabla\chi_1\right)\cdot Y(\cdot;j), \quad \nabla\cdot Z_2(\cdot;j) = \chi_2\left(\nabla\chi_1\right)\cdot Z(\cdot;j) \quad \text{in } Q.\end{gathered}$$

Therefore, we can apply Lemma 3.1 (i) so that, by $\left(\nabla\chi_1\right)(x) = 0$ for $x \in \Omega\setminus\omega$ and $\left(\partial_t\chi_2\right)(t) = 0$ for $t \in (-T, -T+\delta)\cup(-T+2\delta, T-2\delta)\cup(T-\delta, T)$,

$$\begin{aligned}&\int_\Omega \left[|Y_2(\cdot,0;j)|^2 + |Z_2(\cdot,0;j)|^2\right]\mathrm{e}^{2s\varphi(\cdot,0)}\mathrm{d}x\\ &\le C_{19}\int_Q \Big[\left|-\chi_1\chi_2\Psi_2(\cdot;j) - \chi_2\left(\nabla\chi_1\right)\times V(\cdot;j) + \left(\partial_t\chi_2\right)Y_1(\cdot;j)\right|^2\\ &\qquad + \left|\chi_1\chi_2\Psi_1(\cdot;j) + \chi_2\left(\nabla\chi_1\right)\times U(\cdot;j) + \left(\partial_t\chi_2\right)Z_1(\cdot;j)\right|^2\\ &\qquad + \left|\chi_2\left(\nabla\chi_1\right)\cdot Y(\cdot;j)\right|^2 + \left|\chi_2\left(\nabla\chi_1\right)\cdot Z(\cdot;j)\right|^2\Big]\mathrm{e}^{2s\varphi}\mathrm{d}x\mathrm{d}t\\ &\le C_{20}\Bigg\{\int_Q \left[|\Psi_1(\cdot;j)|^2 + |\Psi_2(\cdot;j)|^2\right]\mathrm{e}^{2s\varphi}\mathrm{d}x\mathrm{d}t\\ &\qquad + \int_{Q_\omega}|\nabla\chi_1|^2\left[|U(\cdot;j)|^2 + |V(\cdot;j)|^2 + |Y(\cdot;j)|^2 + |Z(\cdot;j)|^2\right]\mathrm{e}^{2s\varphi}\mathrm{d}x\mathrm{d}t\\ &\qquad + \left(\int_{-T+\delta}^{-T+2\delta} + \int_{T-2\delta}^{T-\delta}\right)\int_\Omega \left(\partial_t\chi_2\right)^2\left[|Y_1(\cdot;j)|^2 + |Z_1(\cdot;j)|^2\right]\mathrm{e}^{2s\varphi}\mathrm{d}x\mathrm{d}t\Bigg\}\end{aligned} \tag{3.47}$$

for all large $s > 0$. By (2.1) and the definition of λ, we have

$$\varphi(x,t) - \varphi(x,0) = \mathrm{e}^{\varrho\left(|x|^2-\lambda^2\right)}\left(\mathrm{e}^{-\varrho\beta^2t^2} - 1\right) \le \mathrm{e}^{-\varrho\beta^2t^2} - 1 \le 0 \quad \text{for } x \in \overline{\Omega}.$$

Therefore, by (3.23) and (3.24) and $\|R_3(\cdot;j)\|_{(W^{1,\infty}(Q))^3}$, $\|R_4(\cdot;j)\|_{(W^{1,\infty}(Q))^3} \le M$, we have

$$
\begin{aligned}
&\int_Q \left[|\Psi_1(\cdot;j)|^2 + |\Psi_2(\cdot;j)|^2\right] \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t \\
&\quad \le C_{21} \sum_{k=1}^{3} \int_Q \left(|f_k|^2 + |\nabla f_k|^2\right) \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t \\
&\quad \le C_{22} \sum_{k=1}^{3} \int_\Omega \left(|f_k|^2 + |\nabla f_k|^2\right) \mathrm{e}^{2s\varphi(\cdot,0)} \left\{ \int_{-T}^{T} \mathrm{e}^{2s[\varphi(\cdot,t)-\varphi(\cdot,0)]} \mathrm{d}t \right\} \mathrm{d}x \\
&\quad \le C_{23} \kappa_1(s) \sum_{k=1}^{3} \int_\Omega \left(|f_k|^2 + |\nabla f_k|^2\right) \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x
\end{aligned}
\tag{3.48}
$$

for all $s > 0$, where

$$
\kappa_1(s) = \int_{-T}^{T} \mathrm{e}^{2s\left(\mathrm{e}^{-\varrho\beta^2 t^2}-1\right)} \mathrm{d}t. \tag{3.49}
$$

By (1.8) and (3.18)–(3.20), we have

$$
\begin{aligned}
&\int_{Q_\omega} |\nabla \chi_1|^2 \left[|U(\cdot;j)|^2 + |V(\cdot;j)|^2 + |Y(\cdot;j)|^2 + |Z(\cdot;j)|^2\right] \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t \\
&\quad \le C_{24} \mathrm{e}^{2s\Gamma} \int_{Q_\omega} \left[|Y(\cdot;j)|^2 + |Z(\cdot;j)|^2\right] \mathrm{d}x\mathrm{d}t \le C_{25} \mathrm{e}^{2s\Gamma} \Theta^2
\end{aligned}
\tag{3.50}
$$

for all $s > 0$, where $\Gamma = \sup_{(x,t)\in Q} \varphi(x,t) \ge 1$. By (3.36), (3.43), and (3.46), we have

$$
\begin{aligned}
&\left(\int_{-T+\delta}^{-T+2\delta} + \int_{T-2\delta}^{T-\delta}\right) \int_\Omega (\partial_t \chi_2)^2 \left[|Y_1(\cdot;j)|^2 + |Z_1(\cdot;j)|^2\right] \mathrm{e}^{2s\varphi} \mathrm{d}x\mathrm{d}t \\
&\quad \le C_{26} \mathrm{e}^{2s(1-\eta)} \left(\int_{-T+\delta}^{-T+2\delta} + \int_{T-2\delta}^{T-\delta}\right) w(t;j) \mathrm{d}t \\
&\quad \le 2C_{27} \delta \mathrm{e}^{2s(1-\eta)} \left[w(0;j) + \sum_{k=1}^{3} \int_\Omega \left(|f_k|^2 + |\nabla f_k|^2\right) \mathrm{d}x + \Theta^2 \right] \\
&\quad \le C_{28} \delta \mathrm{e}^{-2s\eta} \left\{ \int_\Omega \left[|Y_1(\cdot,0;j)|^2 + |Z_1(\cdot,0;j)|^2\right] \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x \right. \\
&\qquad\qquad \left. + \sum_{k=1}^{3} \int_\Omega \left(|f_k|^2 + |\nabla f_k|^2\right) \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x \right\} + C_{29} \mathrm{e}^{2s\Gamma} \Theta^2
\end{aligned}
\tag{3.51}
$$

for all $s > 0$. For the last inequality, we have used the first inequality in (3.35) and definition (3.43) of $w(t;j)$. Moreover, by $\chi_2(0) = 1$ and the definition of $Y_2(\cdot;j)$, $Z_2(\cdot;j)$, we have $Y_2(x,0;j) = Y_1(x,0;j)$ and $Z_2(x,0;j) = Z_1(x,0;j)$ for $x \in \overline{\Omega}$. Hence, substituting (3.48), (3.50), and (3.51) into (3.47), we have

$$
\begin{aligned}
&\int_\Omega \left[|Y_1(\cdot,0;j)|^2 + |Z_1(\cdot,0;j)|^2\right] \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x \\
&\quad = \int_\Omega \left[|Y_2(\cdot,0;j)|^2 + |Z_2(\cdot,0;j)|^2\right] \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x \\
&\quad \le C_{30} \left\{ \delta \mathrm{e}^{-2s\eta} \int_\Omega \left[|Y_1(\cdot,0;j)|^2 + |Z_1(\cdot,0;j)|^2\right] \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x \right. \\
&\qquad\qquad \left. + \kappa_2(s) \sum_{k=1}^{3} \int_\Omega \left(|f_k|^2 + |\nabla f_k|^2\right) \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x + \mathrm{e}^{2s\Gamma} \Theta^2 \right\}
\end{aligned}
$$

for all large $s > 0$, where

$$\kappa_2(s) = \kappa_1(s) + \delta e^{-2s\eta}. \tag{3.52}$$

Furthermore, there exists $s_4 > 0$ such that $C_{30}\delta e^{-2s\eta} \le 1/2$ for all $s > s_4$. Therefore, we have

$$\begin{aligned}
&\int_\Omega \left[|Y_1(\cdot,0;j)|^2 + |Z_1(\cdot,0;j)|^2\right] e^{2s\varphi(\cdot,0)}dx \\
&\le 2C_{30}\left[\kappa_2(s)\sum_{k=1}^3\int_\Omega \left(|f_k|^2 + |\nabla f_k|^2\right) e^{2s\varphi(\cdot,0)}dx + e^{2s\Gamma}\Theta^2\right] \\
&\le C_{31}\left[\kappa_2(s)\sum_{k=1}^3\int_\Omega |\nabla f_k|^2 e^{2s\varphi(\cdot,0)}dx + e^{2s\Gamma}\Theta^2\right]
\end{aligned} \tag{3.53}$$

for all sufficiently large $s > 0$. For the last inequality in (3.53), we have used (3.34).

Moreover, by the definition of $Y_1(\cdot;j)$, $Z_1(\cdot;j)$, we have

$$Y(\cdot,0;j) = Y_1(\cdot,0;j) + (1-\chi_1)Y(\cdot,0;j),\ Z(\cdot,0;j) = Z_1(\cdot,0;j) + (1-\chi_1)Z(\cdot,0;j) \text{ in } \Omega.$$

Therefore, by $1 - \chi_1(x) = 0$ for $x \in \Omega \setminus \omega$, we have

$$\begin{aligned}
&\int_\Omega \left[|Y(\cdot,0;j)|^2 + |Z(\cdot,0;j)|^2\right] e^{2s\varphi(\cdot,0)}dx \\
&\le C_{32}\left\{\int_\Omega \left[|Y_1(\cdot,0;j)|^2 + |Z_1(\cdot,0;j)|^2\right] e^{2s\varphi(\cdot,0)}dx \right.\\
&\qquad \left. + \int_\omega (1-\chi_1)^2\left[|Y(\cdot,0;j)|^2 + |Z(\cdot,0;j)|^2\right] e^{2s\varphi(\cdot,0)}dx\right\}
\end{aligned} \tag{3.54}$$

for $s > 0$. By the Sobolev embedding theorem (e.g., [1]) and using (1.8) and (3.18) and (3.19), we have

$$\begin{aligned}
&\int_\omega (1-\chi_1)^2\left[|Y(\cdot,0;j)|^2 + |Z(\cdot,0;j)|^2\right] e^{2s\varphi(\cdot,0)}dx \\
&\le C_{33}e^{2s\Gamma}\int_\omega \left[|Y(\cdot,0;j)|^2 + |Z(\cdot,0;j)|^2\right] dx \le C_{34}e^{2s\Gamma}\Theta^2
\end{aligned} \tag{3.55}$$

for all $s > 0$. Substituting (3.53) and (3.55) into (3.54) and summing over $j = 1, 2$, we obtain

$$\begin{aligned}
&\sum_{j=1}^2\int_\Omega \left[|Y(\cdot,0;j)|^2 + |Z(\cdot,0;j)|^2\right] e^{2s\varphi(\cdot,0)}dx \\
&\le C_{35}\left[\kappa_2(s)\sum_{k=1}^3\int_\Omega |\nabla f_k|^2 e^{2s\varphi(\cdot,0)}dx + e^{2s\Gamma}\Theta^2\right]
\end{aligned} \tag{3.56}$$

for all sufficiently large $s > 0$.

Step 5. On the other hand, by (1.1), we have $D[\epsilon;\Phi(j)](\cdot,0) = D[\widetilde{\epsilon};\Phi(j)](\cdot,0) = d^j$, $B[\epsilon;\Phi(j)](\cdot,0) = B[\widetilde{\epsilon};\Phi(j)](\cdot,0) = b^j$ in Ω. Therefore, by (3.28) and (3.29), we have

$$\{E[\epsilon;\Phi(j)] - E[\widetilde{\epsilon};\Phi(j)]\}(\cdot,0) = -f_1 d^j - f_2 b^j \quad \text{in } \Omega, \tag{3.57}$$

$$\{H[\epsilon;\Phi(j)] - H[\widetilde{\epsilon};\Phi(j)]\}(\cdot,0) = -f_2 d^j - f_3 b^j \quad \text{in } \Omega. \tag{3.58}$$

By (3.32), (3.33), (3.57), and (3.58), direct calculations yield

$$\begin{aligned} Y(\cdot,0;j) &= -\nabla\times\left(f_2 d^j + f_3 b^j\right) \\ &= -f_2\left(\nabla\times d^j\right) - f_3\left(\nabla\times b^j\right) - (\partial_1 f_2)\left(e_1\times d^j\right) - (\partial_1 f_3)\left(e_1\times b^j\right) \\ &\quad - (\partial_2 f_2)\left(e_2\times d^j\right) - (\partial_2 f_3)\left(e_2\times b^j\right) - (\partial_3 f_2)\left(e_3\times d^j\right) - (\partial_3 f_3)\left(e_3\times b^j\right) \quad \text{in } \Omega, \\ Z(\cdot,0;j) &= \nabla\times\left(f_1 d^j + f_2 b^j\right) \\ &= f_1\left(\nabla\times d^j\right) + f_2\left(\nabla\times b^j\right) + (\partial_1 f_1)\left(e_1\times d^j\right) + (\partial_1 f_2)\left(e_1\times b^j\right) \\ &\quad + (\partial_2 f_1)\left(e_2\times d^j\right) + (\partial_2 f_2)\left(e_2\times b^j\right) + (\partial_3 f_1)\left(e_3\times d^j\right) + (\partial_3 f_2)\left(e_3\times b^j\right) \quad \text{in } \Omega. \end{aligned}$$

Therefore, by definition of $\mathbb{G}$, we have

$$\mathbb{G}F = \begin{pmatrix} -Y(\cdot,0;1) \\ Z(\cdot,0;1) \\ -Y(\cdot,0;2) \\ Z(\cdot,0;2) \end{pmatrix} - \begin{pmatrix} 0 & \nabla\times d^1 & \nabla\times b^1 \\ \nabla\times d^1 & \nabla\times b^1 & 0 \\ 0 & \nabla\times d^2 & \nabla\times b^2 \\ \nabla\times d^2 & \nabla\times b^2 & 0 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \quad \text{in } \Omega, \tag{3.59}$$

where $F = (\partial_1 f_1, \partial_1 f_2, \partial_1 f_3, \partial_2 f_1, \partial_2 f_2, \partial_2 f_3, \partial_3 f_1, \partial_3 f_2, \partial_3 f_3)^{\mathrm{T}}$. By (1.6) and (3.59), we see that

$$\sum_{k=1}^{3} |\nabla f_k|^2 \le C_{36} |\mathbb{G}F|^2 \le C_{37} \left\{ \sum_{j=1}^{2} \left[|Y(\cdot,0;j)|^2 + |Z(\cdot,0;j)|^2\right] + \sum_{k=1}^{3} |f_k|^2 \right\} \quad \text{in } \Omega.$$

Hence, by (3.34) and (3.56), we have

$$\begin{aligned} &\sum_{k=1}^{3} \int_\Omega \left(|f_k|^2 + |\nabla f_k|^2\right) \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x \\ &\le C_{38} \int_\Omega \left\{ \sum_{j=1}^{2} \left[|Y(\cdot,0;j)|^2 + |Z(\cdot,0;j)|^2\right] + \sum_{k=1}^{3} |f_k|^2 \right\} \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x \\ &\le C_{39} \left[\kappa_3(s) \sum_{k=1}^{3} \int_\Omega |\nabla f_k|^2 \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x + \mathrm{e}^{2s\Gamma} \Theta^2 \right] \end{aligned} \tag{3.60}$$

for all sufficiently large $s > 0$, where

$$\kappa_3(s) = \kappa_2(s) + \frac{1}{s}. \tag{3.61}$$

Furthermore, by definitions (3.49), (3.52), and (3.61) of $\kappa_3(s)$, we see that $\lim_{s\to\infty} \kappa_3(s) = 0$. Therefore, there exists $s_5 > 0$ such that $C_{39}\kappa_3(s) \le 1/2$ for $s > s_5$. Therefore, by (3.60), we can obtain that

$$\sum_{k=1}^{3} \int_\Omega \left(|f_k|^2 + |\nabla f_k|^2\right) \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x \le 2C_{39} \mathrm{e}^{2s\Gamma} \Theta^2$$

for all sufficiently large $s > 0$. Consequently, by the first inequality in (3.35), we have

$$\begin{aligned} \sum_{k=1}^{3} \int_\Omega \left(|f_k|^2 + |\nabla f_k|^2\right) \mathrm{d}x &\le \mathrm{e}^{-2s} \sum_{k=1}^{3} \int_\Omega \left(|f_k|^2 + |\nabla f_k|^2\right) \mathrm{e}^{2s\varphi(\cdot,0)} \mathrm{d}x \\ &\le 2C_{39} \mathrm{e}^{2s(\Gamma-1)} \Theta^2 \end{aligned} \tag{3.62}$$

for all sufficiently large $s > 0$. Hence, taking $s > 0$ sufficiently large and fixing it, we obtain that

$$\sum_{k=1}^{3} \int_\Omega |f_k|^2 \mathrm{d}x \le \sum_{k=1}^{3} \int_\Omega \left(|f_k|^2 + |\nabla f_k|^2\right) \mathrm{d}x \le C_{40}\Theta^2. \tag{3.63}$$

Moreover, by direct calculations, we can verify that

$$\begin{aligned}
&\widetilde{\varepsilon} - \varepsilon = \widetilde{\xi} f_3 + \gamma_3 \xi \widetilde{\xi} \left[(\widetilde{\gamma}_2 + \gamma_2) f_2 - \gamma_1 f_3 - \widetilde{\gamma}_3 f_1 \right], \\
&\widetilde{\zeta} - \zeta = -\widetilde{\xi} f_2 - \gamma_2 \xi \widetilde{\xi} \left[(\widetilde{\gamma}_2 + \gamma_2) f_2 - \gamma_1 f_3 - \widetilde{\gamma}_3 f_1 \right], \\
&\widetilde{\mu} - \mu = \widetilde{\xi} f_1 + \gamma_1 \xi \widetilde{\xi} \left[(\widetilde{\gamma}_2 + \gamma_2) f_2 - \gamma_1 f_3 - \widetilde{\gamma}_3 f_1 \right] \quad \text{on } \overline{\Omega}.
\end{aligned}$$

Therefore, by (3.63), we have

$$\|\widetilde{\epsilon} - \epsilon\|_{L^2(\Omega)} + \left\|\widetilde{\zeta} - \zeta\right\|_{L^2(\Omega)} + \|\widetilde{\mu} - \mu\|_{L^2(\Omega)} \le c_{41} \sum_{k=1}^{3} \|f_k\|_{L^2(\Omega)} \le C_{42}\Theta. \tag{3.64}$$

The proof of Theorem 1.1 is complete. $\square$

Acknowledgments. The author thanks Professor Victor Isakov (Wichita State University, Wichita, KS, USA) for his valuable comments. Finally, the author would like to express her sincerest thanks to Professor Masahiro Yamamoto for his patient guidance and useful discussions, and thanks referees for invaluable comments.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] A. L. Bukhgeim, *Introduction to the Theory of Inverse Problems*, VSP, Utrecht, The Netherlands, 2000.
[3] A. L. Bukhgeim and M. V. Klibanov, *Global uniqueness of a class of multidimensional inverse problems*, Sov. Math.-Dokl., 24 (1981), pp. 244–247.
[4] J. Cheng, V. Isakov, M. Yamamoto, and Q. Zhou, *Lipschitz stability in the lateral Cauchy problem for elasticity system*, J. Math. Kyoto Univ., 43 (2003), pp. 475–501.
[5] I. E. Dzyaloshinskii, *On the Magneto-electrical effect in antiferromagnets*, Soviet Phys. JETP 10 (1960), pp. 628–629. (Translated from J. Exptl. Theoret. Phys. (U.S.S.R.), 37 (1959), pp. 881–882.)
[6] L. Hörmander, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, 1963.
[7] L. Hörmander, *The Analysis of Linear Partial Differential Operators*, I–IV, Springer-Verlag, Berlin, 1983, 1985.
[8] O. Y. Imanuvilov, *On Carleman estimates for hyperbolic equations*, Asymptot. Anal., 32 (2002), pp. 185–220.
[9] O. Y. Imanuvilov, V. Isakov, and M. Yamamoto, *An inverse problem for the dynamical Lamé system with two sets of boundary data*, Comm. Pure Appl. Math., 56 (2003), pp. 1366–1382.
[10] O. Y. Imanuvilov and M. Yamamoto, *Global Lipschitz stability in an inverse hyperbolic problem by interior observations*, Inverse Problems, 17 (2001), pp. 717–728.
[11] O. Y. Imanuvilov and M. Yamamoto, *Carleman estimate for a parabolic equation in a Sobolev space of negative order and its applications*, in Control of Nonlinear Distributed Parameter Systems, Lecture Notes in Pure and Appl. Math., Marcel-Dekker, New York, 218 (2001), pp. 113–137.
[12] O. Y. Imanuvilov and M. Yamamoto, *Determination of a coefficient in an acoustic equation with a single measurement*, Inverse Problems, 19 (2003), pp. 157–171.
[13] O. Y. Imanuvilov and M. Yamamoto, *Stability estimate in a Cauchy problem for a hyperbolic equation with variable coefficients*, Preprint 2004-27, Graduate School of Mathematical Sciences, University of Tokyo, Tokyo, 2004.

[14] V. Isakov, *Uniqueness of the continuation across a time-like hyperplane and related inverse problems for hyperbolic equations*, Comm. Partial Differential Equations, 14 (1989), pp. 465–478.

[15] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer-Verlag, New York, 1998.

[16] V. Isakov, *Carleman type estimates and their applications*, in New Analytic and Geometric Methods in Inverse Problems, Springer-Verlag, Berlin, 2004, pp. 93–125.

[17] A. Khaĭdarov, *Carleman estimates and inverse problems for second order hyperbolic equations*, Math. USSR Sb., 58 (1987), pp. 267–277.

[18] A. Khaĭdarov, *On stability estimates in multidimensional inverse problems for differential equations* (English translation), Soviet Math. Dokl., 38 (1989), pp. 614–617.

[19] M. V. Klibanov, *Inverse problems and Carleman estimates*, Inverse Problems, 8 (1992), pp. 575–596.

[20] M. V. Klibanov and A. Timonov, *Carleman Estimates for Coefficient Inverse Problems and Numerical Applications*, VSP, Utrecht, The Netherlands, 2004.

[21] J. A. Kong, *Electromagnetic Wave Theory*, 2nd ed., John-Wiley, New York, 1990.

[22] L. D. Landau and E. M. Lifshitz, *Electrodynamics of Continuous Media*, Addison-Wesley, Reading, MA, 1960.

[23] S. M. Li, and M. Yamamoto, *Inverse source problem for Maxwell's equations in anisotropic media*, Applicable Analysis, 84 (2005), pp.1051–1067.

[24] T. H. O'dell, *The Electrodynamics of Magneto-electric Media*, Series of Monographs on Selected Topics in Solid State Physics, E. P. Wohlfarth, ed., North-Holland Publishing Company, Amsterdam, The Netherlands, 1970.

[25] A. Ruiz, *Unique continuation for weak solutions of the wave equation plus a potential*, J. Math. Pures Appl., 71 (1992), pp. 455–467.

[26] V. G. Romanov, *Inverse Problem of Mathematical Physics*, VNU Science Press, Utrecht, The Netherlands, 1987.

[27] V. G. Romanov and S. I. Kabanikhin, *Inverse Problems for Maxwell's Equations*, VSP, Utrecht, The Netherlands, 1994.

[28] Z. Sun and G. Uhlmann, *An inverse boundary problem for Maxwell's Equations*, Arch. Rational Mech. Anal., 119 (1992), pp. 71–93.

[29] M. Yamamoto, *A mathematical aspect of inverse problems for non-stationary Maxwell's equations*, Int. J. of Applied Electromagnetics and Mechanics, 8 (1997), pp. 77–98.

[30] M. Yamamoto, *On an inverse problem of determining source terms in Maxwell's equations with a single measurement*, Inverse Problems, Tomography, and Image Processing, Plenum Press, New York, 15 (1998), pp. 241–256.

[31] M. Yamamoto, *Uniqueness and stability in multidimensional hyperbolic inverse problems*, J. Math. Pure. Appl., 78 (1999), pp. 65–98.

SIAM J. MATH. ANAL.
Vol. 37, No. 4, pp. 1044–1069

GLOBAL WEAK SOLUTIONS TO A GENERALIZED HYPERELASTIC-ROD WAVE EQUATION*

G. M. COCLITE†, H. HOLDEN‡, AND K. H. KARLSEN†

Abstract. We consider a generalized hyperelastic-rod wave equation (or generalized Camassa–Holm equation) describing nonlinear dispersive waves in compressible hyperelastic rods. We establish existence of a strongly continuous semigroup of global weak solutions for any initial data from $H^1(\mathbb{R})$. We also present a "weak equals strong" uniqueness result.

Key words. Hyperelastic-rod wave equation, Camassa–Holm equation, weak solutions, existence, uniqueness

AMS subject classifications. 35G25, 35L05, 35A05

DOI. 10.1137/040616711

1. Introduction and statement of main results. In recent years the so-called Camassa–Holm equation [3] has caught a great deal of attention. It is a nonlinear dispersive wave equation that takes the form

$$\frac{\partial u}{\partial t} - \frac{\partial^3 u}{\partial t \partial x^2} + 2\kappa \frac{\partial u}{\partial x} + 3u \frac{\partial u}{\partial x} = 2 \frac{\partial u}{\partial x} \frac{\partial^2 u}{\partial x^2} + u \frac{\partial^3 u}{\partial x^3}, \quad t > 0, \ x \in \mathbb{R}. \tag{1.1}$$

When $\kappa > 0$, this equation models the propagation of unidirectional shallow water waves on a flat bottom, and $u(t,x)$ represents the fluid velocity at time t in the horizontal direction x [3, 22]. The Camassa–Holm equation possesses a bi-Hamiltonian structure (and thus an infinite number of conservation laws) [20, 3] and is completely integrable [3, 1, 11, 6]. Moreover, when $\kappa = 0$ it has an infinite number of solitary wave solutions, called *peakons* due to the discontinuity of their first derivatives at the wave peak, interacting like solitons:

$$u(t,x) = ce^{-|x-ct|}, \qquad c \in \mathbb{R}.$$

The solitary waves with $\kappa > 0$ are smooth, while they become peaked when $\kappa \to 0$. From a mathematical point of view the Camassa–Holm equation is well studied. Local well-posedness results are proved in [7, 21, 24, 30]. It is also known that there exist global solutions for a particular class of initial data and also solutions that blow up in finite time for a large class of initial data [5, 7, 10]. Here blow up means that the slope of the solution becomes unbounded while the solution itself stays bounded. More relevant for the present paper, we recall that existence and uniqueness results for global weak solutions of (1.1) with $\kappa = 0$ have been proved by Constantin and Escher [8], Constantin and Molinet [12], and Xin and Zhang [32, 33], see also Danchin [17, 18].

*Received by the editors October 11, 2004; accepted for publication (in revised form) May 9, 2005; published electronically December 30, 2005. This work was supported by the BeMatA program of the Research Council of Norway and the European network HYKE, funded by the EC as contract HPRN-CT-2002-00282.

http://www.siam.org/journals/sima/37-4/61671.html

†Centre of Mathematics for Applications, University of Oslo, P.O. Box 1053, Blindern, NO–0316 Oslo, Norway (giusepc@math.uio.no, kennethk@math.uio.no). The third author's research was also supported by an Outstanding Young Investigators Award from the Research Council of Norway.

‡Department of Mathematical Sciences, Norwegian University of Science and Technology, NO–7491 Trondheim, Norway (holden@math.ntnu.no).

Here we are interested in the Cauchy problem for the nonlinear equation

$$\frac{\partial u}{\partial t} - \frac{\partial^3 u}{\partial t \partial x^2} + \frac{\partial}{\partial x}\left(\frac{g(u)}{2}\right) = \gamma\left(2\frac{\partial u}{\partial x}\frac{\partial^2 u}{\partial x^2} + u\frac{\partial^3 u}{\partial x^3}\right), \quad t > 0, \ x \in \mathbb{R}, \tag{1.2}$$

where the function $g : \mathbb{R} \to \mathbb{R}$ and the constant $\gamma \in \mathbb{R}$ are given. Observe that if $g(u) = 2\kappa u + 3u^2$ and $\gamma = 1$, then (1) is the classical Camassa–Holm equation. With $g(u) = 3u^2$, Dai [15, 14, 16] derived (1) as an equation describing finite length, small amplitude radial deformation waves in cylindrical compressible hyperelastic rods, and often referred to it as the hyperelastic-rod wave equation. Stability of solitary waves for this equation was studied in [13]. The constant γ is given in terms of the material constants and the prestress of the rod; we coin (1) the *generalized hyperelastic-rod wave equation*.

In the derivation of the Camassa–Holm equation in the context of the shallow water waves [3, 22], the constant κ is proportional to the square root of water depth. Thus under normal circumstances it is not physical to set $\kappa = 0$. Although strictly speaking one does not have peakons in the shallow water model ($\kappa > 0$), one has them in Dai's model for compressible hyperelastic rods, since in this model $g(u) = 3u^2$ and $\gamma \in \mathbb{R}$. For $\gamma = 0$ and $g(u) = 3u^2$, (1) becomes the regularized wave equation describing surface waves in channel [2]; the solutions are global, the equation has an Hamiltonian structure but is not integrable, and its solitary waves are not solitons.

A difference between the Camassa–Holm equation (1.1) (with $\kappa = 0$) and the generalized hyperelastic-rod wave equation (1) is that (the slope of) solitary wave solutions to (1) can blow up, while they cannot for (1.1). Solitary waves are bounded solutions of (1) of the form $u(t, x) = \varphi(x - ct)$, where c is the wave speed. It is not hard to check that $\varphi(\zeta)$, $\zeta = x - ct$, satisfies the ordinary differential equation $(\varphi')^2 = \frac{c\varphi^2 - G(\phi)}{c - \gamma\varphi}$, where $G(\xi) = \int_0^\xi g(\xi) d\xi$. From this expression it is clear that $|\varphi'|$ can become infinite. Notice, however, that for the Camassa–Holm equation (1.1) (with $\kappa = 0$), for which $G(u) = u^3$, it follows from the above equation that $(\varphi')^2 = \varphi$ (if $\varphi \neq c/\gamma$) and thus any solitary wave (peakon) φ belongs to $W^{1,\infty}$. Notice also that for (1) with $g(u) = 2\kappa u + 3u^2$, the above ordinary differential equation becomes $(\varphi')^2 = \phi^2 \frac{(c-\kappa) - \varphi}{c - \gamma\varphi}$, and choosing $\gamma = \frac{c}{c-\kappa}$, $c \neq \kappa$, we find the peakon solution

$$\varphi(\xi) = (c - \kappa)e^{-\sqrt{\frac{c-\kappa}{c}}|\xi|}. \tag{1.3}$$

From a mathematical point of view the generalized hyperelastic-rod wave equation (1) is much less studied than (1.1). Recently, Yin [34, 35, 36] (see also Constantin and Escher [9]) proved local well-posedness, global well-posedness for a particular class of initial data, and in particular that smooth solutions blow up in finite time (with a precise estimate of the blow-up time) for a large class of initial data. Lopes [28] proved stability of solitary waves for (1) with $\gamma = 1$, while Kalisch [23] studied the stability when $g(u) = 2\kappa u + 3u^3$ and $\gamma \in \mathbb{R}$. Qian and Tang [29] used the bifurcation method to study peakons and periodic cusp waves for (1) with $g(u) = 2\kappa u + au^2$, $\kappa, a \in \mathbb{R}$, $\gamma = 1$. When $a \neq 3$, $a > 0$, $\kappa \neq 0$, they found the following two peakon type solutions: $u(t, x) = \frac{6\kappa}{3-a} \exp\left(-\sqrt{\frac{a}{3}}|x - \frac{6\kappa t}{3-a}|\right)$ and $u(t, x) = \frac{2\kappa}{a+1}\left(3a \exp\left(-\sqrt{\frac{a}{3}}|x - \frac{2\kappa t}{a+1}|\right) - 2\right)$. When $a = 3$ and $\kappa \neq 0$ they also found a peakon type solution of the form $u(t, x) = \frac{3\kappa}{2} \exp(-|x - \frac{\kappa t}{2}|) - \kappa$. For (1) with $g(u) = 3u^2$, Dai [16] has constructed explicitly a variety of traveling waves, including solitary shock (or peakon like) waves. To give an example, suppose $0 < \gamma < 3$ and pick any constant $c > 0$. Then the following peakon like function is a travelling wave solution:

$$u(t,x) = \frac{1}{2}\left(1 - \frac{1}{\gamma}\right)c + \frac{c}{2}\left(\frac{3}{\gamma} - 1\right)\exp\left(-\frac{1}{\sqrt{\gamma}}|x - ct - \zeta|\right),$$

where ζ is a particular constant. Dai refers to this as a supersonic solitary shock wave. Although all the above displayed peakon type solutions belong to $W^{1,\infty}$ they do not all belong to $H^1(\mathbb{R})$ (some of them do not decay to zero at $\pm\infty$) and these cannot be encompassed by our theory.

Up to now there are no global existence results for weak solutions to the generalized hyperelastic-rod wave equation (1). Here we establish the existence of a global weak solution to (1) for any initial function u_0 belonging to $H^1(\mathbb{R})$. Furthermore, we prove the existence of a strongly continuous semigroup, which in particular implies stability of the solution with respect to perturbations of data in the equation as well as perturbation in the initial data. Our approach is based on a vanishing viscosity argument, showing stability of the solution when a regularizing term vanishes. This stability result is new even for the Camassa–Holm equation (1.1). Finally, we prove a "weak equals strong" uniqueness result. Here we follow closely the approach of Xin and Zhang [32] for the Camassa–Holm equation (1.1) with $\kappa = 0$.

Let us be more precise about our results. We shall assume that

$$u|_{t=0} = u_0 \in H^1(\mathbb{R}), \tag{1.4}$$

and

$$g \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}), \qquad g(0) = 0, \qquad \gamma > 0. \tag{1.5}$$

Observe that the case $\gamma = 0$ is much simpler than the one we are considering. Moreover, if $\gamma < 0$, peakons become antipeakons, so we can use a similar argument. The assumption of infinite differentiability of g is made just for convenience. In fact, locally Lipschitz continuity would be sufficient. Define

$$h(\xi) := \frac{1}{2}\big(g(\xi) - \gamma\xi^2\big)$$

for $\xi \in \mathbb{R}$. Rewriting (1) as

$$(1 - \partial_x^2)u_t + \gamma(1 - \partial_x^2)uu_x + \left(h(u) + \frac{\gamma}{2}u_x^2\right)_x = 0, \tag{1.6}$$

we see that (1) formally is equivalent to the elliptic-hyperbolic system

$$\frac{\partial u}{\partial t} + \gamma u\frac{\partial u}{\partial x} + \frac{\partial P}{\partial x} = 0, \quad -\frac{\partial^2 P}{\partial x^2} + P = h(u) + \frac{\gamma}{2}\left(\frac{\partial u}{\partial x}\right)^2. \tag{1.7}$$

Moreover, since $e^{-|x|}/2$ is the Green's function of the operator $-\frac{\partial^2}{\partial x^2} + 1$, (1) is equivalent to the integro-differential equation

$$\frac{\partial u}{\partial t} + \gamma u\frac{\partial u}{\partial x} + \frac{\partial P}{\partial x} = 0, \quad P(t,x) = \frac{1}{2}\int_{\mathbb{R}} e^{-|x-y|}\left(h(u(t,y)) + \frac{\gamma}{2}\left(\frac{\partial u}{\partial x}(t,y)\right)^2\right)dy. \tag{1.8}$$

Motivated by this, we shall use the following definition of weak solution.

DEFINITION 1.1. *We call $u\colon [0,\infty)\times\mathbb{R}\to\mathbb{R}$ a weak solution of the Cauchy problem for* (1) *if*

(*i*) $u\in C([0,\infty)\times\mathbb{R})\cap L^\infty\big((0,\infty);H^1(\mathbb{R})\big)$;

(*ii*) *u satisfies* (1.7) *in the sense of distributions for some* $P\in L^\infty([0,\infty):W^{1,\infty}(\mathbb{R}))$;

(*iii*) $u(0,x)=u_0(x)$, *for every* $x\in\mathbb{R}$;

(*iv*) $\|u(t,\cdot)\|_{H^1(\mathbb{R})}\le\|u_0\|_{H^1(\mathbb{R})}$, *for each* $t>0$.

If, in addition, there exists a positive constant K_1 depending only on $\|u_0\|_{H^1(\mathbb{R})}$ such that

$$\frac{\partial u}{\partial x}(t,x)\le\frac{2}{\gamma t}+K_1,\qquad (t,x)\in(0,\infty)\times\mathbb{R},\tag{1.9}$$

then we call u an admissible weak solution of the Cauchy problem for (1).

Our existence results are collected in the following theorem.

THEOREM 1.2. *There exists a strongly continuous semigroup of solutions associated to the Cauchy problem* (1). *More precisely, there exists a map*

$$\begin{aligned}S\colon[0,\infty)\times(0,\infty)\times\mathcal{E}\times H^1(\mathbb{R})&\longrightarrow C([0,\infty)\times\mathbb{R})\cap L^\infty\big([0,\infty);H^1(\mathbb{R})\big),\\ (t,\gamma,g,u_0)&\mapsto S_t(\gamma,g,u_0),\end{aligned}$$

where

$$\mathcal{E}:=\big\{g\in\mathrm{Lip}_{\mathrm{loc}}(\mathbb{R})\mid g(0)=0\big\}$$

with the following properties:

(*j*) *for each $u_0\in H^1(\mathbb{R})$, $\gamma>0$, $g\in\mathcal{E}$ the map $u(t,x)=S_t(\gamma,g,u_0)(x)$ is an admissible weak solution of* (1);

(*jj*) *it is stable with respect to the initial condition in the following sense, if*

$$u_{0,n}\longrightarrow u_0\ \text{ in } H^1(\mathbb{R}),\ \ \gamma_n\longrightarrow\gamma,\ \ g_n'\longrightarrow g'\ \text{ in } L^\infty(\mathcal{I}),\tag{1.10}$$

then

$$S_t(\gamma_n,g_n,u_{0,n})\longrightarrow S_t(\gamma,g,u_0)\ \text{ in } L^\infty([0,T];H^1(\mathbb{R})),\tag{1.11}$$

for every $\{u_{0,n}\}_{n\in\mathbb{N}}\subset H^1(\mathbb{R})$, $\{\gamma_n\}_{n\in\mathbb{N}}\subset(0,\infty)$, $\{g_n\}_{n\in\mathbb{N}}\subset\mathcal{E}$, $u_0\in H^1(\mathbb{R})$, $\gamma>0$, $g\in\mathcal{E}$, $T>0$, where

$$\mathcal{I}:=\frac{1}{\sqrt{2}}\Big[-\sup_n\|u_{0,n}\|_{H^1(\mathbb{R})},\sup_n\|u_{0,n}\|_{H^1(\mathbb{R})}\Big].$$

Moreover, the following statements hold:

(*k*) *Estimate* (1.9) *is valid with*

$$K_1:=\sqrt{\frac{2}{\gamma}}\bigg(2\max_{|\xi|\le\sqrt{2}\|u_0\|_{H^1(\mathbb{R})}}\big|h(\xi)\big|+\frac{\gamma}{2}\|u_0\|^2_{H^1(\mathbb{R})}\bigg)^{1/2}.$$

(*kk*) *There results*

$$\frac{\partial}{\partial x}S_t(\gamma,g,u_0)\in L^p_{\mathrm{loc}}([0,\infty)\times\mathbb{R}),\tag{1.12}$$

for each $1\le p<3$.

(*kkk*) *The following identity holds in the sense of distributions on* $[0,\infty)\times\mathbb{R}$*:*

$$\frac{\partial}{\partial t}\left(\frac{1}{2}\left[u^2+q^2\right]\right)+\frac{\partial}{\partial x}\left(u\left[\frac{\gamma}{2}q^2+P\right]+\frac{\gamma}{3}u^3-H(u)\right)=-\mu, \tag{1.13}$$

where $u=S_t(\gamma,g,u_0)$*,* $q=\frac{\partial}{\partial x}S_t(\gamma,g,u_0)$*,* $H'=h$*, the defect measure* μ *is a nonnegative Radon measure such that as* $R\to\infty$ *there holds* $Rq\,(q+R)\,\chi_{(-\infty,-R)}(q)\overset{\star}{\rightharpoonup}\mu$ *in the sense of measures and* $\mu([0,\infty)\times\mathbb{R})\le\frac{1}{2}\|u_0\|_{H^1(\mathbb{R})}$*.*

We stress that the existence of a strongly continuous semigroup is new, even for the Camassa–Holm equation itself. In particular, this includes the stability of the solution with respect to perturbations in the initial data and the coefficients in the equation.

As in Xin and Zhang [32, 33] and their study of the Camassa–Holm equation (1.1) with $\kappa=0$, we prove existence of a global weak solution by establishing convergence as $\varepsilon\to 0$ of a sequence of smooth viscous approximate solutions u_ε (see (2.1)). Regarding the limiting process there is a an interesting mathematical issue: we need to prove that the derivative $q_\varepsilon=\partial u_\varepsilon/\partial x$, which a priori is only weakly compact, in fact converges strongly (along a subsequence). Strong convergence of q_ε is needed if we want to send ε to zero in the viscous problem and recover (1). To improve the weak convergence of q_ε to strong convergence we follow [32] closely when using renormalization theory for linear transport equations with nonsmooth coefficients. The idea of renormalization goes back to DiPerna and Lions [19] and it has been developed further and applied by many authors (see Lions [26, 27], Xin and Zhang [32], and the references given therein for relevant information). In the process of improving weak convergence to strong convergence, the higher integrability estimate (1.12) for q_ε is crucial. It ensures that the weak limit of q_ε^2 does not contain singular measures (there are no concentration effects).

Regarding the optimality of (1.12), one should keep in mind that when a solution u blows up (necessarily in the sense that $|\partial u/\partial x|\to\infty$), say at $x=0$, then u must behave like $x^{2/3}$ and $\partial u/\partial x$ like $x^{-1/3}$, since $u(t,\cdot)\in H^1(\mathbb{R})$, in which case $\partial u/\partial x$ belongs to L^p_{loc} if and only if $1\le p<3$.

Denote by u an (admissible) weak solution. If the associated defect measure μ defined in (1.13) vanishes, then we call u an *energy conservative (admissible) weak solution*. Xin and Zhang [33] proved a "weak equals strong" uniqueness result for energy conservative admissible weak solutions of the Camassa–Holm equation (1.1) when $\kappa=0$. Their result also contains the uniqueness result of Constantin and Molinet [12] as a special case.

By slightly adapting the arguments of Xin and Zhang we can also have a "weak equals strong" uniqueness result for the generalized hyperelastic-rod wave equation.

THEOREM 1.3. *Suppose there exists a function* u *such that* (*i*)*,* (*ii*)*, and* (*iii*) *of Definition* 1.1 *hold and that there exists a function* $\beta\in L^2([0,T))$ *for all* $T>0$ *such that* $\left\|\frac{\partial u}{\partial x}(t,\cdot)\right\|_{L^\infty(\mathbb{R})}\le\beta(t)$ *for any* $t\ge 0$*. Then the energy conservative admissible weak solution is unique.*

Due to strong similarities with the proof of Xin and Zhang, we do not prove Theorem 1.3 here and instead ask that the reader consult their paper [33].

Whenever a sufficiently regular solution to (1) can be found (see [9, 16, 29, 34, 35, 36] for some situations where this happens), then Theorem 1.3 ensures that this solution is unique in the class energy conservative admissible weak solutions. Note that peakons are "sufficiently regular." For example, the peakon solution (1.3) is covered by our theory. One should compare Theorem 1.3 with the uniqueness/stability assertion in Theorem 1.2, which states that there is uniqueness in the class of vanishing viscosity solutions.

In passing, we mention that it is apparently not easy to prove existence and uniqueness results for (1) by adapting the methods in [8, 12] for the Camassa–Holm equation, which are based on studying the equation for the "vorticity" $m := (1-\frac{\partial^2}{\partial x^2})u$. In the present context the equation for m reads

$$\frac{\partial m}{\partial t} + \gamma u \frac{\partial m}{\partial x} + 2\gamma \frac{\partial u}{\partial x} m = -\frac{1}{2}\frac{\partial}{\partial x}\big(g(u) - 3\gamma u^2\big). \tag{1.14}$$

In the case of the Camassa–Holm equation (that is, $g(u) = 3u^2$ and $\gamma = 1$), the right-hand side of (1.14) vanishes, and assuming that $m|_{t=0}$ is a bounded nonnegative measure it is not difficult to see that $m(t,\cdot) \in L^1$ remains nonnegative at later times and consequently one can bound $\partial u/\partial x$ in L^∞ and $\partial^2 u/\partial x^2$ in L^1. Using these bounds one can in fact prove the existence and uniqueness of an energy conservative weak solution [8, 12]. In the general case ($g(u)$ is not equal to $3\gamma u^2$), it seems difficult to implement this strategy for proving existence and uniqueness results, and this fact has motivated us to use the "weak convergence" approach.

The remainder of this paper is organized as follows. Section 2 is devoted to stating the viscous problem and a corresponding well-posedness result. In sections 3 and 4 we establish, respectively, an Oleinik type estimate and a higher integrability estimate for the viscous approximants. Section 5 is devoted to proving basic compactness properties for the viscous approximants. Strong compactness of the derivative of the viscous approximants is obtained in section 6, where an existence result for (1) is also stated. In section 7 we prove the uniqueness of the vanishing viscosity limit, this defines a semigroup of solutions as stated in Theorem 1.2. In section 8 we prove the continuity properties of the semigroup.

2. Viscous approximants: Existence and energy estimate. We will prove existence of a weak solution to the Cauchy problem for (1) by proving compactness of a sequence of smooth functions $\{u_\varepsilon\}_{\varepsilon>0}$ solving the following viscous problems (see [4, Theorem 2.3]):

$$\begin{cases} \dfrac{\partial u_\varepsilon}{\partial t} + \gamma u_\varepsilon \dfrac{\partial u_\varepsilon}{\partial x} + \dfrac{\partial P_\varepsilon}{\partial x} = \varepsilon \dfrac{\partial^2 u_\varepsilon}{\partial x^2}, & t>0,\ x\in\mathbb{R},\\ -\dfrac{\partial^2 P_\varepsilon}{\partial x^2} + P_\varepsilon = h(u_\varepsilon) + \dfrac{\gamma}{2}\left(\dfrac{\partial u_\varepsilon}{\partial x}\right)^2, & t>0,\ x\in\mathbb{R},\\ u_\varepsilon(0,x) = u_{\varepsilon,0}(x), & x\in\mathbb{R}. \end{cases} \tag{2.1}$$

We shall assume that

$$\|u_{\varepsilon,0}\|_{H^1(\mathbb{R})} \le \|u_0\|_{H^1(\mathbb{R})},\ \ \varepsilon>0, \quad \text{and} \quad u_{\varepsilon,0}\to u_0 \ \text{in}\ H^1(\mathbb{R}). \tag{2.2}$$

The starting point of our analysis is the following well-posedness result for (2.1).

THEOREM 2.1. *Assume* (1.4) *and* (2.2). *Let* $\varepsilon > 0$, $u_{\varepsilon,0} \in H^\ell(\mathbb{R})$ *and* $\ell \ge 2$. *Then there exists a unique solution* $u_\varepsilon \in C\big(\mathbb{R}; H^\ell(\mathbb{R})\big)$ *to the Cauchy problem* (2.1). *Moreover, for each* $t \ge 0$,

$$\begin{aligned}&\int_{\mathbb{R}}\left(u_\varepsilon^2 + \left(\frac{\partial u_\varepsilon}{\partial x}\right)^2\right)(t,x)dx\\ &\quad + 2\varepsilon\int_0^t\int_{\mathbb{R}}\left(\left(\frac{\partial u_\varepsilon}{\partial x}\right)^2 + \left(\frac{\partial^2 u_\varepsilon}{\partial x^2}\right)^2\right)(s,x)dxds = \|u_{\varepsilon,0}\|^2_{H^1(\mathbb{R})},\end{aligned} \tag{2.3}$$

or

$$\|u_\varepsilon(t,\cdot)\|^2_{H^1(\mathbb{R})} + 2\varepsilon\int_0^t \left\|\frac{\partial u_\varepsilon}{\partial x}(s,\cdot)\right\|^2_{H^1(\mathbb{R})} ds = \|u_{\varepsilon,0}\|^2_{H^1(\mathbb{R})}.$$

Remark 2.2. Due to [25, Theorem 8.5], (2.2) and (2.3), we have for each $t \geq 0$

$$\|u_\varepsilon(t,\cdot)\|_{L^\infty(\mathbb{R})} \leq \frac{1}{\sqrt{2}}\,\|u_\varepsilon(t,\cdot)\|_{H^1(\mathbb{R})} \leq \frac{1}{\sqrt{2}}\|u_0\|_{H^1(\mathbb{R})}. \tag{2.4}$$

Proof of Theorem 2.1. From Theorem 2.3 in [4] we infer that (2.1) has a solution $u_\varepsilon \in C(\mathbb{R}; H^\ell(\mathbb{R}))$. Define

$$q_\varepsilon(t,x) := \frac{\partial u_\varepsilon}{\partial x}(t,x).$$

By (2.1), $q_\varepsilon = q_\varepsilon(t,x)$ is the solution of

$$\frac{\partial q_\varepsilon}{\partial t} + \gamma u_\varepsilon \frac{\partial q_\varepsilon}{\partial x} - \varepsilon\frac{\partial^2 q_\varepsilon}{\partial x^2} + \frac{\gamma}{2}q_\varepsilon^2 = h(u_\varepsilon) - P_\varepsilon, \quad q_\varepsilon(0,x) = \frac{\partial u_{\varepsilon,0}}{\partial x}(x), \tag{2.5}$$

for $t > 0$ and $x \in \mathbb{R}$. Multiply (2.1) by u_ε, (2.5) by q_ε, and add the resulting equations. After rearranging a bit, we derive the conservation law

$$\begin{aligned}\frac{\partial}{\partial t}\left(\frac{1}{2}\left[u_\varepsilon^2 + q_\varepsilon^2\right]\right) &+ \frac{\partial}{\partial x}\left(u_\varepsilon\left[\frac{\gamma}{2}q_\varepsilon^2 + P_\varepsilon\right] + \frac{\gamma}{3}u^3 - H(u)\right)\\ &= \frac{\varepsilon}{2}(u_\varepsilon^2 + q_\varepsilon^2)_{xx} - \varepsilon q_\varepsilon^2 - \varepsilon\left(\frac{\partial q_\varepsilon}{\partial x}\right)^2,\end{aligned}$$

where $H' = h$. From this (2.3) follows easily. ∎

3. Viscous approximants: Oleinik type estimate.

LEMMA 3.1. *For each $t > 0$ and $x \in \mathbb{R}$,*

$$\frac{\partial u_\varepsilon}{\partial x}(t,x) \leq \frac{2}{\gamma t} + C_2, \tag{3.1}$$

where $u_\varepsilon = u_\varepsilon(t,x)$ is the unique solution of (2.1), and

$$C_2 := \sqrt{\frac{2}{\gamma}}\left(2\max_{|\xi|\leq\sqrt{2}\|u_0\|_{H^1(\mathbb{R})}}\left|h(\xi)\right| + \frac{\gamma}{2}\|u_0\|^2_{H^1(\mathbb{R})}\right)^{1/2}.$$

Proof. From (2.4),

$$\left\|h(u_\varepsilon)\right\|_{L^\infty([0,\infty)\times\mathbb{R})} \leq \max_{|\xi|\leq\sqrt{2}\|u_0\|_{H^1(\mathbb{R})}}\left|h(\xi)\right| := L_1 < \infty. \tag{3.2}$$

Moreover, since

$$\int_{\mathbb{R}} e^{-|x-y|}dy = 2, \qquad x \in \mathbb{R}, \tag{3.3}$$

again using (2.4), for each $t \geq 0$ and $x \in \mathbb{R}$,

$$|P_\varepsilon(t,x)| \leq L_1 + \frac{\gamma}{4}\left\|\frac{\partial u_\varepsilon}{\partial x}(t,\cdot)\right\|^2_{L^2(\mathbb{R})} \leq L_1 + \frac{\gamma}{4}\|u_0\|^2_{H^1(\mathbb{R})} := L_2.$$

So, denoting $L := L_1 + L_2$, we have, from (2.5),

$$\frac{\partial q_\varepsilon}{\partial t} + \gamma u_\varepsilon \frac{\partial q_\varepsilon}{\partial x} - \varepsilon \frac{\partial^2 q_\varepsilon}{\partial x^2} + \frac{\gamma}{2} q_\varepsilon^2 \leq L. \tag{3.4}$$

Let $f = f(t)$ be the solution of

$$\frac{df}{dt} + \frac{\gamma}{2} f^2 = L, \quad t > 0, \qquad f(0) = \left\|\frac{\partial u_{\varepsilon,0}}{\partial x}\right\|_{L^\infty(\mathbb{R})}. \tag{3.5}$$

Since, by (2.4) and (3.4), $f = f(t)$ is a supersolution of the parabolic initial value problem (2.5), due to the comparison principle for parabolic equations, we get

$$q_\varepsilon(t,x) \leq f(t), \qquad t \geq 0,\ \ x \in \mathbb{R}. \tag{3.6}$$

Finally, consider the map $F(t) := \frac{2}{\gamma t} + \sqrt{\frac{2}{\gamma} L}$, $t > 0$. Observe that $\frac{dF}{dt}(t) + \frac{\gamma}{2} F^2(t) - L = \frac{2\sqrt{2L/\gamma}}{t} > 0$, for any $t > 0$, so that $F = F(t)$ is a supersolution of (3.5). Due to the comparison principle for ordinary differential equations, we get $f(t) \leq F(t)$ for all $t > 0$. Therefore, by this and (3.6), the estimate (3.1) is proven. □

4. Viscous approximants: Higher integrability estimate.

LEMMA 4.1. *Let $0 < \alpha < 1$, $T > 0$, and $a, b \in \mathbb{R}$, $a < b$. Then there exists a positive constant C_3 depending only on $\|u_0\|_{H^1(\mathbb{R})}$, α, T, a and b, but independent of ε, such that*

$$\int_0^T \int_a^b \left|\frac{\partial u_\varepsilon}{\partial x}(t,x)\right|^{2+\alpha} dtdx \leq C_3, \tag{4.1}$$

where $u_\varepsilon = u_\varepsilon(t,x)$ is the unique solution of (2.1).

Proof. The proof is a variant of the proof found in Xin and Zhang [32]. Let $\chi \in C^\infty(\mathbb{R})$ be a cut-off function such that

$$0 \leq \chi \leq 1, \qquad \chi(x) = \begin{cases} 1, & \text{if } x \in [a,b], \\ 0, & \text{if } x \in (-\infty, a-1] \cup [b+1, \infty). \end{cases}$$

Consider also the map $\theta(\xi) := \xi\big(|\xi| + 1\big)^\alpha$, $\xi \in \mathbb{R}$, and observe that, since $0 < \alpha < 1$,

$$\begin{aligned} \theta'(\xi) &= \big((\alpha+1)|\xi| + 1\big)\big(|\xi|+1\big)^{\alpha-1}, \\ \theta''(\xi) &= \alpha \operatorname{sign}(\xi)\, \big(|\xi|+1\big)^{\alpha-2}\big((\alpha+1)|\xi| + 2\big) \\ &= \alpha(\alpha+1)\operatorname{sign}(\xi)\, \big(|\xi|+1\big)^{\alpha-1} + (1-\alpha)\alpha \operatorname{sign}(\xi)\, \big(|\xi|+1\big)^{\alpha-2}, \end{aligned}$$

$$\big|\theta(\xi)\big| \leq |\xi|^{\alpha+1} + |\xi|, \quad \big|\theta'(\xi)\big| \leq (\alpha+1)|\xi| + 1, \quad \big|\theta''(\xi)\big| \leq 2\alpha, \tag{4.2}$$

(4.3)

$$\begin{aligned} \xi\theta(\xi) - \frac{1}{2}\xi^2\theta'(\xi) &= \frac{1-\alpha}{2}\xi^2\big(|\xi|+1\big)^\alpha + \frac{\alpha}{2}\xi^2\big(|\xi|+1\big)^{\alpha-1} \\ &\geq \frac{1-\alpha}{2}\xi^2\big(|\xi|+1\big)^\alpha. \end{aligned}$$

Multiplying (2.5) by $\chi\theta'(q_\varepsilon)$, using the chain rule, and integrating over $\Pi_T := [0,T]\times\mathbb{R}$, we get

$$
\begin{aligned}
(4.4)\qquad &\int_{\Pi_T} \gamma\chi(x) q_\varepsilon \theta(q_\varepsilon)dtdx - \frac{\gamma}{2}\int_{\Pi_T} q_\varepsilon^2 \chi(x)\theta'(q_\varepsilon)dtdx\\
&= \int_{\mathbb{R}} \chi(x)\left(\theta\big(q_\varepsilon(T,x)\big) - \theta\big(q_\varepsilon(0,x)\big)\right)dx - \int_{\Pi_T}\gamma u_\varepsilon \chi'(x)\theta(q_\varepsilon)dtdx\\
&+ \ \varepsilon\int_{\Pi_T} \frac{\partial q_\varepsilon}{\partial x}\chi'(x)\theta'(q_\varepsilon)dtdx + \varepsilon\int_{\Pi_T}\left(\frac{\partial q_\varepsilon}{\partial x}\right)^2\chi(x)\theta''(q_\varepsilon)dtdx\\
&- \int_{\Pi_T}\left(h(u_\varepsilon) - P_\varepsilon\right)\chi(x)\theta'(q_\varepsilon)dtdx.
\end{aligned}
$$

Observe that, by (4.3),

$$
\begin{aligned}
&\int_{\Pi_T} \gamma\chi(x) q_\varepsilon \theta(q_\varepsilon)dtdx - \frac{\gamma}{2}\int_{\Pi_T} q_\varepsilon^2 \chi(x)\theta'(q_\varepsilon)dtdx\\
&= \int_{\Pi_T}\gamma\chi(x)\Big(q_\varepsilon\theta(q_\varepsilon) - \frac{1}{2}q_\varepsilon^2\theta'(q_\varepsilon)\Big)dtdx\\
(4.5)\qquad &\geq \frac{\gamma(1-\alpha)}{2}\int_{\Pi_T}\chi(x) q_\varepsilon^2\big(|q_\varepsilon|+1\big)^\alpha dtdx.
\end{aligned}
$$

Let $t \geq 0$, since $0 < \alpha < 1$, using the Hölder inequality, (2.4) and the first part of (4.2),

$$
\begin{aligned}
(4.6)\qquad \left|\int_{\mathbb{R}}\chi(x)\theta(q_\varepsilon)dx\right| &\leq \int_{\mathbb{R}}\chi(x)\left(|q_\varepsilon|^{\alpha+1} + |q_\varepsilon|\right)dx\\
&\leq \|\chi\|_{L^{2/(1-\alpha)}(\mathbb{R})}\|q_\varepsilon(t,\cdot)\|_{L^2(\mathbb{R})}^{\alpha+1} + \|\chi\|_{L^2(\mathbb{R})}\|q_\varepsilon(t,\cdot)\|_{L^2(\mathbb{R})}\\
&\leq (b-a+2)^{(1-\alpha)/2}\|u_0\|_{H^1(\mathbb{R})}^{\alpha+1} + (b-a+2)^{1/2}\|u_0\|_{H^1(\mathbb{R})},
\end{aligned}
$$

and

(4.7)
$$
\begin{aligned}
\left|\int_{\Pi_T}\gamma u_\varepsilon\chi'(x)\theta(q_\varepsilon)dtdx\right| &\leq \int_{\Pi_T}\gamma|u_\varepsilon||\chi'(x)|\left(|q_\varepsilon|^{\alpha+1}+|q_\varepsilon|\right)dtdx\\
&\leq \int_{\Pi_T}\gamma\|u_\varepsilon(t,\cdot)\|_{L^\infty(\mathbb{R})}|\chi'(x)|\left(|q_\varepsilon|^{\alpha+1}+|q_\varepsilon|\right)dtdx\\
&\leq \gamma\frac{\|u_0\|_{H^1(\mathbb{R})}}{\sqrt{2}}\int_0^T\big(\|\chi'\|_{L^{2/(1-\alpha)}(\mathbb{R})}\|q_\varepsilon(t,\cdot)\|_{L^2(\mathbb{R})}^{\alpha+1}\\
&\qquad + \|\chi'\|_{L^2(\mathbb{R})}\|q_\varepsilon(t,\cdot)\|_{L^2(\mathbb{R})}\big)dt\\
&\leq \gamma T\frac{\|u_0\|_{H^1(\mathbb{R})}}{\sqrt{2}}\big(\|\chi'\|_{L^{2/(1-\alpha)}(\mathbb{R})}\|u_0\|_{H^1(\mathbb{R})}^{\alpha+1}\\
&\qquad + \|\chi'\|_{L^2(\mathbb{R})}\|u_0\|_{H^1(\mathbb{R})}\big).
\end{aligned}
$$

Moreover, observe that

$$
\varepsilon\int_{\Pi_T}\frac{\partial q_\varepsilon}{\partial x}\chi'(x)\theta'(q_\varepsilon)dtdx = -\varepsilon\int_{\Pi_T}\theta(q_\varepsilon)\chi''(x)dtdx,
$$

so, again by the Hölder inequality (2.4) and the first part of (4.2),

$$\begin{aligned}\left|\varepsilon\int_{\Pi_T}\frac{\partial q_\varepsilon}{\partial x}\chi'(x)\theta(q_\varepsilon)dtdx\right| &\le \varepsilon\int_{\Pi_T}|\theta(q_\varepsilon)||\chi''(x)|dtdx\\ &\le \varepsilon\int_{\Pi_T}\left(|q_\varepsilon|^{\alpha+1}+|q_\varepsilon|\right)|\chi''(x)|dtdx\\ &\le \varepsilon\int_0^T\Big(\|\chi''\|_{L^{2/(1-\alpha)}(\mathbb{R})}\|q_\varepsilon(t,\,\cdot\,)\|^{\alpha+1}_{L^2(\mathbb{R})}\\ &\qquad+\|\chi''\|_{L^2(\mathbb{R})}\|q_\varepsilon(t,\,\cdot\,)\|_{L^2(\mathbb{R})}\Big)dt\\ &\le \varepsilon T\Big(\|\chi''\|_{L^{2/(1-\alpha)}(\mathbb{R})}\|u_0\|^{\alpha+1}_{H^1(\mathbb{R})}+\|\chi''\|_{L^2(\mathbb{R})}\|u_0\|_{H^1(\mathbb{R})}\Big).\end{aligned} \tag{4.8}$$

Since $0<\alpha<1$, using (2.3) and the third part of (4.2),

$$\varepsilon\left|\int_{\Pi_T}\left(\frac{\partial q_\varepsilon}{\partial x}\right)^2\chi(x)\theta''(q_\varepsilon)dtdx\right|\le 2\alpha\varepsilon\int_{\Pi_T}\left(\frac{\partial q_\varepsilon}{\partial x}\right)^2dtdx\le\alpha\|u_0\|^2_{H^1(\mathbb{R})}. \tag{4.9}$$

As we showed in the proof of Lemma 3.1, there exists a constant $L>0$ depending only on $\|u_0\|_{H^1(\mathbb{R})}$ such that $\left\|h(u_\varepsilon)-P_\varepsilon\right\|_{L^\infty([0,\infty)\times\mathbb{R})}\le L$, so, since $0<\alpha<1$, using the second part of (4.2),

$$\begin{aligned}&\left|\int_{\Pi_T}\left(h(u_\varepsilon)-P_\varepsilon\right)\chi(x)\theta'(q_\varepsilon)dtdx\right|\\ &\quad\le L\int_{\Pi_T}\chi(x)\left((\alpha+1)|q_\varepsilon|+1\right)dtdx\\ &\quad\le L\int_0^T\left((\alpha+1)\|\chi\|_{L^2(\mathbb{R})}\|q_\varepsilon(t,\,\cdot\,)\|_{L^2(\mathbb{R})}+\|\chi\|_{L^1(\mathbb{R})}\right)dt\\ &\quad\le LT\Big((\alpha+1)(b-a+2)^{1/2}\|u_0\|_{H^1(\mathbb{R})}+(b-a+2)\Big).\end{aligned} \tag{4.10}$$

From (4.4)–(4.10), there exists a constant $c>0$ depending only on $\|u_0\|_{H^1(\mathbb{R})}$, α, $T>0$, a, and b, but independent of ε, such that

$$\frac{\gamma(1-\alpha)}{2}\int_{\Pi_T}|q_\varepsilon|^2\chi(x)\big(|q_\varepsilon|+1\big)^\alpha dtdx\le c. \tag{4.11}$$

Then

$$\int_0^T\int_a^b\left|\frac{\partial u_\varepsilon}{\partial x}(t,x)\right|^{2+\alpha}dtdx\le\int_{\Pi_T}|q_\varepsilon|\chi(x)\left(|q_\varepsilon|+1\right)^{\alpha+1}dtdx\le\frac{2c}{\gamma(1-\alpha)},$$

hence estimate (4.1) is proved. ∎

5. Viscous approximants: Basic compactness.

LEMMA 5.1. *There exists a positive constant C_4 depending only on $\|u_0\|_{H^1(\mathbb{R})}$ such that*

$$\|P_\varepsilon(t,\,\cdot\,)\|_{L^\infty(\mathbb{R})},\ \|P_\varepsilon(t,\,\cdot\,)\|_{L^2(\mathbb{R})},\ \left\|\frac{\partial P_\varepsilon}{\partial x}(t,\,\cdot\,)\right\|_{L^\infty(\mathbb{R})},\ \left\|\frac{\partial P_\varepsilon}{\partial x}(t,\,\cdot\,)\right\|_{L^2(\mathbb{R})}\le C_4, \tag{5.1}$$

where $u_\varepsilon = u_\varepsilon(t,x)$ is the unique solution of (2.1). In particular, $\{P_\varepsilon\}_\varepsilon$ is uniformly bounded in $L^\infty([0,\infty); W^{1,\infty}(\mathbb{R}))$ and $L^\infty([0,\infty); H^1(\mathbb{R}))$.

Proof. Define

$$P_{1,\varepsilon}(t,x) := \frac{\gamma}{4}\int_{\mathbb{R}} e^{-|x-y|} q_\varepsilon^2 dy, \qquad P_{2,\varepsilon}(t,x) := \frac{1}{2}\int_{\mathbb{R}} e^{-|x-y|} h\big(u_\varepsilon(t,y)\big) dy, \tag{5.2}$$

and notice that $P_\varepsilon = P_{1,\varepsilon} + P_{2,\varepsilon}$. By (2.4) and (3.3),

$$P_{1,\varepsilon}(t,x)| \le \frac{\gamma}{4}\|u_\varepsilon(t,\cdot)\|^2_{H^1(\mathbb{R})} \le \frac{\gamma}{4}\|u_0\|^2_{H^1(\mathbb{R})}, \tag{5.3}$$

$$|P_{2,\varepsilon}(t,x)| \le \max_{|\xi| \le \|u_0\|_{H^1(\mathbb{R})}/\sqrt{2}} \big|h(\xi)\big|. \tag{5.4}$$

Moreover, using (3.3) and the Tonelli theorem,

$$\int_{\mathbb{R}} |P_{1,\varepsilon}(t,x)| dx \le \frac{\gamma}{2}\|u_\varepsilon(t,\cdot)\|^2_{H^1(\mathbb{R})} \le \frac{\gamma}{2}\|u_0\|^2_{H^1(\mathbb{R})}. \tag{5.5}$$

From (3.3), (5.3), (5.5), and the Hölder inequality,

$$\int_{\mathbb{R}} |P_{1,\varepsilon}(t,x)|^2 dx \le \|P_{1,\varepsilon}\|_{L^\infty([0,\infty)\times\mathbb{R})}\|P_{1,\varepsilon}(t,\cdot)\|_{L^1(\mathbb{R})} \le \frac{\gamma^2}{8}\|u_0\|^4_{H^1(\mathbb{R})},$$

so that

$$\|P_{1,\varepsilon}(t,\cdot)\|_{L^2(\mathbb{R})} \le \frac{\gamma}{2\sqrt{2}}\|u_0\|^2_{H^1(\mathbb{R})}. \tag{5.6}$$

Using (1.4), (2.4), (3.3), the Tonelli theorem, and the Hölder inequality,

$$\begin{aligned}
\int_{\mathbb{R}} |P_{2,\varepsilon}(t,x)|^2 dx &\le \frac{1}{2}\int_{\mathbb{R}} \left(\int_{\mathbb{R}} e^{-|x-y|} dx\right) \left(h(u_\varepsilon(t,y))\right)^2 dy \\
&\le \left(\max_{|\xi| \le \|u_0\|_{H^1(\mathbb{R})}/\sqrt{2}} \left(h'(\xi)\right)^2\right) \int_{\mathbb{R}} u_\varepsilon^2(t,y) dy \\
&\le \left(\max_{|\xi| \le \|u_0\|_{H^1(\mathbb{R})}/\sqrt{2}} \left(h'(\xi)\right)^2\right) \|u_0\|^2_{H^1(\mathbb{R})}.
\end{aligned} \tag{5.7}$$

Finally, observing

$$\begin{aligned}
\frac{\partial P_{1,\varepsilon}}{\partial x}(t,x) &= \frac{\gamma}{4}\int_{\mathbb{R}} \operatorname{sign}(y-x)\, e^{-|x-y|} \left(q_\varepsilon(t,y)\right)^2 dy, \\
\frac{\partial P_{2,\varepsilon}}{\partial x}(t,x) &= \frac{1}{2}\int_{\mathbb{R}} \operatorname{sign}(y-x)\, e^{-|x-y|} h\big(u_\varepsilon(t,y)\big) dy,
\end{aligned}$$

and recalling $P_\varepsilon = P_{1,\varepsilon} + P_{2,\varepsilon}$, the claim is a direct consequence of (5.3), (5.4), (5.6), and (5.7). ◻

LEMMA 5.2. *There exists a sequence $\{\varepsilon_j\}_{j\in\mathbb{N}}$ tending to zero and a function $u \in L^\infty([0,\infty); H^1(\mathbb{R})) \cap H^1([0,T]\times\mathbb{R})$, for each $T \ge 0$, such that*

$$u_{\varepsilon_j} \rightharpoonup u \quad \text{in } H^1([0,T]\times\mathbb{R}), \text{ for each } T \ge 0, \tag{5.8}$$

$$u_{\varepsilon_j} \to u \quad \text{in } L^\infty_{\mathrm{loc}}([0,\infty)\times\mathbb{R}), \tag{5.9}$$

where $u_\varepsilon = u_\varepsilon(t,x)$ is the unique solution of (2.1).

Proof. Fix $T > 0$. Observe that, from (2.1), $\frac{\partial u_\varepsilon}{\partial t} = \varepsilon \frac{\partial^2 u_\varepsilon}{\partial x^2} - \gamma u_\varepsilon \frac{\partial u_\varepsilon}{\partial x} - \frac{\partial P_\varepsilon}{\partial x}$, so, by (2.4), (2.3), (5.1), and the Hölder inequality,

$$\left\| \frac{\partial u_\varepsilon}{\partial t} \right\|_{L^2([0,T]\times\mathbb{R})} \le \sqrt{\frac{\varepsilon}{2}} \|u_0\|_{L^\infty(\mathbb{R})} + \frac{\gamma\sqrt{T}}{\sqrt{2}} \|u_0\|^2_{L^\infty(\mathbb{R})} + C_4\sqrt{T}. \tag{5.10}$$

Hence $\{u_\varepsilon\}$ is uniformly bounded in $H^1([0,T]\times\mathbb{R}) \cap L^\infty([0,\infty); H^1(\mathbb{R}))$, and (5.8) follows.

Observe that, for each $0 \le s,\, t \le T$,

$$\|u_\varepsilon(t,\cdot) - u_\varepsilon(s,\cdot)\|^2_{L^2(\mathbb{R})} = \int_{\mathbb{R}} \left(\int_s^t \frac{\partial u_\varepsilon}{\partial t}(\tau, x) d\tau \right)^2 dx \le \sqrt{|t-s|} \int_{\Pi_T} \left(\frac{\partial u_\varepsilon}{\partial t}(\tau, x) \right)^2 d\tau dx.$$

Moreover, $\{u_\varepsilon\}$ is uniformly bounded in $L^\infty([0,T]; H^1(\mathbb{R}))$ and $H^1(\mathbb{R}) \subset\subset L^\infty_{\mathrm{loc}}(\mathbb{R}) \subset L^2_{\mathrm{loc}}(\mathbb{R})$, then (5.9) is consequence of [31, Theorem 5]. $\square$

LEMMA 5.3. *The sequence $\{P_\varepsilon\}_\varepsilon$ is uniformly bounded in $W^{1,1}_{\mathrm{loc}}([0,\infty)\times\mathbb{R})$. In particular, there exists a sequence $\{\varepsilon_j\}_{j\in\mathbb{N}}$ tending to zero and a function $P \in L^\infty([0,\infty); W^{1,\infty}(\mathbb{R}))$ such that for each $1 < p < \infty$*

$$P_{\varepsilon_j} \to P \quad \textit{strongly in } L^p_{\mathrm{loc}}([0,\infty)\times\mathbb{R}). \tag{5.11}$$

Proof. We begin by proving that $\left\{\frac{\partial P_\varepsilon}{\partial t}\right\}_\varepsilon$ is uniformly bounded in $L^1_{\mathrm{loc}}([0,\infty)\times\mathbb{R})$. Fix $T > 0$. We claim that

$$\left\{ \frac{\partial P_{1,\varepsilon}}{\partial t} \right\}_\varepsilon \text{ is uniformly bounded in } L^1([0,T]\times\mathbb{R}), \tag{5.12}$$

$$\left\{ \frac{\partial P_{2,\varepsilon}}{\partial t} \right\}_\varepsilon \text{ is uniformly bounded in } L^2([0,T]\times\mathbb{R}), \tag{5.13}$$

where $P_{1,\varepsilon}$ and $P_{2,\varepsilon}$ are defined in (5.2). We begin by proving (5.12). Observe that, from (2.5),

(5.14)
$$\begin{aligned} \frac{\partial P_{1,\varepsilon}}{\partial t}(t,x) &= \frac{\gamma}{2} \int_{\mathbb{R}} e^{-|x-y|} q_\varepsilon \frac{\partial q_\varepsilon}{\partial t} dy \\ &= \frac{\gamma}{2} \int_{\mathbb{R}} e^{-|x-y|} \left(-\gamma q_\varepsilon u_\varepsilon \frac{\partial q_\varepsilon}{\partial x} + \varepsilon q_\varepsilon \frac{\partial^2 q_\varepsilon}{\partial x^2} - \frac{\gamma}{2} q_\varepsilon^3 + q_\varepsilon \left(h(u_\varepsilon) - P_\varepsilon \right) \right) dy. \end{aligned}$$

Using $\frac{\gamma}{2}\frac{\partial}{\partial x}(u_\varepsilon q_\varepsilon^2) = \frac{\gamma}{2} q_\varepsilon^3 + \gamma q_\varepsilon u_\varepsilon \frac{\partial q_\varepsilon}{\partial x}$, $\frac{\partial}{\partial x}\left(q_\varepsilon \frac{\partial q_\varepsilon}{\partial x}\right) = q_\varepsilon \frac{\partial^2 q_\varepsilon}{\partial x^2} + \left(\frac{\partial q_\varepsilon}{\partial x}\right)^2$, (5.14), and integration by parts, we get

$$\begin{aligned} &\frac{\partial P_{1,\varepsilon}}{\partial t}(t,x) \\ &\quad = \frac{\gamma}{4} \int_{\mathbb{R}} e^{-|x-y|} \left(-\frac{\gamma}{2} \frac{\partial}{\partial x}(u_\varepsilon q_\varepsilon^2) + \varepsilon \frac{\partial}{\partial x}\left(q_\varepsilon \frac{\partial q_\varepsilon}{\partial x} \right) - \varepsilon \left(\frac{\partial q_\varepsilon}{\partial x} \right)^2 + q_\varepsilon \left(h(u_\varepsilon) - P_\varepsilon \right) \right) dy \\ &\quad = \frac{\gamma}{4} \int_{\mathbb{R}} e^{-|x-y|} \left(\operatorname{sign}(y-x) \left[\frac{\gamma}{2} u_\varepsilon q_\varepsilon^2 - \varepsilon q_\varepsilon \frac{\partial q_\varepsilon}{\partial x} \right] - \varepsilon \left(\frac{\partial q_\varepsilon}{\partial x} \right)^2 + q_\varepsilon \left(h(u_\varepsilon) - P_\varepsilon \right) \right) dy. \end{aligned}$$

Using (1.4), (2.3), (2.4), (5.1), the Tonelli theorem, and the Hölder inequality,

$$\int_{\mathbb{R}\times\mathbb{R}} e^{-|x-y|}|u_\varepsilon|q_\varepsilon^2 dxdy \le \sqrt{2}\|u_0\|_{H^1(\mathbb{R})}\|u_\varepsilon(t,\cdot)\|^2_{H^1(\mathbb{R})} \le \sqrt{2}\|u_0\|^3_{H^1(\mathbb{R})},$$

$$\begin{aligned}\varepsilon\int_{\Pi_T\times\mathbb{R}} e^{-|x-y|}|q_\varepsilon|\left|\frac{\partial q_\varepsilon}{\partial x}\right| dtdxdy &\le \varepsilon\int_0^T \|u_\varepsilon(t,\cdot)\|^2_{H^1(\mathbb{R})}dt + \varepsilon\int_0^T \left\|\frac{\partial u_\varepsilon}{\partial x}(t,\cdot)\right\|^2_{H^1(\mathbb{R})} dt\\ &\le \left(\varepsilon T + \frac{1}{2}\right)\|u_0\|^2_{H^1(\mathbb{R})},\end{aligned}$$

$$\varepsilon\int_{\Pi_T\times\mathbb{R}} e^{-|x-y|}\left(\frac{\partial q_\varepsilon}{\partial x}\right)^2 dtdxdy \le 2\varepsilon\int_0^T \left\|\frac{\partial u_\varepsilon}{\partial x}(t,\cdot)\right\|^2_{H^1(\mathbb{R})} dt \le \|u_0\|^2_{H^1(\mathbb{R})},$$

$$\begin{aligned}\int_{\mathbb{R}\times\mathbb{R}} e^{-|x-y|}|q_\varepsilon||h(u_\varepsilon)|dxdy &\le \int_{\mathbb{R}} q_\varepsilon^2 dy + \max_{|\xi|\le\|u_0\|_{H^1(\mathbb{R})}/\sqrt{2}} (h'(\xi))^2\int_{\mathbb{R}} u_\varepsilon^2 dy\\ &\le \left(1 + \max_{|\xi|\le\|u_0\|_{H^1(\mathbb{R})}/\sqrt{2}} (h'(\xi))^2\right)\|u_0\|^2_{H^1(\mathbb{R})},\end{aligned}$$

$$\int_{\mathbb{R}\times\mathbb{R}} e^{-|x-y|}|q_\varepsilon||P_\varepsilon|dxdy \le \|u_\varepsilon(t,\cdot)\|^2_{H^1(\mathbb{R})} + \|P_\varepsilon(t,\cdot)\|^2_{L^2(\mathbb{R})} \le \|u_0\|^2_{H^1(\mathbb{R})} + C_4^2.$$

It follows from these estimates that (5.12) holds.

We continue by proving (5.13). Observe that

$$\frac{\partial P_{2,\varepsilon}}{\partial t}(t,x) = \frac{1}{2}\int_{\mathbb{R}} e^{-|x-y|}h'(u_\varepsilon)\frac{\partial u_\varepsilon}{\partial t}dy, \tag{5.15}$$

so, using (1.4), (2.4), the Tonelli theorem, and the Hölder inequality,

$$\left\|\frac{\partial P_{2,\varepsilon}}{\partial t}\right\|^2_{L^2(\Pi_T)} \le \max_{|\xi|\le\|u_0\|_{H^1(\mathbb{R})}/\sqrt{2}} (h'(\xi))^2\left\|\frac{\partial u_\varepsilon}{\partial t}\right\|^2_{L^2(\Pi_T)}. \tag{5.16}$$

Then (5.13) is a direct consequence of (5.10).

Since the bound on $\left\{\frac{\partial P_\varepsilon}{\partial t}\right\}_\varepsilon$ is a consequence of (5.12) and (5.13), the family $\{P_\varepsilon\}_\varepsilon$ is bounded in $W^{1,1}_{\mathrm{loc}}([0,\infty)\times\mathbb{R})$.

Finally, using also Lemma 5.1, we have the existence of a pointwise converging subsequence that is uniformly bounded in $L^\infty([0,\infty)\times\mathbb{R})$. Clearly, this implies (5.11). $\square$

Throughout this paper we use overbars to denote weak limits (the spaces in which these weak limits are taken should be clear from the context and thus they are not always explicitly stated).

LEMMA 5.4. *There exists a sequence $\{\varepsilon_j\}_{j\in\mathbb{N}}$ tending to zero and two functions $q \in L^p_{\mathrm{loc}}([0,\infty)\times\mathbb{R})$, $\overline{q^2} \in L^r_{\mathrm{loc}}([0,\infty)\times\mathbb{R})$ such that*

$$q_{\varepsilon_j} \rightharpoonup q \quad \text{in } L^p_{\mathrm{loc}}([0,\infty)\times\mathbb{R}), \qquad q_{\varepsilon_j} \overset{\star}{\rightharpoonup} q \quad \text{in } L^\infty_{\mathrm{loc}}([0,\infty); L^2(\mathbb{R})), \tag{5.17}$$

$$q^2_{\varepsilon_j} \rightharpoonup \overline{q^2} \quad \text{in } L^r_{\mathrm{loc}}([0,\infty)\times\mathbb{R}), \tag{5.18}$$

for each $1 < p < 3$ and $1 < r < \frac{3}{2}$. Moreover,

$$q^2(t,x) \le \overline{q^2}(t,x) \quad \text{for almost every } (t,x) \in [0,\infty)\times\mathbb{R} \tag{5.19}$$

and

$$\frac{\partial u}{\partial x} = q \quad \text{in the sense of distributions on } [0,\infty)\times\mathbb{R}. \tag{5.20}$$

Proof. Formulas (5.17) and (5.18) are direct consequences of Theorem 2.1 and Lemma 4.1. Inequality (5.19) is true thanks to the weak convergence in (5.18). Finally, (5.20) is a consequence of the definition of q_ε, Lemma 5.2, and (5.17). ∎

In the following, for notational convenience, we replace the sequences $\{u_{\varepsilon_j}\}_{j\in\mathbb{N}}$, $\{q_{\varepsilon_j}\}_{j\in\mathbb{N}}$, $\{P_{\varepsilon_j}\}_{j\in\mathbb{N}}$ by $\{u_\varepsilon\}_{\varepsilon>0}$, $\{q_\varepsilon\}_{\varepsilon>0}$, $\{P_\varepsilon\}_{\varepsilon>0}$, respectively.

In view of (5.17), we conclude that for any $\eta\in C^1(\mathbb{R})$ with η' bounded, Lipschitz continuous on $\mathbb{R}$ and any $1<p<3$ we have

$$\eta(q_\varepsilon) \rightharpoonup \overline{\eta(q)} \quad \text{in } L^p_{\mathrm{loc}}([0,\infty)\times\mathbb{R}), \qquad \eta(q_\varepsilon) \overset{\star}{\rightharpoonup} \overline{\eta(q)} \quad \text{in } L^\infty_{\mathrm{loc}}([0,\infty);L^2(\mathbb{R})). \tag{5.21}$$

Multiplying the equation in (2.5) by $\eta'(q_\varepsilon)$, we get

$$\begin{aligned} &\frac{\partial}{\partial t}\eta(q_\varepsilon) + \frac{\partial}{\partial x}\left(\gamma u_\varepsilon \eta(q_\varepsilon)\right) - \varepsilon\frac{\partial^2}{\partial x^2}\eta(q_\varepsilon) - \varepsilon\eta''(q_\varepsilon)\left(\frac{\partial}{\partial x}\eta(q_\varepsilon)\right)^2 \\ &\quad = \gamma q_\varepsilon \eta(q_\varepsilon) - \frac{\gamma}{2}\eta'(q_\varepsilon)q_\varepsilon^2 + \left(h(u_\varepsilon)-P_\varepsilon\right)\eta'(q_\varepsilon). \end{aligned} \tag{5.22}$$

Lemma 5.5. *For any convex $\eta\in C^1(\mathbb{R})$ with η' bounded, Lipschitz continuous on $\mathbb{R}$, we have*

$$\frac{\partial\overline{\eta(q)}}{\partial t} + \frac{\partial}{\partial x}\left(\gamma u\overline{\eta(q)}\right) \le \gamma\overline{q\eta(q)} - \frac{\gamma}{2}\overline{\eta'(q)q^2} + (h(u)-P)\overline{\eta'(q)}, \tag{5.23}$$

in the sense of distributions on $[0,\infty)\times\mathbb{R}$. Here $\overline{q\eta(q)}$ and $\overline{\eta'(q)q^2}$ denote the weak limits of $q_\varepsilon\eta(q_\varepsilon)$ and $\eta'(q_\varepsilon)q_\varepsilon^2$ in $L^r_{\mathrm{loc}}([0,\infty)\times\mathbb{R})$, $1<r<\frac{3}{2}$, respectively.

Proof. In (5.22), by convexity of η, (1.4), (5.9), (5.17), and (5.18), sending $\varepsilon\to 0$ yields (5.23). ∎

Remark 5.6. From (5.17) and (5.18), it is clear that

$$q = q_+ + q_- = \overline{q_+} + \overline{q_-}, \quad q^2 = (q_+)^2 + (q_-)^2, \quad \overline{q^2} = \overline{(q_+)^2} + \overline{(q_-)^2},$$

almost everywhere in $[0,\infty)\times\mathbb{R}$, where $\xi_+ := \xi\chi_{[0,+\infty)}(\xi)$, $\xi_- := \xi\chi_{(-\infty,0]}(\xi)$, $\xi\in\mathbb{R}$. Moreover, by (3.1) and (5.17),

$$q_\varepsilon(t,x), q(t,x) \le \frac{2}{\gamma t} + C_2, \qquad t\ge 0,\ x\in\mathbb{R}. \tag{5.24}$$

Lemma 5.7. *There holds*

$$\frac{\partial q}{\partial t} + \frac{\partial}{\partial x}(\gamma u q) = \frac{\gamma}{2}\overline{q^2} + h(u) - P \quad \text{in the sense of distributions on } [0,\infty)\times\mathbb{R}. \tag{5.25}$$

Proof. Using (2.5), (5.9), (5.11), (5.17), and (5.18), the result (5.25) follows by $\varepsilon\to 0$ in (2.5). ∎

The next lemma contains a renormalized formulation of (5.25).

LEMMA 5.8. *For any $\eta \in C^1(\mathbb{R})$ with $\eta' \in L^\infty(\mathbb{R})$,*

$$\frac{\partial \eta(q)}{\partial t} + \frac{\partial}{\partial x}\left(\gamma u \eta(q)\right) = \gamma q \eta(q) + \left(\frac{\gamma}{2}\overline{q^2} - \gamma q^2\right)\eta'(q) + \left(h(u) - P\right)\eta'(q), \tag{5.26}$$

in the sense of distributions on $[0,\infty) \times \mathbb{R}$.

Proof. Let $\{\omega_\delta\}_\delta$ be a family of mollifiers defined on $\mathbb{R}$. Denote $q_\delta(t,x) := (q(t,\cdot) \star \omega_\delta)(x)$. Here and in the following all convolutions are with respect to the x variable. According to Lemma II.1 of [19], it follows from (5.25) that q_δ solves

$$\frac{\partial q_\delta}{\partial t} + \gamma u \frac{\partial q_\delta}{\partial x} = \frac{\gamma}{2}\overline{q^2} \star \omega_\delta - \gamma q^2 \star \omega_\delta + h(u) \star \omega_\delta - P \star \omega_\delta + \rho_\delta, \tag{5.27}$$

where the error ρ_δ tends to zero in $L^1_{\mathrm{loc}}([0,\infty) \times \mathbb{R})$. Multiplying (5.27) by $\eta'(q_\delta)$, we get

$$\begin{aligned} \frac{\partial \eta(q_\delta)}{\partial t} + \frac{\partial}{\partial x}\left(\gamma u \eta(q_\delta)\right) = q\eta(q_\delta) &+ \frac{\gamma}{2}\left(\overline{q^2} \star \omega_\delta\right)\eta'(q_\delta) - \gamma\left(q^2 \star \omega_\delta\right)\eta'(q_\delta) \\ &+ \left(h(u) \star \omega_\delta\right)\eta'(q_\delta) - \left(P \star \omega_\delta\right)\eta'(q_\delta) + \rho_\delta \eta'(q_\delta). \end{aligned} \tag{5.28}$$

Using the boundedness of η, η', we can send $\delta \to 0$ in (5.28) to obtain (5.26). The weak time continuity is standard. $\square$

6. Strong convergence of ${}_{\varepsilon}$ and existence for (1). Following [32], in this section we wish to improve the weak convergence of q_ε in (5.17) to strong convergence (and then we have an existence result for (1)). Roughly speaking, the idea is to derive a "transport equation" for the evolution of the defect measure $(\overline{q^2} - q^2)(t,\cdot) \geq 0$, so that if it is zero initially then it will continue to be zero at all later times $t > 0$. The proof is complicated by the fact that we do not have a uniform bound on q_ε from below but merely (5.24) and that in Lemma 4.1 we have only $\alpha < 1$.

LEMMA 6.1. *There holds*

$$\lim_{t\to 0+}\int_{\mathbb{R}} q^2(t,x)dx = \lim_{t\to 0+}\int_{\mathbb{R}} \overline{q^2}(t,x)dx = \int_{\mathbb{R}}\left(\frac{\partial u_0}{\partial x}\right)^2 dx. \tag{6.1}$$

Proof. Since $u \in C(\mathbb{R}_+ \times \mathbb{R})$ (see Lemma 5.2), from (5.20),

$$\begin{aligned} \lim_{t\to 0}\int_{\mathbb{R}} q(t,x)\varphi(x)dx &= -\lim_{t\to 0}\int_{\mathbb{R}} u(t,x)\frac{\partial \varphi}{\partial x}(x)dx \\ &= -\int_{\mathbb{R}} u_0(x)\frac{\partial \varphi}{\partial x}(x)dx = \int_{\mathbb{R}}\frac{\partial u_0}{\partial x}(x)\varphi(x)dx, \end{aligned}$$

for each test function $\varphi \in C^\infty(\mathbb{R})$ with compact support. Due to the boundedness of $\{q_\varepsilon\}_{\varepsilon>0}$ in $L^\infty((0,\infty);L^2(\mathbb{R}))$ we get

$$q(t,\cdot) \rightharpoonup \frac{\partial u_0}{\partial x} \qquad \text{weakly in } L^2(\mathbb{R}) \text{ as } t \to 0+,$$

so

$$\liminf_{t\to 0+}\int_{\mathbb{R}} q^2(t,x)dx \geq \int_{\mathbb{R}}\left(\frac{\partial u_0}{\partial x}(x)\right)^2 dx. \tag{6.2}$$

Moreover, from (2.2), (2.3), (5.9), and (5.18),

$$\int_{\mathbb{R}} u^2(t,x)dx + \int_{\mathbb{R}} \overline{q^2}(t,x)dx \le \int_{\mathbb{R}} u_0^2(x)dx + \int_{\mathbb{R}} \left(\frac{\partial u_0}{\partial x}\right)^2 dx,$$

and, again using the continuity of u (see Lemma 5.2), $\lim_{t\to 0+}\int_{\mathbb{R}} u^2(t,x)dx = \int_{\mathbb{R}} u_0^2 dx$. Hence

$$\limsup_{t\to 0+} \int_{\mathbb{R}} \overline{q^2}(t,x)dx \le \int_{\mathbb{R}} \left(\frac{\partial u_0}{\partial x}\right)^2 dx. \tag{6.3}$$

Clearly, (5.19), (6.2), and (6.3) imply (6.1). ◻

LEMMA 6.2. *For each $R>0$,*

$$\lim_{t\to 0+} \int_{\mathbb{R}} \left(\overline{\eta_R^{\pm}(q)}(t,x) - \eta_R^{\pm}(q(t,x))\right) dx = 0, \tag{6.4}$$

where

$$\eta_R(\xi) := \begin{cases} \frac{1}{2}\xi^2, & \text{if } |\xi| \le R, \\ R|\xi| - \frac{1}{2}R^2, & \text{if } |\xi| > R, \end{cases} \tag{6.5}$$

and $\eta_R^+(\xi) := \eta_R(\xi)\chi_{[0,+\infty)}(\xi)$, $\eta_R^-(\xi) := \eta_R(\xi)\chi_{(-\infty,0]}(\xi)$, $\xi \in \mathbb{R}$.

Proof. Let $R>0$. Observe that

$$\overline{\eta_R(q)} - \eta_R(q) = \frac{1}{2}(\overline{q^2} - q^2) - \left(\overline{f_R(q)} - f_R(q)\right),$$

where $f_R(\xi) := \frac{1}{2}\xi^2 - \eta_R(\xi)$, $\xi\in\mathbb{R}$. Since η_R and f_R are convex,

$$0 \le \overline{\eta_R(q)} - \eta_R(q) = \frac{1}{2}\left(\overline{q^2} - q^2\right) - \left(\overline{f_R(q)} - f_R(q)\right) \le \frac{1}{2}\left(\overline{q^2} - q^2\right).$$

Then, from (6.1), $\lim_{t\to 0+}\int_{\mathbb{R}}\left(\overline{\eta_R(q)}(t,x) - \eta_R(q(t,x))\right)dx = 0$. Since, $\overline{\eta_R^{\pm}(q)} - \eta_R^{\pm}(q) \le \overline{\eta_R(q)} - \eta_R(q)$, the proof is done. ◻

Remark 6.3. Let $R>0$. Then for each $\xi\in\mathbb{R}$

$$\begin{aligned}
\eta_R(\xi) &= \frac{1}{2}\xi^2 - \frac{1}{2}(R-|\xi|)^2\chi_{(-\infty,-R)\cup(R,\infty)}(\xi),\\
\eta_R'(\xi) &= \xi + (R-|\xi|)\operatorname{sign}(\xi)\,\chi_{(-\infty,-R)\cup(R,\infty)}(\xi),\\
\eta_R^+(\xi) &= \frac{1}{2}(\xi_+)^2 - \frac{1}{2}(R-\xi)^2\chi_{(R,\infty)}(\xi),\\
(\eta_R^+)'(\xi) &= \xi_+ + (R-\xi)\chi_{(R,\infty)}(\xi),\\
\eta_R^-(\xi) &= \frac{1}{2}(\xi_-)^2 - \frac{1}{2}(R+\xi)^2\chi_{(-\infty,-R)}(\xi),\\
(\eta_R^-)'(\xi) &= \xi_- - (R+\xi)\chi_{(-\infty,-R)}(\xi).
\end{aligned}$$

LEMMA 6.4. *Assume* (1.4) *and* (2.2). *Then for almost all $t\ge 0$*

$$\int_{\mathbb{R}} \left(\overline{(q_+)^2} - (q_+)^2\right)(t,x)dx \le 2\int_0^t\int_{\mathbb{R}} S(s,x)\left[\overline{q_+}(s,x) - q_+(s,x)\right]dsdx, \tag{6.6}$$

where $S(s,x) := h\big(u(s,x)\big) - P(s,x)$.

Proof. Let $R > C_2$ (see Lemma 3.1). Subtract (5.26) from (5.23) using the entropy η_R^+ (see Lemma 6.2). The result is

$$\begin{aligned}(6.7)\qquad &\frac{\partial}{\partial t}\left(\overline{\eta_R^+(q)}-\eta_R^+(q)\right)+\frac{\partial}{\partial x}\left(\gamma u\left[\overline{\eta_R^+(q)}-\eta_R^+(q)\right]\right)\\ &\quad\le \gamma\left[\overline{q\eta_R^+(q)}-q\eta_R^+(q)\right]-\frac{\gamma}{2}\left[\overline{q^2(\eta_R^+)'(q)}-q^2(\eta_R^+)'(q)\right]\\ &\quad\quad -\frac{\gamma}{2}\left(\overline{q^2}-q^2\right)(\eta_R^+)'(q)+S(t,x)\left[\overline{(\eta_R^+)'(q)}-(\eta_R^+)'(q)\right].\end{aligned}$$

Since η_R^+ is increasing and $\gamma \ge 0$, by (5.19),

$$(6.8)\qquad -\frac{\gamma}{2}\left(\overline{q^2}-q^2\right)(\eta_R^+)'(q)\le 0.$$

Moreover, from Remark 6.3,

$$\begin{aligned}\gamma q\eta_R^+(q)-\frac{\gamma}{2}q^2(\eta_R^+)'(q)&=-\frac{\gamma R}{2}q(R-q)\chi_{(R,\infty)}(q),\\ \overline{\gamma q\eta_R^+(q)}-\frac{\gamma}{2}\overline{q^2(\eta_R^+)'(q)}&=-\frac{\gamma R}{2}\overline{q(R-q)\chi_{(R,\infty)}(q)}.\end{aligned}$$

Therefore, due to (5.24),

$$(6.9)\qquad \gamma q\eta_R^+(q)-\frac{\gamma}{2}q^2(\eta_R^+)'(q)=\overline{q\eta_R^+(q)}-\frac{1}{2}\overline{q^2(\eta_R^+)'(q)}=0,\quad \text{in } \Omega_R:=\left(\tfrac{2}{R-C_2},\infty\right)\times\mathbb{R}.$$

Then from (6.7), (6.8), and (6.9) the following inequality holds in Ω_R:

$$(6.10)\qquad \frac{\partial}{\partial t}\left(\overline{\eta_R^+(q)}-\eta_R^+(q)\right)+\frac{\partial}{\partial x}\left(\gamma u\left[\overline{\eta_R^+(q)}-\eta_R^+(q)\right]\right)\le S(t,x)\left[\overline{(\eta_R^+)'(q)}-(\eta_R^+)'(q)\right].$$

In view of Remark 5.6 and due to (5.24),

$$\eta_R^+(q)=\frac{1}{2}(q_+)^2,\quad (\eta_R^+)'(q)=q_+,\quad \overline{\eta_R^+(q)}=\frac{1}{2}\overline{(q_+)^2},\quad \overline{(\eta_R^+)'(q)}=\overline{q_+},\quad \text{in } \Omega_R.$$

Inserting this into (6.10) and integrating the result over $(\frac{2}{R-C_2},t)\times\mathbb{R}$ gives

$$\begin{aligned}\frac{1}{2}\int_{\mathbb{R}}\left[\overline{(q_+)^2}(t,x)-q_+(t,x)^2\right]dx &\le \int_{\mathbb{R}}\left[\overline{\eta_R^+(q)}(\tfrac{2}{R-C_2},x)-\eta_R^+(q)(\tfrac{2}{R-C_2},x)\right]dx\\ &\quad+\int_{\frac{2}{R-C_2}}^t\int_{\mathbb{R}}S(s,x)\left[\overline{q_+}(s,x)-q_+(s,x)\right]dsdx,\end{aligned}$$

for almost all $t>\frac{2}{R-C_2}$. Sending $R\to\infty$ and using Lemma 6.2, we get (6.6). ∎

Lemma 6.5. *For any $t \ge 0$ and any $R > 0$,*

$$\begin{aligned}(6.11)\qquad &\int_{\mathbb{R}}\left[\overline{\eta_R^-(q)}-\eta_R^-(q)\right](t,x)dx\\ &\quad\le\frac{\gamma R^2}{2}\int_0^t\int_{\mathbb{R}}\overline{(R+q)\chi_{(-\infty,-R)}(q)}dsdx\\ &\quad\quad-\frac{\gamma R^2}{2}\int_0^t\int_{\mathbb{R}}(R+q)\chi_{(-\infty,-R)}(q)dsdx+\gamma R\int_0^t\int_{\mathbb{R}}\left[\overline{\eta_R^-(q)}-\eta_R^-(q)\right]dsdx\\ &\quad\quad+\frac{\gamma R}{2}\int_0^t\int_{\mathbb{R}}\left[\overline{(q_+)^2}-q_+^2\right]dsdx+\int_0^t\int_{\mathbb{R}}S(s,x)\left[\overline{(\eta_R^-)'(q)}-(\eta_R^-)'(q)\right]dsdx.\end{aligned}$$

Proof. Let $R > 0$. By subtracting (5.26) from (5.23), using the entropy η_R^- (see Lemma 6.2), we deduce

$$\begin{aligned}(6.12)\quad &\frac{\partial}{\partial t}\left(\overline{\eta_R^-(q)} - \eta_R^-(q)\right) + \frac{\partial}{\partial x}\left(\gamma u\left[\overline{\eta_R^-(q)} - \eta_R^-(q)\right]\right)\\ &\quad\le \gamma\left[\overline{q\eta_R^-(q)} - q\eta_R^-(q)\right] - \frac{\gamma}{2}\left[\overline{q^2(\eta_R^-)'(q)} - q^2(\eta_R^-)'(q)\right]\\ &\quad\quad - \frac{\gamma}{2}(\overline{q^2} - q^2)(\eta_R^-)'(q) + S(t,x)\left[\overline{(\eta_R^-)'(q)} - (\eta_R^-)'(q)\right].\end{aligned}$$

Since $-R \le (\eta_R^-)' \le 0$ and $\gamma \ge 0$, by (5.19),

$$-\frac{\gamma}{2}\left(\overline{q^2} - q^2\right)(\eta_R^-)'(q) \le \frac{\gamma R}{2}\left(\overline{q^2} - q^2\right). \tag{6.13}$$

Using Remarks 5.6 and 6.3

$$\gamma q\eta_R^-(q) - \frac{\gamma}{2}q^2(\eta_R^-)'(q) = -\frac{\gamma R}{2}q(R+q)\chi_{(-\infty,-R)}(q), \tag{6.14}$$

$$\gamma\overline{q\eta_R^-(q)} - \frac{\gamma}{2}\overline{q^2(\eta_R^-)'(q)} = -\frac{\gamma R}{2}\overline{q(R+q)\chi_{(-\infty,-R)}(q)}. \tag{6.15}$$

Inserting (6.13), (6.14), and (6.15) into (6.12) gives

$$\begin{aligned}&\frac{\partial}{\partial t}\left(\overline{\eta_R^-(q)} - \eta_R^-(q)\right) + \frac{\partial}{\partial x}\left(\gamma u\left[\overline{\eta_R^-(q)} - \eta_R^-(q)\right]\right)\\ &\quad\le -\frac{\gamma R}{2}\overline{q(R+q)\chi_{(-\infty,-R)}(q)} + \frac{\gamma R}{2}q(R+q)\chi_{(-\infty,-R)}(q)\\ &\quad\quad + \frac{\gamma R}{2}\left(\overline{q^2} - q^2\right) + S(t,x)\left[\overline{(\eta_R^-)'(q)} - (\eta_R^-)'(q)\right].\end{aligned}$$

Integrating this inequality over $(0,t)\times\mathbb{R}$ yields

$$\begin{aligned}(6.16)\quad &\int_{\mathbb{R}}\left[\overline{\eta_R^-(q)} - \eta_R^-(q)\right](t,x)dx\\ &\quad\le -\frac{\gamma R}{2}\int_0^t\int_{\mathbb{R}}\overline{q(R+q)\chi_{(-\infty,-R)}(q)}dsdx\\ &\quad\quad + \frac{\gamma R}{2}\int_0^t\int_{\mathbb{R}}q(R+q)\chi_{(-\infty,-R)}(q)dsdx + \frac{R}{2}\int_0^t\int_{\mathbb{R}}\left[\overline{q^2} - q^2\right]dsdx\\ &\quad\quad + \int_0^t\int_{\mathbb{R}}S(s,x)\left[\overline{(\eta_R^-)'(q)} - (\eta_R^-)'(q)\right]dsdx.\end{aligned}$$

Using Remark 6.3,

$$\overline{\eta_R^-(q)} - \eta_R^-(q) = \frac{1}{2}\left(\overline{(q_-)^2} - (q_-)^2\right) + \frac{1}{2}(R+q)^2\chi_{(-\infty,-R)}(q) - \frac{1}{2}\overline{(R+q)^2\chi_{(-\infty,-R)}(q)}.$$

Hence, from Remark 5.6 and (6.16),

$$\begin{aligned}
&\int_{\mathbb{R}} \left[\overline{\eta_R^-(q)} - \eta_R^-(q)\right](t,x)dx \\
&\quad\le -\frac{\gamma R}{2}\int_0^t\int_{\mathbb{R}} \overline{q(R+q)\chi_{(-\infty,-R)}(q)}dsdx \\
&\qquad + \frac{\gamma R}{2}\int_0^t\int_{\mathbb{R}} q(R+q)\chi_{(-\infty,-R)}(q)dsdx + \gamma R\int_0^t\int_{\mathbb{R}}\left[\overline{\eta_R^-(q)} - \eta_R^-(q)\right]dsdx \\
&\qquad - \frac{\gamma R}{2}\int_0^t\int_{\mathbb{R}} (R+q)^2\chi_{(-\infty,-R)}(q)dsdx + \frac{\gamma R}{2}\int_0^t\int_{\mathbb{R}} \overline{(R+q)^2\chi_{(-\infty,-R)}(q)}dsdx \\
&\qquad + \frac{\gamma R}{2}\int_0^t\int_{\mathbb{R}}\left[\overline{(q_+)^2} - q_+^2\right]dsdx + \int_0^t\int_{\mathbb{R}} S(s,x)\left[\overline{(\eta_R^-)'(q)} - (\eta_R^-)'(q)\right]dsdx,
\end{aligned}$$

and applying twice the identity $\frac{R}{2}(R+q)^2 - \frac{R}{2}q(R+q) = \frac{R^2}{2}(R+q)$ we deduce (6.11). ∎

Lemma 6.6. *There holds*

$$\overline{q^2} = q^2 \quad \textit{almost everywhere in } [0,\infty)\times\mathbb{R}. \tag{6.17}$$

Proof. Adding (6.6) and (6.11) yields

$$\begin{aligned}
&\int_{\mathbb{R}}\left(\frac{1}{2}\left[\overline{(q_+)^2} - (q_+)^2\right] + \left[\overline{\eta_R^-(q)} - \eta_R^-(q)\right]\right)(t,x)dx \\
&\quad\le \frac{\gamma R^2}{2}\int_0^t\int_{\mathbb{R}} \overline{(R+q)\chi_{(-\infty,-R)}(q)}dsdx - \frac{\gamma R^2}{2}\int_0^t\int_{\mathbb{R}} (R+q)\chi_{(-\infty,-R)}(q)dsdx \\
&\qquad + \gamma R\int_0^t\int_{\mathbb{R}}\left[\overline{\eta_R^-(q)} - \eta_R^-(q)\right]dsdx + \frac{\gamma R}{2}\int_0^t\int_{\mathbb{R}}\left[\overline{(q_+)^2} - q_+^2\right]dsdx \\
&\qquad + \int_0^t\int_{\mathbb{R}} S(s,x)\left(\left[\overline{q_+} - q_+\right] + \left[\overline{(\eta_R^-)'(q)} - (\eta_R^-)'(q)\right]\right)dsdx.
\end{aligned} \tag{6.18}$$

Arguing as in the proof of Lemma 3.1, there exists a constant $L>0$, depending only on $\|u_0\|_{H^1(\mathbb{R})}$, such that

$$\|S\|_{L^\infty([0,\infty)\times\mathbb{R})} = \left\|h(u) - P\right\|_{L^\infty([0,\infty)\times\mathbb{R})} \le L. \tag{6.19}$$

By Remarks 5.6 and 6.3,

$$q_+ + (\eta_R^-)'(q) = q - (R+q)\chi_{(-\infty,-R)}(q), \quad \overline{q_+} + \overline{(\eta_R^-)'(q)} = q - \overline{(R+q)\chi_{(-\infty,-R)}(q)},$$

so by the convexity of the map $\xi \mapsto \xi_+ + (\eta_R^-)'(\xi)$,

$$0 \le \left[\overline{q_+} - q_+\right] + \left[\overline{(\eta_R^-)'(q)} - (\eta_R^-)'(q)\right] = (R+q)\chi_{(-\infty,-R)}(q) - \overline{(R+q)\chi_{(-\infty,-R)}(q)},$$

and, by (6.19),

$$\begin{aligned}
&S(s,x)\left(\left[\overline{q_+}(s,x) - q_+(s,x)\right] + \left[\overline{(\eta_R^-)'(q)} - (\eta_R^-)'(q)\right]\right) \\
&\quad\le -L\left(\overline{(R+q)\chi_{(-\infty,-R)}(q)} - (R+q)\chi_{(-\infty,-R)}(q)\right).
\end{aligned}$$

Since $\xi \mapsto (R+\xi)\chi_{(-\infty,-R)}(\xi)$ is concave and choosing R large enough,

$$
\begin{aligned}
&\frac{\gamma R^2}{2}\overline{(R+q)\chi_{(-\infty,-R)}(q)} - \frac{\gamma R^2}{2}(R+q)\chi_{(-\infty,-R)}(q)\\
&\quad + S(s,x)\left([\overline{q_+}(s,x) - q_+(s,x)] + \left[\overline{(\eta_R^-)'(q)} - (\eta_R^-)'(q)\right]\right) \qquad (6.20)\\
&\le \left(\frac{\gamma R^2}{2} - L\right)\left(\overline{(R+q)\chi_{(-\infty,-R)}(q)} - (R+q)\chi_{(-\infty,-R)}(q)\right) \le 0.
\end{aligned}
$$

Then, from (6.18) and (6.20),

$$
\begin{aligned}
0 &\le \int_{\mathbb{R}} \left(\frac{1}{2}\left[\overline{(q_+)^2} - (q_+)^2\right] + \left[\overline{\eta_R^-(q)} - \eta_R^-(q)\right]\right)(t,x)dx\\
&\le \gamma R \int_0^t \int_{\mathbb{R}} \left(\frac{1}{2}\left[\overline{(q_+)^2} - q_+^2\right] + \left[\overline{\eta_R^-(q)} - \eta_R^-(q)\right]\right) dsdx,
\end{aligned}
$$

and using the Gronwall inequality and Lemmas 6.1 and 6.2 we conclude that

$$
\int_{\mathbb{R}} \left(\frac{1}{2}\left[\overline{(q_+)^2} - (q_+)^2\right] + \left[\overline{\eta_R^-(q)} - \eta_R^-(q)\right]\right)(t,x)dx = 0, \qquad \text{for each } t > 0.
$$

By the Fatou lemma, Remark 5.6, and (5.19), sending $R \to \infty$ yields

$$
0 \le \int_{\mathbb{R}} \left(\overline{q^2} - q^2\right)(t,x)dx \le 0, \qquad t > 0, \tag{6.21}
$$

and we see that (6.17) holds. ∎

LEMMA 6.7. *Assume* (1.4) *and* (2.2). *Then there exists an admissible weak solution of* (1)*, satisfying conditions* (*k*)*,* (*kk*)*, and* (*kkk*) *of Theorem* 1.2.

Proof. The conditions (*i*), (*iii*), (*iv*) of Definition 1.1 are satisfied, due to (2.2), (2.3) and Lemma 5.2. We have to verify (*ii*). Due to (6.17), we have

$$
q_\varepsilon \to q \qquad \text{in } L^2_{\mathrm{loc}}([0,\infty)\times\mathbb{R}). \tag{6.22}
$$

Clearly (5.9), (5.11), and (6.22) imply that u is a distributional solution of (1.7). Finally, (*k*) and (*kk*) are consequences of Lemmas 3.1 and 4.1, respectively. For (*kkk*) we can argue as in [32], so let us just sketch the proof. From Lemmas 5.8 and 6.6 we get

$$
\frac{\partial \eta_R(q)}{\partial t} + \frac{\partial}{\partial x}\left(\gamma u \eta_R(q)\right) = \gamma q \eta_R(q) - \frac{\gamma}{2} q^2 \eta_R'(q) + (h(u) - P)\,\eta_R'(q). \tag{6.23}
$$

From the definition of η_R,

$$
\begin{aligned}
\gamma\left(q\eta_R(q) - \frac{1}{2}q^2\eta_R'(q)\right) &= \frac{\gamma R}{2}\left(q^2 - Rq\right)\chi_{(R,\infty)}(q) - \frac{\gamma R}{2}\left(q^2 + Rq\right)\chi_{(-\infty,-R)}(q)\\
&=: S_-^R + S_+^R.
\end{aligned}
$$

By (1.9), it follows as in [32] that $\iint_{[0,\infty)\times\mathbb{R}} S_+^R dxdt \le C\|u_0\|_{H^1(\mathbb{R})}$ and thus, by integrating (6.23), $\iint_{[0,\infty)\times\mathbb{R}} S_-^R dxdt \le C$. The latter bound implies that along a subsequence $S_-^R \overset{\star}{\rightharpoonup} \mu$ in the sense of measures as $R \to \infty$, for some nonnegative Radon measure μ. By (1.9), $\iint_{[0,\infty)\times\mathbb{R}} S_+^R dxdt \to 0$ as $R \to \infty$. Hence sending

$R \to \infty$ in (6.23) and adding the result to the equation obtained by multiplying (1) by u, we get (1.13). Finally, integrating (1.13) shows that the total mass of μ is bounded by $\|u_0\|_{H^1(\mathbb{R})}$. $\square$

Remark 6.8. It is possible to prove results similar to those obtained for (1.7) for slightly more general equations of the form

$$\frac{\partial u}{\partial t} + \gamma u \frac{\partial u}{\partial x} + \frac{\partial P}{\partial x} = 0, \quad -\alpha^2 \frac{\partial^2 P}{\partial x^2} + P = h(u) + \frac{\gamma\alpha^2}{2}\left(\frac{\partial u}{\partial x}\right)^2, \tag{6.24}$$

where $\gamma \geq 0$, $\alpha > 0$, and $h\colon \mathbb{R} \to \mathbb{R}$ is any locally Lipschitz continuous function with $h(0) = 0$. The Green function of the operator $-\alpha^2 \frac{\partial^2}{\partial x^2} + 1$ is $e^{-|x|/\alpha}/2$. Formally, by letting $\alpha \to 0$, we recover the conservation law $\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} F(u) = 0$, where the flux $F(u)$ is given by $F'(u) = \gamma u - h'(u)$. Hence (6.24) may be viewed as a new type of regularization for one-dimensional conservation laws. We are currently investigating this singular limit problem.

7. Uniqueness of the viscous limit: the semigroup. Here we prove the existence of the semigroup.

LEMMA 7.1. *There exists a strongly continuous semigroup of solutions associated with the Cauchy problem* (1)

$$S\colon [0,\infty) \times (0,\infty) \times (\mathcal{E} \cap C^\infty(\mathbb{R})) \times H^1(\mathbb{R}) \longrightarrow C([0,\infty) \times \mathbb{R}) \cap L^\infty\big([0,\infty); H^1(\mathbb{R})\big),$$

namely, for each $u_0 \in H^1(\mathbb{R})$, $\gamma > 0$, $g \in \mathcal{E}$ the map $u(t,x) = S_t(\gamma, g, u_0)(x)$ is an admissible weak solution of (1). *Moreover,* (k), (kk), *and* (kkk) *of Theorem* 1.2 *are satisfied.*

Clearly, this lemma is a direct consequence of the following lemma and of the lemmas in the previous sections.

LEMMA 7.2. *Assume* (1.4), (1.5). *Let $\{\varepsilon_n\}_{n\in\mathbb{N}}$, $\{\mu_n\}_{n\in\mathbb{N}} \subset (0,\infty)$ and u, $v \in L^\infty([0,\infty); H^1(\mathbb{R})) \cap H^1([0,T] \times \mathbb{R})$, for each $T \geq 0$, be such that ε_n, $\mu_n \to 0$ and*

$$u_{\varepsilon_n} \to u, \quad u_{\mu_n} \to v, \quad \textit{strongly in } L^\infty_{\mathrm{loc}}([0,\infty); H^1(\mathbb{R})), \tag{7.1}$$

then

$$u = v.$$

Proof. Let $t > 0$, it is not restrictive to assume that

$$\|u_{0,\varepsilon} - u_{0,\mu}\|_{H^1(\mathbb{R})} \leq |\varepsilon - \mu|, \qquad \varepsilon,\ \mu > 0. \tag{7.2}$$

Moreover, passing to subsequences, we can assume that

$$0 < \mu_n < \varepsilon_n < \mu_{n-1}, \qquad n \in \mathbb{N}. \tag{7.3}$$

Indeed, we can argue in the following way: we begin by considering two strictly decreasing subsequences $\{\varepsilon_{n_k}\}_{k\in\mathbb{N}}$, $\{\mu_{n_k}\}_{k\in\mathbb{N}}$. Then we start by defining $\mu_{n_{k_0}} = \mu_{n_0}$, then we continue with $\varepsilon_{n_{k_1}}$ and $\mu_{n_{k_1}}$ in the following way:

$$\varepsilon_{n_{k_1}} = \max\big\{\varepsilon_{n_k};\ \varepsilon_{n_k} < \mu_{n_{k_0}}\big\}, \qquad \mu_{n_{k_1}} = \max\big\{\mu_{n_k};\ \mu_{n_k} < \varepsilon_{n_{k_1}}\big\}.$$

Arguing inductively in this way we find two subsequences $\{\varepsilon_{n_{k_h}}\}_{h\in\mathbb{N}}$, $\{\mu_{n_{k_h}}\}_{h\in\mathbb{N}}$ satisfying (7.3).

From [4, Theorem 3.1] and (7.2), we have that

$$\begin{aligned}
&\|u_\varepsilon(t,\cdot\,) - u_\mu(t,\cdot\,)\|_{H^1(\mathbb{R})}\\
&\quad \le A(t,\varepsilon+\mu)\|u_{0,\varepsilon} - u_{0,\mu}\|_{H^1(\mathbb{R})} + B(t,\varepsilon+\mu)|\varepsilon-\mu|\\
&\quad \le \big(A(t,\varepsilon+\mu) + B(t,\varepsilon+\mu)\big)|\varepsilon-\mu|,
\end{aligned}$$

with

$$A(t,\varepsilon+\mu) = \mathcal{O}\big(e^{t/(\varepsilon+\mu)}\big), \qquad B(t,\varepsilon+\mu) = \mathcal{O}\big(e^{t/(\varepsilon+\mu)}\big),$$

for each ε, $\mu > 0$. Hence

$$\|u_\varepsilon(t,\cdot\,) - u_\mu(t,\cdot\,)\|_{H^1(\mathbb{R})} \le c_1 e^{t/(\varepsilon+\mu)}|\varepsilon-\mu|, \qquad \varepsilon,\mu>0, \tag{7.4}$$

for some constant $c_1 > 0$. Define

$$\varepsilon_{k,n} := \varepsilon_n - k e^{-1/\varepsilon_n^2}, \quad N_n := \big[(\varepsilon_n - \mu_n)e^{1/\varepsilon_n^2}\big], \qquad k,n\in\mathbb{N},$$

where $[\cdot]$ denotes the integer part. Observe that

$$\varepsilon_{N_n,n} \le \mu_n \le \varepsilon_{N_n-1,n}, \quad \mu_n - \varepsilon_{N_n,n} \le e^{-1/\varepsilon_n^2}, \qquad n\in\mathbb{N}, \tag{7.5}$$

and

$$\begin{aligned}
\lim_n \|u_{\varepsilon_{k,n}}(t,\cdot\,) - u_{\varepsilon_n}(t,\cdot\,)\|_{H^1(\mathbb{R})}
&\le c_1 \lim_n e^{t/(\varepsilon_{k,n}+\varepsilon_n)}|\varepsilon_{k,n}-\varepsilon_n|\\
&\le c_1 k \lim_n e^{t/\varepsilon_n} e^{-1/\varepsilon_n^2} = 0,
\end{aligned} \tag{7.6}$$

for each $k\in\mathbb{N}$, in other terms

$$u_{\varepsilon_{k,n}} \to u, \quad \text{strongly in } L^\infty_{\mathrm{loc}}([0,\infty);H^1(\mathbb{R})) \text{ as } n\to\infty, \quad \text{for each } k\in\mathbb{N}. \tag{7.7}$$

Since, from (7.3),

$$\varepsilon_{n+1} < \mu_n < \mu_n + \varepsilon_{N_n,n}, \qquad n\in\mathbb{N},$$

employing (7.5), we have that

$$\begin{aligned}
\lim_n \|u_{\varepsilon_{N_n,n}}(t,\cdot\,) - u_{\mu_n}(t,\cdot\,)\|_{H^1(\mathbb{R})}
&\le c_1 \lim_n e^{t/(\varepsilon_{N_n,n}+\mu_n)}|\varepsilon_{N_n,n}-\mu_n|\\
&\le c_1 \lim_n e^{t/\varepsilon_{n+1}} e^{-1/\varepsilon_n^2} = 0.
\end{aligned} \tag{7.8}$$

Hence

$$u_{\varepsilon_{N_n,n}} \to v, \quad \text{strongly in } L^\infty_{\mathrm{loc}}([0,\infty);H^1(\mathbb{R})) \text{ as } n\to\infty. \tag{7.9}$$

If $\{N_n\}_{n\in\mathbb{N}}$ is bounded, the claim is a direct consequence of (7.7) and (7.9). So we consider the case

$$\lim_n N_n = \infty. \tag{7.10}$$

As before, we define the sequences

$$\mu_{h,n} := \mu_n + he^{-1/\mu_n^2}, \quad M_n := \left[(\varepsilon_n - \mu_n)e^{1/\mu_n^2}\right], \qquad h, n \in \mathbb{N}.$$

Due to (7.3) and (7.10),

$$\lim_n M_n = \infty, \tag{7.11}$$

and, arguing as for (7.7) and (7.9), we are able to prove that

$$\lim_n \|u_{\mu_{h,n}}(t,\cdot) - u_{\mu_n}(t,\cdot)\|_{H^1(\mathbb{R})} = \lim_n \|u_{\mu_{h,n}}(t,\cdot) - v(t,\cdot)\|_{H^1(\mathbb{R})} = 0, \quad h \in \mathbb{N}, \tag{7.12}$$

$$\lim_n \|u_{\mu_{M_n,n}}(t,\cdot) - u_{\varepsilon_n}(t,\cdot)\|_{H^1(\mathbb{R})} = \lim_n \|u_{\mu_{M_n,n}}(t,\cdot) - u(t,\cdot)\|_{H^1(\mathbb{R})} = 0. \tag{7.13}$$

Due to (7.10) and (7.11), we can choose two sequences $\{k_n\}_{n\in\mathbb{N}}$, $\{h_n\}_{n\in\mathbb{N}}$, such that

$$\mu_n \le \mu_{h_n,n}, \varepsilon_{k_n,n} \le \varepsilon_n, \quad |\mu_{h_n,n} - \varepsilon_{k_n,n}| \le c_2 e^{-1/\mu_n^2}, \qquad n \in \mathbb{N}, \tag{7.14}$$

$$\lim_n h_n = \lim_n k_n = \infty, \tag{7.15}$$

for some constant $c_2 > 0$. Observe that

$$\begin{aligned}\|u(t,\cdot) - v(t,\cdot)\|_{H^1(\mathbb{R})} \le\ & \|u(t,\cdot) - u_{\mu_{h,n}}(t,\cdot)\|_{H^1(\mathbb{R})} \\ & + \|u_{\mu_{h,n}}(t,\cdot) - u_{\varepsilon_{k,n}}(t,\cdot)\|_{H^1(\mathbb{R})} \\ & + \|u_{\varepsilon_{k,n}}(t,\cdot) - v(t,\cdot)\|_{H^1(\mathbb{R})}.\end{aligned} \tag{7.16}$$

From (7.9) and (7.13), we have

$$\liminf_{h,n} \|u(t,\cdot) - u_{\mu_{h,n}}(t,\cdot)\|_{H^1(\mathbb{R})} \le \lim_n \|u(t,\cdot) - u_{\mu_{M_n,n}}(t,\cdot)\|_{H^1(\mathbb{R})} = 0, \tag{7.17}$$

$$\liminf_{k,n} \|u_{\varepsilon_{k,n}}(t,\cdot) - v(t,\cdot)\|_{H^1(\mathbb{R})} \le \lim_n \|u_{\varepsilon_{N_n,n}}(t,\cdot) - v(t,\cdot)\|_{H^1(\mathbb{R})} = 0, \tag{7.18}$$

respectively. Finally, from (7.4), (7.14), and (7.15),

$$\begin{aligned}&\liminf_{h,k,n} \|u_{\mu_{h,n}}(t,\cdot) - u_{\varepsilon_{k,n}}(t,\cdot)\|_{H^1(\mathbb{R})} \\ &\quad \le \liminf_n \|u_{\mu_{h_n,n}}(t,\cdot) - u_{\varepsilon_{k_n,n}}(t,\cdot)\|_{H^1(\mathbb{R})} \\ &\quad \le c_1 c_2 \liminf_n e^{-1/\mu_n^2} e^{t(\mu_{h_n,n} + \varepsilon_{k_n,n})} \end{aligned} \tag{7.19}$$

$$\le c_1 c_2 \lim_n e^{-1/\mu_n^2} e^{t(\mu_n)} = 0. \tag{7.20}$$

Clearly, (7.16), (7.17), (7.18), (7.19), imply $u = v$. ∎

8. Stability of the semigroup and proof of Theorem 1.2. Here we prove the stability of the semigroup and then conclude the proof of Theorem 1.2.

LEMMA 8.1. *The semigroup S defined on $[0,\infty) \times (0,\infty) \times (\mathcal{E} \cap C^\infty(\mathbb{R})) \times H^1(\mathbb{R})$ satisfies the stability property* (jj) *of Theorem* 1.2.

Proof. Fix $\varepsilon > 0$. Denote S^ε the semigroup associated to the viscous problem (2.1). Choose $\{u_{0,n}\}_{n\in\mathbb{N}} \subset H^1(\mathbb{R})$, $\{\gamma_n\}_{n\in\mathbb{N}} \subset (0,\infty)$, $\{g_n\}_{n\in\mathbb{N}} \subset \mathcal{E} \cap C^\infty(\mathbb{R})$,

$u_0 \in H^1(\mathbb{R})$, $\gamma > 0$, $g \in \mathcal{E} \cap C^\infty(\mathbb{R})$ satisfying (1.10). The initial data satisfy $u_{0,\varepsilon,n}$, $u_{0,\varepsilon} \in H^\ell(\mathbb{R})$, $\ell \geq 2$, the condition (2.2), and

$$\|u_{0,\varepsilon,n} - u_{0,\varepsilon}\|_{H^1(\mathbb{R})} \leq \|u_{0,n} - u_0\|_{H^1(\mathbb{R})}. \tag{8.1}$$

Finally, write

$$u_{\varepsilon,n} := S^\varepsilon(\gamma_n, g_n, u_{0,n}), \qquad u_n := S(\gamma_n, g_n, u_{0,n}), \qquad u := S(\gamma, g, u_0).$$

Let $t > 0$, then

$$\begin{aligned} \|u_n(t,\cdot) - u(t,\cdot)\|_{H^1(\mathbb{R})} &\leq \|u_n(t,\cdot) - u_{\varepsilon,n}(t,\cdot)\|_{H^1(\mathbb{R})} \\ &\quad + \|u_{\varepsilon,n}(t,\cdot) - u_\varepsilon(t,\cdot)\|_{H^1(\mathbb{R})} + \|u_\varepsilon(t,\cdot) - u(t,\cdot)\|_{H^1(\mathbb{R})}, \end{aligned} \tag{8.2}$$

so

$$\begin{aligned} 0 \leq \liminf_n \|u_n(t,\cdot) - u(t,\cdot)\|_{H^1(\mathbb{R})} &\leq \liminf_{\varepsilon,n} \|u_n(t,\cdot) - u_{\varepsilon,n}(t,\cdot)\|_{H^1(\mathbb{R})} \\ &\quad + \liminf_{\varepsilon,n} \|u_{\varepsilon,n}(t,\cdot) - u_\varepsilon(t,\cdot)\|_{H^1(\mathbb{R})} \\ &\quad + \liminf_{\varepsilon} \|u_\varepsilon(t,\cdot) - u(t,\cdot)\|_{H^1(\mathbb{R})}. \end{aligned} \tag{8.3}$$

From Lemma 5.2 we know that

$$\liminf_{\varepsilon,n} \|u_n(t,\cdot) - u_{\varepsilon,n}(t,\cdot)\|_{H^1(\mathbb{R})} = 0, \tag{8.4}$$

$$\liminf_{\varepsilon} \|u_\varepsilon(t,\cdot) - u(t,\cdot)\|_{H^1(\mathbb{R})} = 0. \tag{8.5}$$

We claim that

$$\liminf_{\varepsilon,n} \|u_{\varepsilon,n}(t,\cdot) - u_\varepsilon(t,\cdot)\|_{H^1(\mathbb{R})} = 0. \tag{8.6}$$

Using [4, Theorem 3.1] and (8.1), we have that

$$\begin{aligned} \|u_{\varepsilon,n}(t,\cdot) - u_\varepsilon(t,\cdot)\|_{H^1(\mathbb{R})} &\leq A(t,\varepsilon)\|u_{0,n} - u_0\|_{H^1(\mathbb{R})} \\ &\quad + B(t,\varepsilon)\big(\|g_n - g\|_{L^\infty(\mathcal{I})} + |\gamma_n - \gamma|\big), \end{aligned} \tag{8.7}$$

with

$$A(t,\varepsilon) = \mathcal{O}(e^{T/\varepsilon}), \qquad B(t,\varepsilon) = \mathcal{O}(e^{T/\varepsilon}), \quad t \in [0,T].$$

Define

$$\varepsilon_n := \frac{T}{|\log(k_n)|}, \qquad k_n := \max\left\{\|u_{0.n} - u_0\|_{H^1(\mathbb{R})}^{1/2}, \|g_n - g\|_{L^\infty(\mathcal{I})}^{1/2}, |\gamma_n - \gamma|^{1/2}\right\},$$

clearly

$$\liminf_{\varepsilon,n} \|u_{\varepsilon,n}(t,\cdot) - u_\varepsilon(t,\cdot)\|_{H^1(\mathbb{R})} \leq \liminf_n \|u_{\varepsilon_n,n}(t,\cdot) - u_{\varepsilon_n}(t,\cdot)\|_{H^1(\mathbb{R})}, \tag{8.8}$$

and

$$\lim_n A(\varepsilon_n, t)\|u_{0,n} - u_0\|_{H^1(\mathbb{R})} = \lim_n B(\varepsilon_n, t)\|g_n - g\|_{L^\infty(\mathcal{I})} = 0. \tag{8.9}$$

Then (8.6) is a consequence of (8.7), (8.8), and (8.9). From (8.3)–(8.6), we get

$$\lim_n \|u_n(t,\cdot) - u(t,\cdot)\|_{H^1(\mathbb{R})} = 0. \qquad \square$$

Proof of Theorem 1.2. The proof is a direct consequence of Lemmas 7.1 and 8.1. $\square$

REFERENCES

[1] R. Beals, D. H. Sattinger, and J. Szmigielski, *Acoustic scattering and the extended Korteweg–de Vries hierarchy*, Adv. Math., 140 (1998), pp. 190–206.

[2] T. B. Benjamin, J. L. Bona, and J. J. Mahony, *Model equations for long waves in nonlinear dispersive systems*, Philos. Trans. Roy. Soc. London Ser. A, 272 (1972), pp. 47–78.

[3] R. Camassa and D. D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.

[4] G. M. Coclite, H. Holden, and K. H. Karlsen, *Wellposedness for a parabolic-elliptic system*, Discrete Cont. in Dyn. Syst., 13 (2005), pp. 659–682.

[5] A. Constantin, *Existence of permanent and breaking waves for a shallow water equation: a geometric approach*, Ann. Inst. Fourier (Grenoble), 50 (2000), pp. 321–362.

[6] A. Constantin, *On the scattering problem for the Camassa–Holm equation*, R. Soc. Lond. Proc. Ser. A. Math. Phys. Eng. Sci., 457 (2001), pp. 953–970.

[7] A. Constantin and J. Escher, *Global existence and blow-up for a shallow water equation*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 26 (1998), pp. 303–328.

[8] A. Constantin and J. Escher, *Global weak solutions for a shallow water equation*, Indiana Univ. Math. J., 47 (1998), pp. 1527–1545.

[9] A. Constantin and J. Escher, *On the Cauchy problem for a family of quasilinear hyperbolic equations*, Comm. Partial Differential Equations, 23 (1998), pp. 1449–1458.

[10] A. Constantin and J. Escher, *Wave breaking for nonlinear nonlocal shallow water equations*, Acta Math., 181 (1998), pp. 229–243.

[11] A. Constantin and H. P. McKean, *A shallow water equation on the circle*, Comm. Pure Appl. Math., 52 (1999), pp. 949–982.

[12] A. Constantin and L. Molinet, *Global weak solutions for a shallow water equation*, Comm. Math. Phys., 211 (2000), pp. 45–61.

[13] A. Constantin and W. A. Strauss, *Stability of a class of solitary waves in compressible elastic rods*, Phys. Lett. A, 270 (2000), pp. 140–148.

[14] H.-H. Dai, *Exact travelling-wave solutions of an integrable equation arising in hyperelastic rods*, Wave Motion, 28 (1998), pp. 367–381.

[15] H.-H. Dai, *Model equations for nonlinear dispersive waves in a compressible Mooney–Rivlin rod*, Acta Mech., 127 (1998), pp. 193–207.

[16] H.-H. Dai and Y. Huo, *Solitary shock waves and other travelling waves in a general compressible hyperelastic rod*, R. Soc. Lond. Proc. Ser. A, Math. Phys. Eng. Sci., 456 (2000), pp. 331–363.

[17] R. Danchin, *A few remarks on the Camassa–Holm equation*, Differential Integral Equations, 14 (2001), pp. 953–988.

[18] R. Danchin, *A note on well-posedness for Camassa–Holm equation*, J. Differential Equations, 192 (2003), pp. 429–444.

[19] R. J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[20] B. Fuchssteiner and A. S. Fokas, *Symplectic structures, their Bäcklund transformations and hereditary symmetries*, Phys. D, 4 (1981/82), pp. 47–66.

[21] A. A. Himonas and G. Misiołek, *The Cauchy problem for an integrable shallow-water equation*, Differential Integral Equations, 14 (2001), pp. 821–831.

[22] R. S. Johnson, *Camassa–Holm, Korteweg–de Vries and related models for water waves*, J. Fluid Mech., 455 (2002), pp. 63–82.

[23] H. Kalisch, *Stability of solitary waves for a nonlinearly dispersive equation*, Discrete Contin. Dyn. Syst., 10 (2004), pp. 709–717.

[24] Y. A. Li and P. J. Olver, *Well-posedness and blow-up solutions for an integrable nonlinearly dispersive model wave equation*, J. Differential Equations, 162 (2000), pp. 27–63.

[25] E. H. Lieb and M. Loss, *Analysis*, Graduate Studies in Mathematics 14, 2nd ed., American Mathematical Society, Providence, RI, 2001.

[26] P.-L. Lions, *Mathematical Topics in Fluid Mechanics, Vol.* 1. *Incompressible Models*, in *Oxford Lecture Series in Mathematics and its Applications*, 3, Oxford University Press, New York, 1996.

[27] P.-L. Lions, *Mathematical Topics in Fluid Mechanics, Vol.* 2. *Compressible Models*, in *Oxford Lecture Series in Mathematics and its Applications*, 10, Oxford University Press, New York, 1998.

[28] O. Lopes, *Stability of peakons for the generalized Camassa–Holm equation*, Electron. J. Differential Equations, (electronic), 2003.

[29] T. Qian and M. Tang, *Peakons and periodic cusp waves in a generalized Camassa–Holm equation*, Chaos Solitons Fractals, 12 (2001), pp. 1347–1360.

[30] G. Rodríguez-Blanco, *On the Cauchy problem for the Camassa–Holm equation*, Nonlinear Anal., 46 (2001), pp. 309–327.

[31] J. Simon, *Compact sets in the space* $L^p(0,T;B)$, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.

[32] Z. Xin and P. Zhang, *On the weak solutions to a shallow water equation*, Comm. Pure Appl. Math., 53 (2000), pp. 1411–1433.

[33] Z. Xin and P. Zhang, *On the uniqueness and large time behavior of the weak solutions to a shallow water equation*, Comm. Partial Differential Equations, 27 (2002), pp. 1815–1844.

[34] Z. Yin, *On the blow-up scenario for the generalized Camassa–Holm equation*, Comm. Partial Differential Equations, 29 (2004), pp. 867–877.

[35] Z. Yin, *On the Cauchy problem for the generalized Camassa–Holm equation*, Nonlinear Anal., to appear.

[36] Z. Yin, *On the Cauchy problem for a nonlinearly dispersive wave equation*, J. Nonlinear Math. Phys., 10 (2003), pp. 10–15.

SIAM J. MATH. ANAL.
Vol. 37, No. 4, pp. 1070–1097

AN ACCELERATED SPLITTING-UP METHOD FOR PARABOLIC EQUATIONS*

ISTVÁN GYÖNGY† AND NICOLAI KRYLOV‡

Abstract. We approximate the solution u of the Cauchy problem

$$\frac{\partial}{\partial t}u(t,x) = Lu(t,x) + f(t,x), \quad (t,x) \in (0,T] \times \mathbb{R}^d,$$

$$u(0,x) = u_0(x), \quad x \in \mathbb{R}^d,$$

by splitting the equation into the system

$$\frac{\partial}{\partial t}v_r(t,x) = L_r v_r(t,x) + f_r(t,x), \qquad r = 1,2,\dots,d_1,$$

where L, L_r are second order differential operators; f, f_r are functions of t,x such that $L = \sum_r L_r$, $f = \sum_r f_r$. Under natural conditions on solvability in the Sobolev spaces W^m_p, we show that for any $k > 1$ one can approximate the solution u with an error of order δ^k, by an appropriate combination of the solutions v_r along a sequence of time discretization, where δ is proportional to the step size of the grid. This result is obtained by using the time change introduced in [I. Gyöngy and N. Krylov, *Ann. Probab.*, 31 (2003), pp. 564–591], together with Richardson's method and a power series expansion of the error of splitting-up approximations in terms of δ.

Key words. Cauchy problem, parabolic partial differential equations, splitting-up, method of alternative direction, Richardson's method

AMS subject classifications. 35K65, 65B99

DOI. 10.1137/S0036141003437903

1. Introduction.

In this paper we are interested in the rate of convergence of splitting-up approximations to the solution of the parabolic, possibly *degenerate,* differential equation with time dependent coefficients

$$\frac{\partial}{\partial t}u(t,x) = Lu(t,x) + f(t,x), \quad (t,x) \in (0,T] \times \mathbb{R}^d, \tag{1.1}$$

with initial condition

$$u(0,x) = u_0(x), \quad x \in \mathbb{R}^d, \tag{1.2}$$

where L is a differential operator of the form

$$L = a^{ij}(t,x)\frac{\partial^2}{\partial x^i x^j} + a^i(t,x)\frac{\partial}{\partial x^i} + a(t,x)$$

and f is a function of $t \geq 0$ and $x \in \mathbb{R}^d$. The first step in the splitting methods is to choose suitable decompositions $L = L_1 + L_2 + \cdots + L_{d_1}$ and $f = f_1 + f_2 + \cdots + f_{d_1}$ for the operator L and the free term f such that each equation

$$\frac{\partial}{\partial t}v_r(t,x) = L_r v_r(t,x) + f_r(t,x), \tag{1.3}$$

*Received by the editors November 24, 2003; accepted for publication (in revised form) April 25, 2005; published electronically December 30, 2005.
http://www.siam.org/journals/sima/37-4/43790.html

†School of Mathematics, University of Edinburgh, King's Buildings, Edinburgh, EH9 3JZ, UK (gyongy@maths.ed.ac.uk). This author's research was partially supported by EU Network HARP.

‡School of Mathematics, 127 Vincent Hall, University of Minnesota, Minneapolis, MN, 55455 (krylov@math.umn.edu). This author's research was partially supported by NSF grant DMS-0140405.

$r = 1, 2, \dots, d_1$, is integrable exactly or approximately.

Assume for simplicity that the operators L_r and the free terms f_r, $r = 1, 2, \dots, d_1$, are time independent, and, for fixed $T > 0$ and integer $n \geq 1$, consider the uniform step

$$T_n := \left\{ t_i^n := \frac{iT}{n}, i = 0, 1, 2, \dots, n \right\} \tag{1.4}$$

of step size $\delta := T/n$. Then a splitting-up approximation $u^{(n)}$ for the solution u of (1.1)–(1.2) is defined by

$$u_n(t_i^n) := (\mathbb{S}_\delta^{(d_1)} \dots \mathbb{S}_\delta^{(2)} \mathbb{S}_\delta^{(1)})^i u_0, \qquad i = 0, 1, \dots, n, \tag{1.5}$$

at the grid points. Here $\mathbb{S}_t^{(r)}$ denotes the solution operator of (1.3); i.e., $\mathbb{S}_t^{(r)} \varphi$ is the solution of (1.3) at time t with initial condition φ at $t = 0$. Formula (1.5) means that we take $u^{(n)}(0) = u_0$, and we calculate the approximation at a grid point $t+\delta$ from the approximation $u^{(n)}(t)$ at the previous grid point t, by solving (1.3) for $r = 1, 2, \dots, d_1$ on the same time interval $[0, \delta]$ successively. First we solve the first equation ($r = 1$) on $[0, \delta]$ with initial condition $v_1(0) = u^{(n)}(t)$, and then we solve the second equation, third equation, and so on, on the same interval $[0, \delta]$, by always taking the value at δ of the solution of the previous equation as the initial value for the following equation. Finally we solve the last equation ($r = d_1$) on the interval $[0, \delta]$ with initial condition $v_{d_1}(0) = v_{d_1 - 1}(\delta)$, and the value $v_{d_1}(\delta)$ is the value of the splitting-up approximation at $t + \delta$.

This kind of approximations is well known in numerical analysis, and it has been successfully applied to various types of PDE problems. They are often combined with other numerical methods, such as finite differences, finite elements, etc. Pioneering applications to the heat equation, to hyperbolic equations, and to nonlinear PDEs are presented, for example, in [16], [4], [2], in [29], [9] and in [3], [24], [1], [23], respectively. Many applications and modifications of the splitting-up method have been developed in various applied fields of linear and nonlinear PDEs and ODEs, under a variety of different names, like dimensional splitting, operator splitting, predictor-corrector method, method of alternating directions, fractional step method, Lie–Trotter–Kato formula, Baker–Campbell–Hausdorff formula, Chernoff formula, split Hamiltonian, split-steps, or leapfrog. For guidance in the huge varieties of methods, names, and references we refer the reader to the survey article [13] and books [11], [12].

In the context of semigroups the splitting-up method first appears as Trotter's formula [25], which can be formulated as follows:

$$\lim_{n \to \infty} (e^{tA_{d_1}/n} \dots e^{tA_2/n} e^{tA_1/n})^n z = e^{tA} z \quad \forall z \in \mathbb{B},$$

where $A = A_1 + A_2 + \dots + A_{d_1}$ and A_r are infinitesimal generators of C_0-semigroups of contractions $\{e^{tA} : t \geq 0\}$ and $\{e^{tA_r} : t \geq 0\}$ on a Banach space $\mathbb{B}$, such that the intersection of the domains of the generators is dense in $\mathbb{B}$. Clearly, in the context of Cauchy problems, Trotter's formula states the convergence of the splitting-up approximations defined by the splitting

$$\frac{d}{dt} v_r(t) = A_r v_r(t), \qquad r = 1, 2, \dots, d_1,$$

to the solution of the abstract Cauchy problem

$$\frac{d}{dt} u(t) = Au(t), \quad t \geq 0, \quad u(0) = z.$$

Our main interest in the present paper is to increase the accuracy of the splitting-up approximations for (1.1). It is known that the error of the splitting-up approximations is proportional to δ, the step size. There are, however, modifications of these approximations which are more accurate. A celebrated example is the Strang symmetric scheme

$$u_n(t_i^{(n)}) := (\mathbb{S}_{\delta/2}^{(1)}\mathbb{S}_{\delta/2}^{(2)}\dots\mathbb{S}_{\delta/2}^{(d_1)}\mathbb{S}_{\delta/2}^{(d_1)}\dots\mathbb{S}_{\delta/2}^{(2)}\mathbb{S}_{\delta/2}^{(1)})^i u_0, \qquad i=0,1,\dots,n,$$

whose error is proportional to δ^2. This approximation scheme is presented in [17], [19]. Other symmetric schemes and their generalizations are given in [17], [18], and [6]. All these schemes are of second order accuracy. Inspired by the above example, for given $k \geq 2$ one looks for a composition of splittings

$$\prod_{i=1}^{m}\prod_{j=1}^{d_1}\mathbb{S}_{c^{ij}\delta}^{(j)}, \tag{1.6}$$

with real numbers c^{ij} and integer $m \geq 1$ to be determined, such that

$$u(\delta) - \prod_{i=1}^{m}\prod_{j=1}^{d_1}\mathbb{S}_{c^{ij}\delta}^{(j)}u_0,$$

the local error of the corresponding approximation is proportional to δ^{k+1} in appropriate norms. Such local error leads to a global error, proportional to δ^k; i.e., composition (1.6) represents a method of (at least) *order* k. The conditions on the numbers c^{ij} and m, which lead to splitting methods of high order, have been studied intensively in the literature. Such methods are obtained in [15] for Hamiltonian systems by the Baker–Campbell–Hausdorff formula. Variations of the Trotter formula and the Baker–Campbell–Hausdorff formula are used for linear and for nonlinear equations, respectively, to show the existence of methods of any order (see [13], [21], [22], [26], [28], and the literature therein). An adaptation of the method of *rooted trees* from the theory of Runge–Kutta approximations is used in [14]. By [20] and [27], however, the numbers c^{ij} in each scheme (1.6) of order $k \geq 3$ cannot all be nonnegative. Thus, by [20] and [27], the above splitting methods of order greater than or equal to 3 cannot be used to approximate the solution of partial differential equations of parabolic type. As McLachlan and Quispel write on page 392 of [13]: "... splitting was proposed as a cheap way to retain unconditional stability. Methods with backward time steps can be only conditionally stable; this stumbling block held up the development of high-order compositions for years."

Then the natural question arises as to whether there exists, in the case of parabolic equations, a way different from the multiplicative one to accelerate the convergence to a higher order. One of our main results consists of showing that, using the step size of order δ but organizing the computations differently, it is indeed possible to achieve the accuracy of order δ^k for any k, even if A_r are (degenerate) elliptic operators with coefficients depending on time. In a subsequent article we intend to show that our method is much more universal in the sense that it covers many situations in which method (1.6) works and requires approximately the same amount of work.

In the present paper we use linear combinations of splittings of type (1.5) with different step sizes to achieve arbitrary high accuracy. We prove that for any given $k \geq 0$ there exist absolute constants $b_0, b_1, \dots, b_k$ expressed by simple formulas such that the accuracy of the approximation

$$v_n := b_0 u_n + b_1 u_{2n} + b_2 u_{4n} + \cdots + b_k u_{2^k n} \tag{1.7}$$

is of order δ^{k+1} (see Theorem 2.3 below). Here $u_{2^j n}$ is the splitting-up approximation (1.5) along the grid (1.4), with $2^j n$ in place of n. In particular, if $k = 1$, we have to deal with two step sizes: δ and $\delta/2$, and we get the order of accuracy δ^2. The Strang formula giving the same order of accuracy, generally, also requires working with step size $\delta/2$. By the way, if $A = A_1 + A_2$ and we construct our splitting-up scheme according to $A = (1/2)A_1 + A_2 + (1/2)A_1$, then our approximations just coincide with the Strang one, and there is no need to use linear combinations to get the error of order δ^2. It is also worth noting that the above coefficients $\mathbf{b} = (b_0, \ldots, b_k)$ are given by $\mathbf{b} := \mathbf{e}_1 V^{-1}$, where $\mathbf{e}_1 := (1, 0, 0, \ldots, 0)$ and V^{-1} is the inverse of the $(k+1) \times (k+1)$ Vandermonde matrix $V^{ij} = 2^{-(i-1)(j-1)}$.

Our work is of a purely theoretical nature and, as the referees pointed out, much work yet needs to be done before our results could be used in practical applications. We restricted ourselves to making the first step in attacking Problem 10 on page 492 of [13]: "For systems that evolve in a semigroup, such as the heat equation, develop effective methods of order higher than 2." However, our results show that each time one has *any* algorithm implementing standard splitting-up method to approximate the solutions of the Cauchy problem for degenerate parabolic equations with sufficiently smooth coefficients and free terms, one can improve the rate of convergence to any degree. For instance, we believe that usually in practice one is not doing computations with only one step size, and we show that having, say, three different step sizes, each of which is of order δ of accuracy, and just taking a linear combination of the results, one gets an approximation with error of order δ^3.

We have to admit that we do not know whether our methods can be carried over to quasi-linear equations or to equations in domains. In this connection we note that there is a very active area of developing and applying in practice splitting-up methods for degenerate nonlinear convection-diffusion equations (see, for instance, [5] and references therein). Our equations can be viewed as belonging to this area only if a^{ij} are constant. However, it is perhaps worth mentioning that our methods can be applied to solving systems of (nonlinear) ODEs, and we are in the process of working on this subject.

Inspired by Richardson's method we obtain our results by expanding the error $u - u_n$ of the splitting-up approximation (1.5) in powers of $\delta = T/n$. This is Theorem 2.2, the main theorem of our paper. We use this expansion with $\delta = T/2^j n$, $j = 0, 1, 2, \ldots, k$, and choose the above coefficients $b_0, b_1, \ldots, b_k$ to eliminate the terms of order less than $k + 1$ in the linear combination (1.7).

The main theorem of the present paper is proved by exploiting a new approach of [7] and [8] to splitting-up methods. As we discussed above, the splitting-up approximation (1.5) means that to get the approximation at $t + \delta$ from that at t, one goes back and forth in time d_1 times while solving (1.3), $r = 1, 2, \ldots, d_1$, successively. A basic idea of [7] is to arrange the splitting continuously in forward time direction, and to synchronize it with the original equation by time-scaling. In this way we have differential equations for the rearranged splitting-up approximations and for the time-scaled solution of the original equation, which enables us to use methods of the theory of partial differential equations and not semigroup theory and get an expansion for their difference in terms of powers of δ even if the coefficients depend on time. The method of [7] and [8] appeared in connection with splitting-up for stochastic PDEs.

It is worth mentioning that most likely it is impossible to accelerate the splitting-up method in this more complicated situation.

The paper is organized as follows. In the next section we introduce our general setting but state the results, Theorems 2.3 and 2.2, only for the case of time independent data for the sake of simplicity of presentation. Theorem 2.2 is proved immediately after its formulation on the basis of Theorem 2.3, which in turn is proved in section 4, after we prepare some auxiliary facts in section 3. In section 5 we generalize Theorems 2.3 and 2.2 for time dependent data and derive some consequences valid in the time-homogeneous case as well.

In conclusion we introduce some notation used everywhere below. Throughout the paper $d \geq 1, d_1 \geq 2$ are fixed positive integers, K, T are fixed finite positive constants, and

$$D_i := \frac{\partial}{\partial x^i}, \quad D_{ij} := \frac{\partial^2}{\partial x^i \partial x^j}, \quad D_t := \frac{\partial}{\partial t}.$$

We denote by W_p^m the Sobolev space defined as the closure of C_0^∞ functions $\varphi : \mathbb{R}^d \to \mathbb{R}$ in the norm

$$\|\varphi\|_{m,p} := \left(\sum_{|\gamma| \leq m} \int_{\mathbb{R}^d} |D^\gamma \varphi(x)|^p \, dx \right)^{1/p},$$

where $D^\gamma := D_1^{\gamma_1} \dots D_d^{\gamma_d}$ for multi-indices $\gamma = (\gamma_1, \dots, \gamma_d)$ of length $|\gamma| := \gamma_1 + \gamma_2 + \dots + \gamma_d$. Unless otherwise indicated, we use the summation convention with respect to repeated indices.

2. Formulation of the main results. The case of time independent coefficients. We consider the problem

$$D_t u(t,x) = Lu(t,x) + f(t,x), \qquad t \in (0,T], \ x \in \mathbb{R}^d, \tag{2.1}$$

$$u(0,x) = u_0(x), \quad x \in \mathbb{R}^d, \tag{2.2}$$

where L is an operator of the form

$$L = a^{ij}(t,x) D_{ij} + a^i(t,x) D_i + a(t,x),$$

and f and u_0 are real functions of $(t,x) \in (0,T] \times \mathbb{R}^d$ and of $x \in \mathbb{R}^d$, respectively. We assume that the coefficients a^{ij}, a^i, a and the derivatives $a^{ij}_{x^k}$ of a^{ij} are bounded Borel functions of (t,x). We fix $p \geq 2$ and assume that u_0 and f are measurable and $|u_0|^p$ and $|f|^p$ are integrable over $\mathbb{R}^d$ and over $[0,T] \times \mathbb{R}^d$, respectively.

DEFINITION 2.1. *By a solution of problem* (2.1)–(2.2) *we mean an W_p^1-valued weakly continuous function $u(t) = u(t,\cdot)$ defined on $[0,T]$ such that for all $\phi \in C_0(\mathbb{R}^d)$ and $t \in [0,T]$*

$$\begin{aligned}(u(t,\cdot),\phi) = (u(0,\cdot),\phi) + \int_0^t &[-(a^{ij} D_i u(s), D_j \phi) \\ &+ ((a^i - a^{ij}_{x^j}) D_i u(s) + a u(s) + f_r(s), \phi)] \, ds,\end{aligned}$$

where (,) *denotes the usual inner product in* $L^2(\mathbb{R}^d)$. *Quite often we write* (2.1) *and similar equations in the form*

$$du(t) = (Lu(t) + f(t))\,dt,$$

bearing in mind the differential of u *in* t *only.*

Suppose that we split (2.1) into the equations

$$D_t v(t,x) = L_r v(t,x) + f_r(t,x), \qquad t \in (0,T],\ \ x \in \mathbb{R}^d, \tag{2.3}$$

with

$$L_r := a_r^{ij}(t,x)D_{ij} + a_r^i(t,x)D_i + a_r(t,x), \quad L = \sum_{r=1}^{d_1} L_r, \quad f = \sum_{r=1}^{d_1} f_r,$$

such that these equations are more pleasant from the point of view of numerical methods than the original one. This motivates the multistage splitting method, which we describe below. First we need some assumptions.

Fix an integer $l \geq 1$.

Assumption 2.1 (ellipticity of L_r). For each $r = 1, 2, \ldots, d_1$ for $dt \times dx$-almost every $(t,x) \in [0,T] \times \mathbb{R}^d$

$$a_r^{ij}(t,x)\lambda^i\lambda^j \geq 0$$

for all $(\lambda^1, \lambda^2, \ldots, \lambda^d) \in \mathbb{R}^d$.

Assumption 2.2. (i) The partial derivatives

$$D_t^s D^\rho a_r^{ij}, \quad D_t^s D^\rho a_r^i, \quad D_t^s D^\rho a_r \quad \text{for } i,j = 1,2,\ldots,d,\ r = 1,2,\ldots,d_1$$

exist and by magnitude are bounded by K for all integers $s \geq 0$ and multi-indices ρ, satisfying $2s + |\rho| \leq l$.

(ii) For every integer $s \in [0, l/2]$

$$\sup_{t\in[0,T]} \|D_t^s f_r(t)\|_{l-2s,p} \leq K.$$

(iii) We have $u_0 \in W_p^l$ and $\|u_0\|_{l,p} \leq K$.

It is well known that under the above conditions, (2.1) and (2.3) with initial condition $u(0) = u_0$ admit unique generalized solutions u and v, respectively, which are W_p^l-valued weakly continuous functions of $t \geq 0$ (see, for instance, Theorem 3.1 below). We want to approximate the solution u by using the splitting-up method, i.e., by solving (2.3) successively with appropriate initial conditions on appropriate time intervals. Let us formulate now our splitting-up scheme in the case when the coefficients a_r^{ij}, a_r^i, a_r and free terms f_r are independent of the time variable t.

Set $T_n := \{t_i := iT/n : i = 0,1,2,\ldots,n\}$, $\delta := T/n$ for an integer $n \geq 1$. Then for fixed n we approximate the solution u of (2.1)–(2.2) at $t_i = iT/n$ recursively by $u_n(0) := u_0$,

$$u_n(t_{i+1}) := \mathbb{S}_\delta^{(d_1)} \ldots \mathbb{S}_\delta^{(2)}\mathbb{S}_\delta^{(1)} u_n(t_i), \quad i = 0,1,2,\ldots,n-1, \tag{2.4}$$

where $\mathbb{S}_t^{(r)}\psi := v(t)$ denotes the solution of (2.3) for $t \geq 0$ with initial condition $v(0) = \psi$.

It is known that if Assumptions 2.1 and 2.2 are satisfied with $l = m + 4$, then

$$\max_{t\in T_n} \|u(t) - u_n(t)\|_{m,p} \leq \frac{N}{n}$$

for all $n \geq 1$, where N depends only on d, d_1, T, K, p, m. Moreover, this rate of convergence is sharp (see [8], where this result is a special case of the rate of convergence estimates for stochastic PDEs). In the present paper we want to show that by suitable combinations of splitting-up approximations we can achieve convergence as fast as we wish. We show this by the aid of the following theorem on expansion of u_n in powers of the step size δ.

THEOREM 2.2. *Let $m \geq 0$ and $k \geq 0$ be integers. Let Assumptions 2.1 and 2.2 hold with*

$$l \geq 4 + m + 4k. \tag{2.5}$$

Suppose that the coefficients a_r^{ij}, a_r^i, a_r and the free terms f_r do not depend on t. Then for all $n \geq 1$ and $t \in T_n$ and $x \in \mathbb{R}^d$, the following representation holds:

$$\begin{aligned} u_n(t,x) = {} & u(t,x) + \delta u^{(1)}(t,x) \\ & + \delta^2 u^{(2)}(t,x) + \cdots + \delta^k u^{(k)}(t,x) + R_n^{(k)}(t,x), \end{aligned} \tag{2.6}$$

where the functions $u^{(1)}, \dots, u^{(k)}$ and $R_n^{(k)}$, defined on $[0,T]$, are W_p^m-valued and weakly continuous. Furthermore, $u^{(j)}$, $j = 1, 2, \dots, k$, are independent of n, and

$$\sup_{t\in T_n} \|R_n^{(k)}(t)\|_{m,p} \leq N\delta^{k+1} \tag{2.7}$$

for all n, where N depends only on l, k, d, d_1, K, m, p, T.

Remark 2.1. If $k = 0$ and $p = 2$, the result holds under a weaker restriction on l: $l \geq 3 + m$ (see, for instance [7]). For general $p \geq 2$ and $k = 0$ the result is proved in [8].

We prove Theorem 2.2 in section 4. Now we deduce from it a result on the acceleration of the splitting-up method. Let V denote the square matrix defined by $V^{ij} := 2^{-(i-1)(j-1)}$, $i, j = 1, \dots, k+1$. Notice that the determinant of V is the Vandermonde determinant, generated by $1, 2^{-1}, \dots, 2^{-k}$, and hence it is different from 0. Thus V is invertible. Set $\mathbf{b} := (b_0, b_1, \dots, b_k) := (1, 0, 0, \dots, 0)V^{-1}$, and define

$$v_n(t) := \sum_{j=0}^{k} b_j u_{2^j n}(t), \quad t \in T_n := \left\{ \frac{iT}{n} : i = 0, 1, \dots, n \right\},$$

where $u_{2^j n}$ is the splitting-up approximation based on the grid $T_{2^j n} := \{iT/(2^j n) : i = 0, 1, \dots, 2^j n\}$.

THEOREM 2.3. *Let $m \geq 0$ and $k \geq 0$ be any integers. Let Assumptions* 2.1 and 2.2 *hold with l satisfying* (2.5). *Suppose that the coefficients a_r^{ij}, a_r^i, a_r and the free terms f_r do not depend on t. Then*

$$\max_{t\in T_n} \|v_n(t) - u(t)\|_{m,p} \leq N\delta^{k+1},$$

where N is a constant depending only on l, k, d, d_1, K, m, p, T.

Proof. By Theorem 2.2

$$u_{2^j n} = u + \sum_{i=1}^{k} \frac{\delta^i}{2^{ji}} u^{(i)} + R^{(k)}_{2^j n}, \quad j = 0, 1, \ldots, k.$$

Therefore for all $n \geq 1$

$$v_n = \sum_{j=0}^{k} b_j u_{2^j n} = \left(\sum_{j=0}^{k} b_j \right) u + \sum_{j=0}^{k} \sum_{i=1}^{k} b_j \frac{\delta^i}{2^{ij}} u^{(i)} + \sum_{j=0}^{k} b_j R^{(k)}_{2^j n}$$

$$= u + \sum_{i=1}^{k} \delta^i u^{(i)} \sum_{j=0}^{k} \frac{b_j}{2^{ij}} + \sum_{j=0}^{k} b_j R^{(k)}_{2^{j}n} = u + \sum_{j=0}^{k} b_j R^{(k)}_{2^j n},$$

since $\sum_{j=0}^{k} b_j = 1$ and $\sum_{j=0}^{k} b_j 2^{-ij} = 0$ for $i = 1, 2, \ldots, k$ by the definition of $(b_0, \ldots, b_k)$. Hence $v_n - u = \sum_{j=0}^{k} b_j R^{(k)}_{2^j n}$ and

$$\max_{t \in T_n} \|v_n(t) - u(t)\|_{m,p} = \max_{t \in T_n} \left\| \sum_{j=0}^{k} b_j R^{(k)}_{2^j n}(t) \right\|_{m,p}$$

$$\leq \sum_{j=0}^{k} |b_j| \max_{t \in T_n} \|R^{(k)}_{2^j n}(t)\|_{m,p} \leq N\delta^{k+1},$$

by (2.7), where N is a constant depending only on l, T, K, d, d_1, m, p, k. ◻

Remark 2.2. Assume that $u^{(1)} = 0$ in expansion (2.6). This happens, for example, for Strang's splitting, which is a special case of our splitting-up scheme, as is explained in the Introduction. In this case we need take only k terms in the linear combination to achieve accuracy of order $k + 1$. Namely, we now define $v_n(t)$ by

$$v_n(t) := \sum_{j=0}^{k-1} \lambda_j u_{2^j n}(t), \quad t \in T_n,$$

where

$$(\lambda_0, \lambda_1, \ldots, \lambda_{k-1}) := (1, 0, \ldots, 0)V^{-1},$$

and V is now a $k \times k$ Vandermonde matrix with entries $V_{i1} := 1$, $V_{i,j} := 2^{-(i-1)j}$ for $i = 1, 2, \ldots, k$ and $j = 2, \ldots, k$. Then Theorem 2.3 remains valid, which one can prove in the same way as Theorem 2.3 is proved. For example,

$$v_n(t) := -\frac{1}{3} u_n(t) + \frac{4}{3} u_{2n}(t), \quad t \in T_n,$$

is an approximation of accuracy δ^3 in the case of Strang's splitting.

3. Auxiliary results. Let us consider the PDE

$$du(t, x) = (Lu(t, x) + f(t, x))\, dA(t), \qquad t \in (0, T], \ \ x \in \mathbb{R}^d, \tag{3.1}$$

$$u(0, x) = u_0(x), \quad x \in \mathbb{R}^d, \tag{3.2}$$

where L is an operator of the form

$$L = a^{ij}(t,x)D_{ij} + a^i(t,x)D_i + a(t,x),$$

$A = A(t)$ is a continuous increasing function starting from 0, and f and u_0 are real functions of $(t,x) \in (0,T] \times \mathbb{R}^d$ and of $x \in \mathbb{R}^d$, respectively. Fix an integer $l \geq 0$ and a real number $p \geq 2$. We understand the solution in the spirit of Definition 2.1 and make the following assumptions.

Assumption 3.1 (smoothness of the coefficients). The coefficients of L are measurable. The derivatives in $x \in \mathbb{R}^d$ of the coefficients a^{ij} up to order $2 \vee l$, of the coefficients $a^i(t,x)$ up to order $1 \vee l$, and of $a(t,x)$ up to order l exist for any $t \in (0,\infty)$, and by magnitude are bounded by K.

Assumption 3.2. We have

$$u_0 \in W^l_p, \quad f \in L_p([0,T], W^l_p).$$

Assumption 3.3 (ellipticity of L). For all $t \geq 0$, $x \in \mathbb{R}^d$, and $\lambda \in \mathbb{R}^d$, we have

$$a^{ij}(t,x)\lambda^i\lambda^j \geq 0.$$

Assumption 3.4. The function A is absolutely continuous and

$$\dot{A}(t) := \frac{d}{dt}A(t) \leq K$$

for dt-almost every $t \geq 0$.

The following result is well known in PDE theory (after replacing dA in (3.1) with $\dot{A}\,dt$ we easily get it, for instance, from [10] or from Theorem 3.1 in [7]).

THEOREM 3.1. *Under Assumptions* 3.1, 3.2, 3.3, *and* 3.4 *with* $l \geq 1$ *the Cauchy problem* (3.1)–(3.2) *has a unique generalized solution* u. *If Assumptions* 3.1, 3.2, 3.3, *and* 3.4 *hold with* $l \geq 0$, *and if* u *is a generalized solution of* (3.1)–(3.2), *then for every integer* $l_1 \in [0,l]$

$$\sup_{t\in[0,T]} \|u(t)\|^p_{l_1,p} \leq N\left\{\|u_0\|^p_{l_1,p} + \int_0^T \|f(t)\|^p_{l_1,p}\,dt\right\},$$

where N *is a constant depending only on* T, K, l, p, d.

Under the assumptions of Theorem 3.1 let $\mathcal{R}f$ denote the solution of (3.1) with initial data $u_0 = 0$. Then by virtue of Theorem 3.1

$$\mathcal{R} : L_p([0,T], W^l_p) \to C_w([0,T], W^l_p)$$

is a bounded linear operator, where $C_w([0,T], W^l_p)$ denotes the Banach space of weakly continuous W^l_p-valued functions $u = u(t)$, $t \in [0,T]$ with the norm $\sup_{t\in[0,T]} \|u(t)\|_{l,p}$.

Let us now consider the equation

$$du(t,x) = Lu(t,x)\,dA(t) + g(t,x)\,dH(t), \quad (t,x) \in (0,T] \times \mathbb{R}^d, \tag{3.3}$$

where g is a real-valued function of $(t,x) \in [0,T] \times \mathbb{R}^d$ and H is an absolutely continuous function of $t \in [0,T]$.

Assumption 3.5. We have $g \in L_p([0,T], W_p^{l+2})$, and there exists $g' \in L_p([0,T], W_p^l)$ such that

$$d(g(t),\phi) = (g'(t),\phi)\, dA(t), \quad t \in [0,T],$$

for all $\phi \in C_0(\mathbb{R}^d)$.

LEMMA 3.2. *Under Assumptions* 3.1, 3.3, 3.4, and 3.5 *with* $l \geq 1$, (3.3) *with zero initial data has a unique generalized solution* u. *Moreover,*

$$u = \mathcal{R}(H(Lg - g')) + Hg =: \mathcal{Q}(H,g). \tag{3.4}$$

If Assumptions 3.1, 3.3, 3.4, and 3.5 *hold with* $l \geq 0$, *and* (3.3) *with zero initial data admits a generalized solution* u, *then for every integer* $l_1 \in [0,l]$

$$\begin{aligned} &\sup_{t\in[0,T]} \|u(t)\|_{l_1,p} \\ &\leq N \sup_{t\in[0,T]} |H(t)| \left(\sup_{t\in[0,T]} \|g(t)\|_{l_1,p} + \left\{ \int_0^T (\|g(t)\|_{l_1+2,p}^p + \|g'(t)\|_{l_1,p}^p)\, dt \right\}^{1/p} \right), \end{aligned} \tag{3.5}$$

where N *is a constant depending only on* p, d, K, l, T.

Proof. Note that

$$d((g(t),\phi)H(t)) = (g(t),\phi)\, dH(t) + (g'(t),\phi)H(t)\, dA(t)$$

for all $\phi \in C_0(\mathbb{R}^d)$. Therefore u solves (3.3) when $w(t,x) := u(t,x) - H(t)g(t,x)$ solves

$$dw(t,x) := \{Lw(t,x) + H(t)(Lg(t,x) - g'(t,x))\}\, dA(t).$$

Hence equality (3.4) follows by Theorem 3.1, and it implies (3.5). $\square$

Remark 3.1. We often consider (3.3) when $dH(t)/dt$ is bounded. Then under Assumptions 3.1, 3.3, 3.4, and 3.5 with $l \geq 1$, (3.3) with zero initial data has a unique generalized solution u by Theorem 3.1. By Theorem 3.1 this solution belongs to $C_w([0,T], W_p^l)$, and its norm in this space admits an estimate with a constant depending on the bound for $\dot{H}$ and only the $L_p([0,T], W_p^l)$-norm of g. It is important that the function H enters (3.5) only through $\sup|H|$ and not any characteristic of its derivative; however, for that we pay a price requiring g to have more derivatives.

4. Proof of Theorem 2.2. Throughout this section the assumptions of Theorem 2.2 are supposed to be satisfied. In particular, $l \geq 4$. Fix n and introduce $\delta = T/n$. We use the idea from [7] and [8] of rearranging the splitting method in forward time. We achieve this by considering the equation

$$dw(t,x) = \sum_{r=1}^{d_1} (L_r w(t,x) + f_r)\, dA_r(t), \quad w(0,x) = u_0(x), \tag{4.1}$$

where the time change A_r, $r = 1, \dots, d_1$, is defined by the requirements that $A_r(0) = 0$, $A_r(t)$ be absolutely continuous, and its derivative in time $\dot{A}_r$ be periodic with period $d_1\delta$ and

$$\dot{A}_r(t) = 1_{[r-1,r]}\left(\frac{t}{\delta}\right), \quad t \in [0, d_1\delta] \quad \text{(a.e.)}. \tag{4.2}$$

Instead of the original Cauchy problem (2.1)–(2.2) we consider

$$dv(t,x) = (Lv(t,x) + f)\,dA_0(t), \quad v(0,x) = u_0(x),$$

where

$$A_0(t) := \frac{t}{d_1}. \tag{4.3}$$

Clearly, $v(t) = u(A_0(t))$ and

$$v(d_1 t) = u(t), \quad w(d_1 t) = u_n(t) \quad \text{for all } t \in T_n.$$

Therefore our aim is to show that Theorem 2.2 holds with v and w in place of u, and u_n, respectively, for all $t = id_1\delta$, $i = 0, 1, \dots, n$. To this end first we introduce some notation. We call a sequence of numbers $\alpha = \alpha_1\alpha_2\dots\alpha_i$ a multinumber of length $|\alpha| := i$ if $\alpha_j \in \{0, 1, 2, \dots, d_1\}$. The reader should notice the difference between multinumbers and multi-indices. The set of all multinumbers is denoted by $\mathcal{N}$. For every multinumber α we define a function $B_\alpha : [0,\infty) \to \mathbb{R}$ and a number c_α recursively starting as follows:

$$B_\gamma := \delta^{-1}(A_\gamma - A_0), \quad c_\gamma = 0 \qquad \text{for } \gamma = 0, 1, 2, \dots, d_1. \tag{4.4}$$

If for every multinumber $\beta = \beta_1\dots\beta_i$ of length i the function B_β and the number c_β are defined, then

$$c_{\beta\gamma} := \delta^{-1}\int_0^{d_1\delta} B_\beta(s)\dot{A}_\gamma(s)\,ds, \tag{4.5}$$

$$B_{\beta\gamma}(t) := \delta^{-1}\int_0^t (B_\beta(s)\dot{A}_\gamma(s) - c_{\beta\gamma}\dot{A}_0(s))\,ds \tag{4.6}$$

for $\gamma = 0, 1, 2, \dots, d_1$, where $\dot{A}_\gamma(s) := dA_\gamma(s)/ds$.

Notice that by (4.6) we have

$$B_\beta(t)\,dA_\gamma(t) = c_{\beta\gamma}\,dA_0(t) + \delta\,dB_{\beta\gamma}(t) \tag{4.7}$$

for all multinumbers β and $\gamma = 0, 1, 2, \dots, d_1$. We will often make use of this equality and of the following lemma.

LEMMA 4.1. *For every $\alpha \in \mathcal{N}$ the function B_α is $d_1\delta$-periodic, i.e., $B_\alpha(t + d_1\delta) = B_\alpha(t)$ for all $t \geq 0$, and $B_\alpha(id_1\delta) = 0$ for all integer $i \geq 0$. Moreover, the numbers c_α, the functions $C_\alpha(t) := B_\alpha(\delta t)$, and*

$$\sup_{t\geq 0}|B_\alpha(t)| = \sup_{t\geq 0}|C_\alpha(t)|$$

are finite and do not depend on δ.

Proof. That the first assertion is true for $\alpha = 0, \dots, d_1$ is almost obvious. If it is true for $\alpha = \beta$, where β is a multinumber, then the integrand in (4.6) is $d_1\delta$-periodic, and by definition of $c_{\beta\gamma}$ its integral over the period is zero. It follows that the first assertion holds for $\alpha = \beta\gamma$, so the induction on the length $|\alpha|$ finishes the proof of the first assertion.

To prove the second one we again use the induction on the length $i = |\alpha|$. This statement is true when $|\alpha| = 1$. Assume that it is true for all multinumbers β of length i and notice that, according to (4.2) and (4.3), $\dot{A}_\gamma(\delta s)$ are d_1-periodic in s and independent of δ. Therefore,

$$c_{\beta\gamma} = \frac{1}{\delta}\int_0^{d_1\delta} B_\beta(s)\dot{A}_\gamma(s)\,ds = \int_0^{d_1} C_\beta(s)\dot{A}_\gamma(\delta s)\,ds$$

is independent of δ by the induction hypothesis. Similar argument works for $C_\alpha(t)$. $\square$

We use the notation $\mathcal{R}f$ and $\mathcal{Q}_\alpha g$, $\alpha \in \mathcal{N}$, for the solutions of (3.1) and (3.3), respectively, with zero initial condition and A_0 and B_α in place of A and H, respectively. Notice that, unlike in the case of uniformly parabolic operators, $\mathcal{R}$ and $\mathcal{Q}_\alpha$ do not increase regularity.

The following lemma exhibits our two main technical tools: centering B_α and integrating by parts with respect to t.

LEMMA 4.2. *Take some functions*

$$h \in L_p([0,T], W_p^1), \quad h_r \in L_p([0,T], W_p^1), \quad r = 1, \dots, d_1, \quad h_0 = 0.$$

Let u be a solution of the "equation"

$$du = \sum_{r=1}^{d_1} h_r\,dA_r,$$

$u(0) \in W_p^1$, which is a particular case of (4.1) *when $L_r \equiv 0$. Finally, let $Lu \in L_p([0,T], W_p^1)$. Then for any $\alpha \in \mathcal{N}$*

$$\mathcal{R}(B_\alpha h) = c_{\alpha 0}\mathcal{R}h + \delta\mathcal{Q}_{\alpha 0}h, \tag{4.8}$$

$$\mathcal{Q}_\alpha u = \mathcal{R}(c_{\alpha 0}Lu - c_{\alpha r}h_r) + \delta\mathcal{Q}_{\alpha 0}Lu - \delta\mathcal{Q}_{\alpha r}h_r + B_\alpha u. \tag{4.9}$$

Proof. To prove (4.8) it suffices to use the definitions of $\mathcal{R}$ and $\mathcal{Q}_\beta$ (see Theorem 3.1 and Lemma 3.2) and that by virtue of (4.7) for $\varphi = \mathcal{R}(B_\alpha h)$ we have

$$d\varphi = L\varphi\,dA_0 + B_\alpha h\,dA_0 = L\varphi\,dA_0 + c_{\alpha 0}h\,dA_0 + \delta h\,dB_{\alpha 0}.$$

To prove (4.9) observe that, by definition, $\theta := \mathcal{Q}_\alpha u$ satisfies

$$d\theta = L\theta\,dA_0 + u\,dB_\alpha, \quad \theta(0) = 0.$$

This and (4.7) imply that $\psi := \theta - uB_\alpha$ satisfies $\psi(0) = 0$ and

$$\begin{aligned} d\psi &= L\psi\,dA_0 + LuB_\alpha\,dA_0 - h_rB_\alpha\,dA_r \\ &= L\psi\,dA_0 + c_{\alpha 0}Lu\,dA_0 + \delta Lu\,dB_{\alpha 0} - c_{\alpha r}h_r\,dA_0 - \delta h_r\,dB_{\alpha r}. \end{aligned}$$

Now (4.9) follows from the definitions of $\mathcal{R}$ and $\mathcal{Q}_\beta$. The proof of the lemma is complete. $\square$

Now we introduce some differential operators L_γ and functions f_γ defined for multinumbers γ as follows: $L_0 := 0$, $f_0 := 0$,

$$L_\gamma := L_r, \quad f_\gamma := f_r$$

for $\gamma = r \in \{1, 2, \ldots, d_1\}$, and

$$L_{\gamma 0} := L L_\gamma, \quad L_{\gamma r} := -L_\gamma L_r,$$

$$f_{\gamma 0} := L f_\gamma, \quad f_{\gamma r} := -L_\gamma f_r$$

for $r = 1, 2, \ldots, d_1$.

In this notation we have the following.

LEMMA 4.3. *Let $\alpha, \beta \in \mathcal{N}$ and $2|\beta| + 3 \le l$. Then*

$$\mathcal{Q}_\alpha(L_\beta w + f_\beta) = c_{\alpha r}\mathcal{R}(L_{\beta r} w + f_{\beta r}) + \delta \mathcal{Q}_{\alpha r}(L_{\beta r} w + f_{\beta r}) + B_\alpha(L_\beta w + f_\beta). \tag{4.10}$$

Proof. It follows from formula (4.9) applied to $u := L_\beta w + f_\beta$, when $h_r = L_\beta L_r w + L_\beta f_r$ (remember f_r are independent of t), that the left-hand side of (4.10) equals

$$\mathcal{R}[c_{\alpha 0}(L L_\beta w + L f_\beta) - c_{\alpha r}(L_\beta L_r w + L_\beta f_r)]$$

$$+ \delta \mathcal{Q}_{\alpha 0}(L L_\beta w + L f_\beta) - \delta \mathcal{Q}_{\alpha r}(L_\beta L_r w + L_\beta f_r) + B_\alpha(L_\beta w + f_\beta),$$

which is easily seen to be equal to the right-hand side of (4.10). ∎

We derive from (4.10) one of the most important formulas.

PROPOSITION 4.4. *Let $\kappa \ge 0$ be an integer and $l \ge 2\kappa + 3$. Then*

(4.11)
$$w = v + \sum_{i=1}^{\kappa} \delta^i \sum_{|\alpha|=i} B_\alpha(L_\alpha w + f_\alpha) + \sum_{i=1}^{\kappa} \delta^i \sum_{|\alpha|=i+1} c_\alpha \mathcal{R}(L_\alpha w + f_\alpha) + \delta^{\kappa+1} r^{(\kappa)}$$

for all $t \in [0, d_1 T]$, where

$$r^{(\kappa)} = \sum_{|\alpha|=\kappa+1} \mathcal{Q}_\alpha(L_\alpha w + f_\alpha).$$

Proof. First notice that for $\varphi_0 := w - v$ we have

$$d\varphi_0 = d(w - v) = L\varphi_0 \, dA_0 + \delta(L_r w + f_r)\, dB_r,$$

which proves (4.11) for $\kappa = 0$.

Next we fix a $\kappa \ge 1$ and transform $r^{(i)}$, for $i = 0, \ldots, \kappa - 1$, by applying (4.10) with $\alpha = \beta$ and $|\alpha| = i + 1$ when $f_r \in W_p^{2|\beta|+3}$. Then we get

$$r^{(i)} = \sum_{|\alpha|=i+1, |\beta|=1} c_{\alpha\beta}\mathcal{R}(L_{\alpha\beta} w + f_{\alpha\beta}) + \delta \sum_{|\alpha|=i+1, |\beta|=1} \mathcal{Q}(L_{\alpha\beta} w + f_{\alpha\beta})$$

$$+ \sum_{|\alpha|=i+1} B_\alpha(L_\alpha w + f_\alpha) = \sum_{|\alpha|=i+1} B_\alpha(L_\alpha w + f_\alpha)$$

$$+ \sum_{|\alpha|=i+2} c_\alpha \mathcal{R}(L_\alpha w + f_\alpha) + \delta r^{(i+1)}.$$

This shows how $r^{(0)}, r^{(1)}, \ldots, r^{(\kappa)}$ are related to each other and certainly proves the proposition. ∎

Decomposition (4.11) looks very much like (2.6), the only difference being that the factors of δ^j depend on the approximating function w, and the coefficients of δ^j in the second term on the right contain B_α, which is not a power series in δ. However, observe that we have to estimate the difference $v-w$ only at the points $id_1\delta$ at which all B_α vanish.

Our next step is to "solve" (4.11) with respect to w by the method of successive iterations, that is, by substituting w given by (4.11) into the right-hand side of the same equation. In the process of doing so we encounter only one difficulty, when the second term on the right is plugged into the third and we have to develop expressions like $\mathcal{R}(B_\alpha u)$ into power series in δ. We transform these terms by using (4.8) and (4.10).

First we introduce the notation

$$w_\beta = L_\beta w + f_\beta;$$

observe that in these terms (4.10) is rewritten as

$$\mathcal{Q}_\alpha w_\beta = c_{\alpha r}\mathcal{R}w_{\beta r} + B_\alpha w_\beta + \delta\mathcal{Q}_{\alpha r}w_{\beta r}, \tag{4.12}$$

and note the following.

LEMMA 4.5. *If $\kappa \geq 0$ is an integer and $\alpha,\beta \in \mathcal{N}$ and $2(|\beta|+\kappa)+1 \leq l$, then*

$$\mathcal{R}(B_\alpha w_\beta) = \sum_{i=0}^{\kappa}\delta^i \sum_{|\gamma|=i} c_{\alpha 0\gamma}\mathcal{R}w_{\beta\gamma}$$

$$+\sum_{i=1}^{\kappa}\delta^i \sum_{|\gamma|=i-1} B_{\alpha 0\gamma}w_{\beta\gamma} + \delta^{\kappa+1}\sum_{|\gamma|=\kappa}\mathcal{Q}_{\alpha 0\gamma}w_{\beta\gamma}, \tag{4.13}$$

where for any multinumbers μ,ν

$$\sum_{|\gamma|=0} c_{\nu\gamma}\mathcal{R}w_{\mu\gamma} := c_\nu \mathcal{R}w_\mu, \qquad \sum_{|\gamma|=0} B_{\nu\gamma}w_{\mu\gamma} := B_\nu w_\mu,$$

$$\sum_{|\gamma|=0}\mathcal{Q}_{\nu\gamma}w_{\mu\gamma} := \mathcal{Q}_\nu w_\mu.$$

Proof. If $\kappa = 0$, (4.13) follows from (4.8). If $\kappa \geq 1$, by applying (4.12) repeatedly as in the proof of Proposition 4.4 we find

$$\mathcal{Q}_\alpha w_\beta = \sum_{i=0}^{\kappa-1}\delta^i \sum_{|\gamma|=i+1} c_{\alpha\gamma}\mathcal{R}w_{\beta\gamma}$$

$$+\sum_{i=0}^{\kappa-1}\delta^i \sum_{|\gamma|=i} B_{\alpha\gamma}w_{\beta\gamma} + \delta^\kappa \sum_{|\gamma|=\kappa}\mathcal{Q}_{\alpha\gamma}w_{\beta\gamma}.$$

We use this formula for $\alpha 0$ in place of α and finish the proof by referring to (4.8). ∎

Let $\mathcal{M}$ denote the set of multinumbers $\gamma_1\gamma_2\dots\gamma_i$ with $\gamma_j \in \{1,2,\dots,d_1\}$, $j = 1,2,\dots,i$, and integers $i \geq 1$.

LEMMA 4.6. *The following statements hold:*

(i) *Let* $\gamma = \gamma_1\gamma_2\dots\gamma_i \in \mathcal{M}$ *be such that* $|\gamma| = i \le 1 + l/2$. *Then*

$$L_\gamma = (-1)^{|\gamma|-1}L_{\gamma_1}\dots L_{\gamma_i}, \quad f_\gamma = (-1)^{|\gamma|-1}L_{\gamma_1}\dots L_{\gamma_{i-1}}f_{\gamma_i}.$$

(ii) *Let* $\beta, \gamma \in \mathcal{M}$ *be such that* $|\beta| + |\gamma| \le 1 + l/2$. *Then*

$$L_\beta L_\gamma = -L_{\beta\gamma}, \quad L_\beta f_\gamma = -f_{\beta\gamma}.$$

(iii) *Let* $\alpha \in \mathcal{N}$ *be such that* $\rho := |\alpha| \le 1 + l/2$. *Then there exist constants* $c(\gamma) = c(\alpha, \gamma) \in \{0, \pm 1\}$ *defined for all* $\gamma \in \mathcal{M}$ *with* $|\gamma| = \rho$ *such that*

$$L_\alpha = \sum_{\gamma\in\mathcal{M},|\gamma|=\rho} c(\gamma)L_\gamma, \quad f_\alpha = \sum_{\gamma\in\mathcal{M},|\gamma|=\rho} c(\gamma)f_\gamma. \tag{4.14}$$

Proof. Part (i) follows immediately from the definition of L_γ, f_γ by induction on $|\gamma|$. Part (i) obviously implies part (ii). Part (iii) clearly holds for $\alpha = 0$ and $\alpha = r \in \{1, \dots, d_1\}$. Assume that (4.14) holds for some $\alpha \in \mathcal{N}$, $|\alpha| < 1 + l/2$. Then

$$L_{\alpha r} = -L_\alpha L_r = -\sum_{|\gamma|=|\alpha|} c(\gamma)L_\gamma L_r = \sum_{|\gamma|=|\alpha|} c(\gamma)L_{\gamma r},$$

$$f_{\alpha r} = -L_\alpha f_r = -\sum_{|\gamma|=|\alpha|} c(\gamma)L_\gamma f_r = \sum_{|\gamma|=|\alpha|} c(\gamma)f_{\gamma r}$$

for $r \in \{1, 2, \dots, d_1\}$, and

$$L_{\alpha 0} = LL_\alpha = \sum_{r=1}^{d_1} L_r \sum_{\gamma\in\mathcal{M},|\gamma|=\rho} c(\gamma)L_\gamma = -\sum_{r=1}^{d_1} \sum_{\gamma\in\mathcal{M},|\gamma|=\rho} c(\gamma)L_{r\gamma},$$

$$f_{\alpha 0} = Lf_\alpha = \sum_{r=1}^{d_1} L_r \sum_{\gamma\in\mathcal{M},|\gamma|=\rho} c(\gamma)f_\gamma = -\sum_{r=1}^{d_1} \sum_{\gamma\in\mathcal{M},|\gamma|=\rho} c(\gamma)f_{r\gamma},$$

which proves (iii) by induction on $|\alpha|$. ∎

We introduce sequences $\sigma = (\beta_1, \beta_2, \dots, \beta_i)$ of multinumbers $\beta_j \in \mathcal{M}$, where $i \ge 1$ is any integer, and set $|\sigma| := |\beta_1| + |\beta_2| + \dots + |\beta_i|$. We consider also the "empty sequence" e of length $|e| = 0$, and denote the set of all these sequences by $\mathcal{J}$. For $\sigma = (\beta_1, \beta_2, \dots, \beta_i)$, $i \ge 1$, we define

$$S_\sigma = \mathcal{R}L_{\beta_1} \cdot \ \cdots \ \cdot \mathcal{R}L_{\beta_i},$$

and for $\sigma = e$ we set

$$S_e = \mathcal{R}.$$

Notice that S_σ involves $2|\sigma|$ derivatives with respect to x and a certain number of operators $\mathcal{R}$ which do not increase regularity. Therefore, basically, S_σ has the power of a differential operator of $2|\sigma|$th order. If we have a collection of functions g_ν indexed by a parameter ν taking values in a set A, then we use the notation

$$\sum_{\nu\in A}^{*} g_\nu$$

for any linear combination of g_ν with coefficients independent of the argument of g_ν and of δ. For instance,

$$\sum_A^* S_\sigma w_\gamma = \sum_{(\sigma,\gamma)\in A}^* S_\sigma w_\gamma = \sum_{(\sigma,\gamma)\in A} c(\sigma,\gamma) S_\sigma w_\gamma,$$

where $c(\sigma,\gamma)$ are certain constants independent of δ. These constants are allowed to change from one occurrence to another.

For functions $u = u(t,x) = u(\delta,t,x)$ depending on the parameter δ we write $u = O_m(\delta^\kappa)$ if

$$\sup_\delta \delta^{-\kappa} \sup_{t\in[0,d_1T]} \|u(t)\|_{m,p} < \infty.$$

We also use the following sets:

$$A(i) = \{(\sigma,\beta) : \sigma \in \mathcal{J}, \beta \in \mathcal{M}, |\sigma| + |\beta| \le i\},$$

$$B(i,j) = \{(\alpha,\beta) : \alpha \in \mathcal{N}, \beta \in \mathcal{M}, |\alpha| = i, |\beta| \le j\}.$$

LEMMA 4.7. *Let $\kappa, \mu \ge 0$ be integers and $\alpha \in \mathcal{N}, \beta \in \mathcal{M}$, $\sigma \in \mathcal{J}$. Assume that*

$$2(|\sigma| + |\beta| + \kappa) + \mu + 2 \le l. \tag{4.15}$$

Then

$$S_\sigma(B_\alpha w_\beta) = \sum_{i=0}^{\kappa} \delta^i \sum_{A(|\sigma|+|\beta|+i)}^* S_{\sigma_1} w_{\beta_1}$$

$$+ \sum_{i=1}^{\kappa} \delta^i \sum_{B(|\alpha|+i,|\sigma|+|\beta|+i-1)}^* B_{\alpha_1} w_{\beta_1} + O_\mu(\delta^{\kappa+1}). \tag{4.16}$$

Proof. For $\sigma = e$, when $S_\sigma = \mathcal{R}$, equation (4.16) turns out to be just a different form of (4.13), which is applicable since $2(|\beta| + \kappa) + 1 \le l$. Indeed, owing to Lemma 4.6(iii),

$$\sum_{|\gamma|=i} c_{\alpha 0\gamma} \mathcal{R} w_{\beta\gamma} = \sum_{A(|\sigma|+|\beta|+i)}^* S_{\sigma_1} w_{\beta_1},$$

$$\sum_{|\gamma|=i-1} B_{\alpha 0\gamma} w_{\beta\gamma} = \sum_{B(|\alpha|+i,|\sigma|+|\beta|+i-1)}^* B_{\alpha_1} w_{\beta_1}.$$

Furthermore, for $|\gamma| = \kappa$ (see Remark 3.1)

$$\mathcal{Q}_{\alpha 0\gamma} w_{\beta\gamma} = O_\mu(1) \quad \text{since} \quad 2(|\beta| + \kappa) + \mu + 2 \le l.$$

For $|\sigma| \ge 1$ we proceed by induction on the length $\ell(S_\sigma)$ of $S_\sigma = \mathcal{R}L_{\beta_1} \cdot \cdots \cdot \mathcal{R}L_{\beta_j}$, which we define to be j. If $\ell(S_\sigma) = 1$, then $S_\sigma = \mathcal{R}L_\nu$ for a $\nu \in \mathcal{M}$ with $\nu = \sigma$, and it suffices to notice that

$$S_\sigma(B_\alpha w_\beta) = \mathcal{R}L_\nu(B_\alpha w_\beta) = -\mathcal{R}(B_\alpha w_{\nu\beta}) = -S_e(B_\alpha w_{\beta'}), \tag{4.17}$$

where $\beta' = \nu\beta \in \mathcal{M}$ and $2(|\beta'| + \kappa) + \mu + 2 = 2(|\sigma| + |\beta| + \kappa) + \mu + 2 \le l$.

Assume that (4.16) holds whenever $\ell(S_\sigma) = s$, and take an S_σ such that $\ell(S_\sigma) = s+1$. Then $S_\sigma = \mathcal{R}L_\nu S_{\sigma'}$, where $\nu, \sigma' \in \mathcal{M}$, $|\nu| + |\sigma'| = |\sigma|$, and $\ell(S_{\sigma'}) = s$. Furthermore,

$$2(|\sigma'| + |\beta| + \kappa) + \mu' + 2 \le l,$$

where $\mu' = \mu + 2|\nu|$. By the induction hypothesis,

$$S_{\sigma'}(B_\alpha w_\beta) = \sum_{i=0}^{\kappa} \delta^i \sideset{}{^*}\sum_{A(|\sigma'|+|\beta|+i)} S_{\sigma_1} w_{\beta_1}$$

$$+ \sum_{i=1}^{\kappa} \delta^i \sideset{}{^*}\sum_{B(|\alpha|+i, |\sigma'|+|\beta|+i-1)} B_{\alpha_1} w_{\beta_1} + O_{\mu'}(\delta^{\kappa+1}).$$

We apply $\mathcal{R}L_\nu$ to both parts of this equality and take into account that $L_\nu w_{\beta_1} = -w_{\nu\beta_1}$ and $|\nu| + |\sigma'| = |\sigma|$. Then similarly to (4.17) we get that

$$S_\sigma(B_\alpha w_\beta) = \sum_{i=0}^{\kappa} \delta^i \sideset{}{^*}\sum_{A(|\sigma|+|\beta|+i)} S_{\sigma_1} w_{\beta_1}$$

$$+ \sum_{i=1}^{\kappa} \delta^i \sideset{}{^*}\sum_{B(|\alpha|+i, |\sigma|+|\beta|+i-1)} S_e(B_{\alpha_1} w_{\beta_1}) + O_\mu(\delta^{\kappa+1}). \tag{4.18}$$

Now we transform the second term on the right. Take $(\alpha_1, \beta_1) \in B(|\alpha| + i, |\sigma| + |\beta| + i - 1)$ and notice that then $|\beta_1| \le |\sigma| + |\beta| + i - 1$. Hence, by assumption (4.15),

$$2(|\beta_1| + \kappa - i) + \mu + 2 < l.$$

Therefore, by the result for $\sigma = e$,

$$S_e(B_{\alpha_1} w_{\beta_1}) = \sum_{j=0}^{\kappa-i} \delta^j \sideset{}{^*}\sum_{A(|\beta_1|+j)} S_{\sigma_2} w_{\beta_2}$$

$$+ \sum_{j=1}^{\kappa-i} \delta^j \sideset{}{^*}\sum_{B(|\alpha_1|+j, |\beta_1|+j-1)} B_{\alpha_2} w_{\beta_2} + O_\mu(\delta^{\kappa-i+1}).$$

We substitute this result into (4.18) and obtain (4.16) after collecting the coefficients of δ^{i+j} and noticing that, if $(\alpha_1, \beta_1) \in B(|\alpha| + i, |\sigma| + |\beta| + i - 1)$ and $(\alpha_2, \beta_2) \in B(|\alpha_1| + j, |\beta_1| + j - 1)$, then

$$(\alpha_2, \beta_2) \in B(|\alpha| + i + j, |\sigma| + |\beta| + i + j - 1).$$

This justifies the induction and finishes the proof of the lemma. $\square$

We remind the reader that throughout this section the assumptions of Theorem 2.2 are supposed to be satisfied, and in the following proposition we use the notation

$$B^*(i,j) = \bigcup_{i_1=1}^{i} B(i_1, j).$$

PROPOSITION 4.8. *For any $j = 0, 1, \ldots, k$ we have*

$$w = v + \sum_{i=1}^{j} \delta^i \sum_{A(2i)}^{*} S_\sigma v_\beta + \sum_{i=j+1}^{k} \delta^i \sum_{A(i+j+1)}^{*} S_{\sigma_1} w_{\beta_1}$$

$$+ \sum_{i=1}^{k} \delta^i \sum_{B^*(i,i+j)}^{*} B_{\alpha_1} w_{\beta_1} + O_m(\delta^{k+1}), \tag{4.19}$$

where $v_\beta := L_\beta v + f_\beta$.

Proof. By Proposition 4.4 (notice that, due to (2.5), $l \geq 2k + 3$ and $2(k + 1) + m + 2 \leq l$) we have

$$w = v + \sum_{i=1}^{k} \delta^i \sum_{|\beta|=i} B_\beta w_\beta + \sum_{i=1}^{k} \delta^i \sum_{|\beta|=i+1} c_\beta \mathcal{R} w_\beta + O_m(\delta^{k+1}),$$

which means that (4.19) holds for $j = 0$, since by Lemma 4.6(iii),

$$\sum_{|\beta|=i} B_\beta w_\beta = \sum_{|\beta|=i} B_\beta \sum_{\gamma \in \mathcal{M}, |\gamma|=i} c(\beta, \gamma) w_\gamma = \sum_{B^*(i,i)}^{*} B_{\alpha_1} w_{\beta_1},$$

$$\sum_{|\beta|=i+1} c_\beta \mathcal{R} w_\beta = \sum_{|\beta|=i+1} c_\beta \sum_{\gamma \in \mathcal{M}, |\gamma|=i+1} c(\beta, \gamma) \mathcal{R} w_\gamma = \sum_{A(i+1)}^{*} S_{\sigma_1} w_{\beta_1}.$$

Next, assume that $k \geq 1$ and that (4.19) holds for a $j \in \{0, \ldots, k-1\}$. Transform the first term with $i = j+1$ in the second sum on the right in (4.19) by using Lemma 4.7. To prepare the transformation take $(\sigma_1, \beta_1) \in A(2i) = A(i + j + 1)$ so that $|\sigma_1| + |\beta_1| \leq 2i$ and apply the operator $S_{\sigma_1} L_{\beta_1}$ to both parts of (4.11) with $k - i$ in place of κ. Then we obtain

$$S_{\sigma_1} w_{\beta_1} = S_{\sigma_1} v_{\beta_1} + \sum_{i_1=1}^{k-i} \delta^{i_1} \sum_{|\alpha_1|=i_1} S_{\sigma_1}(B_{\alpha_1} L_{\beta_1} w_{\alpha_1})$$

$$+ \sum_{i_1=1}^{k-i} \delta^{i_1} \sum_{|\alpha_1|=i_1+1} c_{\alpha_1} S_{\sigma_1} L_{\beta_1} \mathcal{R} w_{\alpha_1} + \delta^{k-i+1} r^{(k-i)},$$

where

$$r^{(k-i)} = \sum_{|\alpha|=k-i+1} S_{\sigma_1} L_{\beta_1} \mathcal{Q}_\alpha w_\alpha.$$

Owing to

$$l - 2(k - i + 1 + |\beta_1| + |\sigma_1|) \geq l - 2(k + i + 1)$$
$$\geq l - 2(2k + 1) \geq m + 2,$$

we have $r^{(k-i)} = O_m(1)$. By the way, this is the only place where we need l to be not smaller than $4+m+4k$. Hence by Lemma 4.6(iii),

$$S_{\sigma_1}w_{\beta_1} = S_{\sigma_1}v_{\beta_1} + \sum_{i_1=1}^{k-i}\delta^{i_1} \sideset{}{^*}\sum_{(\alpha_2,\beta_2)\in B(i_1,|\beta_1|+i_1)} S_{\sigma_1}(B_{\alpha_2}w_{\beta_2})$$

$$+\sum_{i_1=1}^{k-i}\delta^{i_1} \sideset{}{^*}\sum_{A(|\sigma_1|+|\beta_1|+i_1+1)} S_{\sigma_2}w_{\beta_2} + O_m(\delta^{k-i+1}) =: J_1+\cdots+J_4. \tag{4.20}$$

Now Lemma 4.7 with $k-i-i_1$ in place of κ and m in place of μ allows us to transform terms entering J_2. For $|\beta_2|\le|\beta_1|+i_1$ we have (remember that $(\sigma_1,\beta_1)\in A(2i)$)

$$2(|\sigma_1|+|\beta_2|+k-i-i_1)+m+2 \le 2(|\sigma_1|+|\beta_1|+k-i)+m+2$$
$$\le 2(i+k)+m+2 \le 4k+m+2 < l.$$

Therefore

$$S_{\sigma_1}(B_{\alpha_2}w_{\beta_2}) = \sum_{i_2=0}^{k-i-i_1}\delta^{i_2} \sideset{}{^*}\sum_{A(|\sigma_1|+|\beta_2|+i_2)} S_{\sigma_3}w_{\beta_3}$$
$$+\sum_{i_2=1}^{k-i-i_1}\delta^{i_2} \sideset{}{^*}\sum_{B(|\alpha_2|+i_2,|\sigma_1|+|\beta_2|+i_2-1)} B_{\alpha_3}w_{\beta_3} + O_m(\delta^{k-i-i_1+1}).$$

We plug this result into J_2 and, in order to collect the coefficients of $\delta^{i_1+i_2}$, notice that for $(\sigma_3,\beta_3)\in A(|\sigma_1|+|\beta_2|+i_2)$ and $(\alpha_2,\beta_2)\in B(i_1,|\beta_1|+i_1)$ it holds that

$$|\sigma_3|+|\beta_3| \le |\sigma_1|+|\beta_2|+i_2 \le |\sigma_1|+|\beta_1|+i_1+i_2.$$

Furthermore, if $(\alpha_3,\beta_3)\in B(|\alpha_2|+i_2,|\sigma_1|+|\beta_2|+i_2-1)$, then

$$|\alpha_3|=|\alpha_2|+i_2=i_1+i_2, \quad |\beta_3|\le|\sigma_1|+|\beta_2|+i_2-1<|\sigma_1|+|\beta_1|+i_1+i_2.$$

It follows that J_2 is written as

$$\sum_{i_1=1}^{k-i}\delta^{i_1}\left(\sideset{}{^*}\sum_{A(|\sigma_1|+|\beta_1|+i_1)} S_{\sigma_2}w_{\beta_2} + \sideset{}{^*}\sum_{B(i_1,|\sigma_1|+|\beta_1|+i_1)} B_{\alpha_2}w_{\beta_2}\right) + O_m(\delta^{k-i+1}),$$

which just amounts to saying that visually in the definition of J_2 one can erase S_{σ_1}, carry all differentiations in it onto w_{β_2}, and still preserve (4.20).

Then we see that

$$\delta^{j+1}\sideset{}{^*}\sum_{A(2j+2)} S_{\sigma_1}w_{\beta_1} = O_m(\delta^{k+1}) + \delta^{j+1}\sideset{}{^*}\sum_{A(2j+2)} S_{\sigma_1}v_{\beta_1}$$
$$+\sum_{i_1=1}^{k-j-1}\delta^{i_1+j+1}\left(\sideset{}{^*}\sum_{A(|\sigma_1|+|\beta_1|+i_1+1)} S_{\sigma_2}w_{\beta_2} + \sideset{}{^*}\sum_{B(i_1,|\sigma_1|+|\beta_1|+i_1)} B_{\alpha_2}w_{\beta_2}\right).$$

Next we notice again that, for $(\sigma_1,\beta_1)\in A(2j+2)$ and $|\sigma_2|+|\beta_2|\le|\sigma_1|+|\beta_1|+i_1+1$, we have $|\sigma_2|+|\beta_2|\le j+2+i_1+j+1$, whereas if $|\beta_2|\le|\sigma_1|+|\beta_1|+i_1$, then $|\beta_2|\le j+1+i_1+j+1$. Therefore, after changing $i_1+j+1\to i$ $(\ge j+2)$, we get

$$\delta^{j+1}\sum_{A(2j+2)}^{*}S_{\sigma_1}w_{\beta_1}=O_m(\delta^{k+1})+\delta^{j+1}\sum_{A(2j+2)}^{*}S_{\sigma_1}v_{\beta_1}$$

$$+\sum_{i=j+2}^{k}\delta^i\left(\sum_{A(i+j+2)}^{*}S_{\sigma_2}w_{\beta_2}+\sum_{B^*(i,i+j+1)}^{*}B_{\alpha_2}w_{\beta_2}\right).$$

This shows that the term with $i=j+1$ in the second sum on the right-hand side in (4.19) can be eliminated at the expense of changing j to $j+1$ in the other terms of (4.19). Thus the induction on j proves the proposition indeed. □

Now we finish the proof of Theorem 2.2. By taking $j=k$ in Proposition 4.8, we find

$$w=v+\sum_{i=1}^{k}\delta^i w^{(i)}+\sum_{B^*(k,2k)}c(\alpha,\beta,\delta)B_\alpha w_\beta+O_m(\delta^{k+1}), \tag{4.21}$$

where

$$w^{(i)}:=\sum_{A(2i)}^{*}S_\sigma v_\beta\in C_w([0,T],W_p^m),\quad i=1,2,\dots,k,$$

are independent of δ and $c(\alpha,\beta,\delta)$ are certain constants. It is not hard to follow our computations in order to see that

$$\sup_{t\in[0,d_1T]}\ \sup_{n,\delta=T/n}\delta^{-(k+1)}\|O_m(\delta^{k+1})(t,\cdot)\|_{m,p}\le N,$$

where the constant N depends only on d,d_1,T,K,k,m,p,l. After that, to finish the proof it remains only to recall that $B_\alpha(jd_1\delta)=0$ for all integers $j\ge0$ and

$$v(d_1t)=u(t),\quad w(d_1t)=u_n(t)\quad\forall t\in T_n=\left\{\frac{iT}{n}:i=0,1,2,\dots,n\right\}.$$

5. The case of time dependent coefficients. We consider here the Cauchy problem (1.1)–(1.2) for time dependent coefficients. We split, as before, the coefficients and the free terms into d_1 terms,

$$a^{ij}=\sum_{r=1}^{d_1}a_r^{ij},\quad a^i=\sum_{r=1}^{d_1}a_r^i,\quad a=\sum_{r=1}^{d_1}a_r,\quad f=\sum_{r=1}^{d_1}f_r;$$

define $\delta=T/n$, $t_i=t_i^n=\delta i$, $T_n=\{t_i:i=0,1,\dots,n\}$; and keep Assumptions 2.1 and 2.2. As before, we also denote $L_r:=a_r^{ij}D_{ij}+a_r^iD_i+a_r$.

One of the splitting-up approximations $u_n(t)$ for $t\in T_n$ is defined as follows. Let $\mathbb{S}_{st}^{(r)}$ be the operator mapping each function φ of an appropriate class into the solution of the problem

$$D_tv(t,x)=d_1L_rv(t,x)+d_1f_r(t,x),\quad t>s,\quad v(s,x)=\varphi(x).$$

Then the approximations are introduced according to

$$u_n(0) := u_0,$$

$$u_n(t_{i+1}) := \mathbb{S}^{(d_1)}_{t_{i\bar{d}},t_{i+1}} \dots \mathbb{S}^{(2)}_{t_{i1},t_{i2}} \mathbb{S}^{(1)}_{t_i,t_{i1}} u_n(t_i), \qquad i = 0, 1, 2, \dots, n, \tag{5.1}$$

where $t_{ij} := t_i + j\delta/d_1$ for $j = 1, 2, \dots, d_1 - 1$, $\bar{d} := d_1 - 1$.

There are many other ways to extend the splitting-up approximations (2.4) to PDEs with time dependent data. Along with (5.1) we also consider another approximation, which has the advantage that in each step we need to solve a time independent PDE, which is usually more convenient in practice than solving time dependent PDEs. This time we define the approximation u_n by

$$u_n(0) := u_0,$$

$$u_n(t^n_{i+1}) := \mathbb{S}^{(d_1)}_{\delta}(t^n_{i+1}) \dots \mathbb{S}^{(2)}_{\delta}(t^n_{i+1}) \mathbb{S}^{(1)}_{\delta}(t^n_{i+1}) u_n(t^n_i), \qquad i = 0, 1, 2, \dots, n, \tag{5.2}$$

where $\mathbb{S}^{(r)}_{\delta}(s)\varphi$ denotes the solution of the problem

$$D_t v(t) = L_r(s) v(t) + f(s), \quad t \geq 0, \qquad v(0) = \varphi, \tag{5.3}$$

with

$$L_r(s) := a^{ij}_r(s,x) D_{ij} + a^i_r(s,x) D_i + a_r(s,x)$$

for $r = 1, 2, \dots, d_1$. Notice that the coefficients of the operators $L_r(s)$ and $f(s)$ are "frozen" at time s; thus (5.3) is a Cauchy problem with time independent data.

We extend Theorem 2.2 as follows.

THEOREM 5.1. *Let $m \geq 0$ and $k \geq 0$ be any integers. Let Assumptions* 2.1 *and* 2.2 *hold with $l \geq 4 + m + 4k$. Let the splitting-up approximation u_n be defined by* (5.1) *or by* (5.2). *Then there exist functions $u^{(j)} \in C_w([0,T], W^m_p)$, $j = 1, 2, \dots, k$, $R^{(k)}_n \in C_w([0,T], W^m_p)$, such that*

$$\begin{aligned} u_n(t,x) &= u(t,x) + \delta u^{(1)}(t,x) \\ &\quad + \delta^2 u^{(2)}(t,x) + \dots + \delta^k u^{(k)}(t,x) + R^{(k)}_n(t,x) \end{aligned} \tag{5.4}$$

for all $t \in T_n$, $x \in \mathbb{R}^d$, and $n \geq 1$. The functions $u^{(j)}$, $j = 1, 2, \dots, k$, are independent of n, and

$$\sup_{t \in [0,T]} \|R^{(k)}_n(t)\|_{m,p} \leq N \delta^{k+1}$$

for all n, where N depends only on k, d, d_1, K, m, p, T.

Clearly Theorem 5.1 implies that we can again accelerate the convergence of the splitting-up approximations by considering

$$v_n(t,\cdot) := \sum_{j=0}^{k} b_j u_{2^j n}(t,\cdot), \quad t \in T_n,$$

where $u_{2^j n}$ for all $j = 0, 1, \dots, k$ are defined by either (5.1) or (5.2).

THEOREM 5.2. *Let $m \geq 0$ and $k \geq 0$ be any integers. Let Assumptions* 2.1 *and* 2.2 *hold with $l \geq 4+m+4k$. Then for all $n \geq 1$*

$$\max_{t\in T_n} \|v_n(t) - u(t)\|_{m,p} \leq N\delta^{k+1},$$

where N is a constant depending only on k, d, d_1, K, m, p, T.

Hence by Sobolev's theorem on embedding W_p^m into C^s, we immediately get the following result.

THEOREM 5.3. *Let $m \geq 0$ and $k \geq 0$ be any integers. Let Assumptions* 2.1 *and* 2.2 *hold with $l \geq 4+m+4k$. Let $s \geq 0$ be an integer such that $m \geq s + d/p$. Then for all $n \geq 1$*

$$\max_{t\in T_n} \sup_{x\in\mathbb{R}^d} \sum_{|\rho|\leq s} |D^\rho v_n(t,x) - D^\rho u(t,x)| \leq N\delta^{k+1},$$

where N is a constant depending only on k, d, d_1, K, m, s, p, T.

We prove Theorem 5.1 by adapting the proof of Theorem 2.2 to the time dependent case. If u_n is defined by (5.1), then we consider the problems

$$dv(t) = \big(L(A_0(t))v(t) + f(A_0(t))\big)\, dA_0(t), \quad v(0) = u_0, \tag{5.5}$$

$$dw(t) = \sum_{r=1}^{d_1} \big(L_r(A_0(t))w(t) + f_r(A_0(t))\big)\, dA_r(t), \quad w(0) = u_0, \tag{5.6}$$

where $A_0(t)$, $A_1(t), A_2(t), \ldots, A_{d_1}(t)$ are defined by (4.2) and (4.3), and $L(A_0(t))$, $L_r(A_0(t))$ mean that we substitute $A_0(t)$ in place of the time variable t of the coefficients of L, L_r.

If u_n is defined by (5.2), then we consider problems (5.5) and (5.6) with absolutely continuous functions $A_0, A_1, \ldots, A_{d_1}$, defined by the following requirements:

$$A_r(0) = 0, \quad \dot{A}_r \text{ is periodic with period } (d_1+1)\delta,$$

$$\dot{A}_r(t) = 1_{[r,r+1]}\left(\frac{t}{\delta}\right), \quad t \in [0, (d_1+1)\delta] \quad \text{for} \quad r = 0, 1, \ldots, d_1. \tag{5.7}$$

By virtue of Theorem 3.1, (5.5) and (5.6) admit unique solutions v and w, respectively. Clearly $v, w \in C_w([0, d'T], W_p^l)$, and

$$v(d't) = u(t), \quad w(d't) = u_n(t) \qquad \forall t \in T_n,$$

where $d' = d_1$ if $A_0, A_1, \ldots, A_{d_1}$ are defined by (4.2) and (4.3), and $d' = d_1 + 1$ if $A_0, A_1, \ldots, A_{d_1}$ are defined by (5.7). Therefore, our aim is to get an equality like (5.1) with v and w in place of u and u_n, respectively.

We treat the cases of two approximations simultaneously and warn the reader that, in order not to repeat the same arguments twice, we are going to use the same notation for some objects that have different meanings in each case. From now on, d' denotes d_1 if we consider the splitting-up approximations u_n defined by (5.1), and it denotes d_1+1 in the case of u_n defined by (5.2). We keep the notation $\mathcal{N}$ for the set of all multinumbers $\alpha = \alpha_1\alpha_2\ldots\alpha_j$ for $\alpha_i \in \{0, 1, 2, \ldots, d_1\}$ and integers $j \geq 1$. We also use the numbers c_α and the functions B_α, defined by (4.4), (4.5), and (4.6), with d' in place of d_1 in (4.5). Observe that, as is easy to see, Lemma 4.1 still holds with d' in place of d_1 in its formulation.

Let $\mathcal{R}f$ and $\bar{\mathcal{R}}f$ denote the solutions of the problems

$$du(t) = (Lu(t) + f(t))\,dt, \quad u(0) = 0,$$

and

$$dv(t) = (\bar{L}v(t) + f(t))\,dA_0(t), \quad v(0) = 0,$$

respectively, where $\bar{L} := L(A_0(t))$, the operator obtained from L by the substitution of $A_0(t)$ in place of t in the coefficients of L. Notice that $\bar{\mathcal{R}}$ depends on n when A_0 is defined by (5.7). Notice also that

$$\overline{\mathcal{R}f}(t,\cdot) := (\mathcal{R}f)(A_0(t),\cdot) = \bar{\mathcal{R}}\bar{f}(t,\cdot), \tag{5.8}$$

where $\bar{f}(t,\cdot) = f(A_0(t),\cdot)$. Let $\bar{Q}_\alpha f$ denote the solution of the problem

$$dv(t) = \bar{L}v(t)\,dA_0(t) + f(t)\,dB_\alpha(t), \quad v(0) = 0.$$

We modify the definition of L_α, f_α, used for time independent operators and free terms, as follows: $L_0 := 0$, $f_0 := 0$,

$$L_\gamma := L_r, \quad f_\gamma := f_r$$

for $\gamma = r \in \{1, 2, \ldots, d_1\}$, and

$$L_{\gamma 0} := LL_\gamma - \dot{L}_\gamma, \quad L_{\gamma r} := -L_\gamma L_r,$$

$$f_{\gamma 0} := Lf_\gamma - \dot{f}_\gamma, \quad f_{\gamma r} := -L_\gamma f_r \tag{5.9}$$

for $r = 1, 2, \ldots, d_1$, where $\dot{f}_\gamma := D_t f_\gamma$ and $\dot{L}_\gamma$ denotes the differential operator which we obtain from L_γ by taking the derivative in t of its coefficients. As is easy to see, L_γ and f_γ are well defined if $2(|\gamma| - 1) \le l$. We use the notation $\bar{L}_\gamma$ and $\bar{f}_\gamma$ for the operator which we obtain from L_γ by substituting $A_0(t)$ in place of t in its coefficients, and for the function obtained from f_γ by the same substitution, respectively. Then we have the following counterpart of Lemma 4.2.

Lemma 5.4. *Take some functions*

$$h \in L_p([0, d'T], W_p^1), \quad h_r \in L_p([0, d'T], W_p^1), \quad r = 0, 1, \ldots, d_1.$$

Let u be a solution of the "equation"

$$du = h_r\,dA_r = \sum_{r=0}^{d_1} h_r\,dA_r$$

with $u(0) \in W_p^1$. Assume that $Lu \in L_p([0, d'T], W_p^1)$. Then for any $\alpha \in \mathcal{N}$

$$\bar{\mathcal{R}}(B_\alpha h) = c_{\alpha 0}\bar{\mathcal{R}}h + \delta\bar{\mathcal{Q}}_{\alpha 0}h,$$

$$\bar{\mathcal{Q}}_\alpha u = \bar{\mathcal{R}}(c_{\alpha 0}\bar{L}u - c_{\alpha r}h_r) + \delta\bar{\mathcal{Q}}_{\alpha 0}\bar{L}u - \delta\bar{\mathcal{Q}}_{\alpha r}h_r + B_\alpha u.$$

The proof of this lemma is an obvious modification of that of Lemma 4.2.

Next, let us use the notation

$$w_\beta = \bar{L}_\beta w + \bar{f}_\beta.$$

Since $w \in C_w([0,d'T],W^l_p)$, the functions w_β are well defined for $2|\beta| \le l$. Under the same condition the coefficients of L_β and f_β have the first derivative in time, and these derivatives are under control. Furthermore, $dw_\beta = h_r\, dA_r$, where, as long as $2|\beta|+3 \le l$, the functions

$$h_0 = (\bar{L}\bar{L}_\beta - \bar{L}_{\beta 0})w + \bar{L}\bar{f}_\beta - \bar{f}_{\beta 0},$$

$$h_r = \bar{L}(\bar{L}_r w + \bar{f}_r), \quad r = 1,\dots,d_1,$$

are bounded W^1_p-valued functions on $[0,d'T]$.

Then in the same way as Lemma 4.2 was used to obtain Lemma 4.3, Proposition 4.4, and Lemma 4.5, we use Lemma 5.4 to get their counterparts, formulated as follows.

LEMMA 5.5. *Let $\alpha,\beta \in \mathcal{N}$. If $2|\beta|+3 \le l$, then*

$$\bar{\mathcal{Q}}_\alpha w_\beta = c_{\alpha r}\bar{\mathcal{R}} w_{\beta r} + \delta \bar{\mathcal{Q}}_{\alpha r} w_{\beta r} + B_\alpha w_\beta.$$

PROPOSITION 5.6. *Let $\kappa \ge 0$ be an integer and $l \ge 2\kappa+3$. Then*

$$w = v + \sum_{i=1}^{\kappa}\delta^i \sum_{|\alpha|=i} B_\alpha w_\alpha + \sum_{i=1}^{\kappa}\delta^i \sum_{|\alpha|=i+1} c_\alpha \bar{\mathcal{R}} w_\alpha + \delta^{\kappa+1}\sum_{|\alpha|=\kappa+1}\bar{\mathcal{Q}}_\alpha w_\alpha. \tag{5.10}$$

LEMMA 5.7. *If $\kappa \ge 0$ is an integer, $\alpha,\beta \in \mathcal{N}$, and $2(|\beta|+\kappa)+1 \le l$, then*

$$\bar{\mathcal{R}}(B_\alpha w_\beta) = \sum_{i=0}^{\kappa}\delta^i \sum_{|\gamma|=i} c_{\alpha 0\gamma}\bar{\mathcal{R}} w_{\beta\gamma}$$

$$+\sum_{i=1}^{\kappa}\delta^i \sum_{|\gamma|=i-1} B_{\alpha 0\gamma} w_{\beta\gamma} + \delta^{\kappa+1}\sum_{|\gamma|=\kappa}\bar{\mathcal{Q}}_{\alpha 0\gamma} w_{\beta\gamma},$$

where for any multinumbers μ,ν

$$\sum_{|\gamma|=0} c_{\nu\gamma}\bar{\mathcal{R}} w_{\mu\gamma} := c_\nu \bar{\mathcal{R}} w_\mu, \quad \sum_{|\gamma|=0} B_{\nu\gamma} w_{\mu\gamma} := B_\nu w_\mu,$$

$$\sum_{|\gamma|=0}\bar{\mathcal{Q}}_{\nu\gamma} w_{\mu\gamma} := \bar{\mathcal{Q}}_\nu w_\mu.$$

In order to iterate (5.10) we introduce the following class of indices. We say that

$$\beta = \gamma^\nu := \gamma_1^{\nu_1}\gamma_2^{\nu_2}\dots\gamma_j^{\nu_j} \tag{5.11}$$

is a *graded multinumber* of length $|\beta| := j+\nu_1+\nu_2+\dots+\nu_j$ if $\gamma_i \in \{1,2,\dots,d_1\}$; $\nu_i \ge 0$ is any integer for $i=1,2,\dots,j$, where $j \ge 1$ is any integer. If $\nu_i = 0$ for some i, then we also write γ_i in place of γ_i^0 in (5.11). Let $\mathcal{K}$ denote the set of all graded multinumbers. For each $\beta = \gamma^\nu = \gamma_1^{\nu_1}\gamma_2^{\nu_2}\dots\gamma_j^{\nu_j} \in \mathcal{K}$ of length $|\beta| \le 1+l/2$ we introduce the following operators and functions:

$$L_\beta = L_{\gamma^\nu} := (-1)^{|\beta|-1} L^{(\nu_1)}_{\gamma_1} L^{(\nu_2)}_{\gamma_2}\cdot\dots\cdot L^{(\nu_j)}_{\gamma_j},$$

$$f_\beta = f_{\gamma^\nu} := (-1)^{|\beta|-1} L^{(\nu_1)}_{\gamma_1}\cdot\dots\cdot L^{(\nu_j-1)}_{\gamma_{j-1}} f^{(\nu_j)}_{\gamma_j}, \tag{5.12}$$

where $f_r^{(s)} := D_t^s f_r$ and $L_r^{(s)}$ denotes the operator which we obtain from L_r by applying the derivation D_t^s to each of its coefficients. By definition, $f_r^{(0)} = f_r$ and $L_r^{(0)} = L_r$. It is easy to see that for $\beta \in \mathcal{N}$, when $\beta = \beta^0 \in \mathcal{K}$, definitions (5.12) are consistent with (5.9).

LEMMA 5.8. *The following statements hold:*

(i) *Let $\beta, \gamma \in \mathcal{K}$ be such that $|\beta| + |\gamma| \le 1 + l/2$. Then*

$$L_\beta L_\gamma = -L_{\beta\gamma}, \quad L_\beta f_\gamma = -f_{\beta\gamma}.$$

(ii) *Let $\alpha \in \mathcal{N}$ be such that $\rho := |\alpha| \le 1 + l/2$. Then there exist constants $c(\gamma) = c(\alpha, \gamma) \in \{0, \pm 1, \pm 2, \dots\}$ defined for all $\gamma \in \mathcal{K}$ with $|\gamma| = \rho$, such that*

$$L_\alpha = \sum_{\gamma \in \mathcal{K}, |\gamma| = \rho} c(\gamma) L_\gamma, \quad f_\alpha = \sum_{\gamma \in \mathcal{K}, |\gamma| = \rho} c(\gamma) f_\gamma. \tag{5.13}$$

Proof. Part (i) follows immediately from the definition (5.12) of L_β, f_β. Part (ii) clearly holds for $\alpha = 0$ and $\alpha = r \in \{1, \dots, d_1\}$. Assume that (5.13) holds for some $\alpha \in \mathcal{N}$, $|\alpha| < 1 + l/2$. Then

$$L_{\alpha r} = -L_\alpha L_r = -\sum_{\beta \in \mathcal{K}, |\beta| = |\alpha|} c(\beta) L_\beta L_r = \sum_{\beta \in \mathcal{K}, |\beta| = |\alpha|} c(\beta) L_{\beta r},$$

$$f_{\alpha r} = -L_\alpha f_r = -\sum_{\beta \in \mathcal{K}, |\beta| = |\alpha|} c(\beta) L_\beta f_r = \sum_{\beta \in \mathcal{K}, |\beta| = |\alpha|} c(\beta) f_{\beta r}$$

for $r \in \{1, 2, \dots, d_1\}$, and

$$L_{\alpha 0} = L L_\alpha - \dot{L}_\alpha = \sum_{r=1}^{d_1} \sum_{\beta \in \mathcal{M}, |\beta| = \rho} c(\beta) L_r L_\beta - \sum_{\gamma^\nu \in \mathcal{M}, |\gamma^\nu| = \rho} c(\gamma^\nu) \dot{L}_{\gamma^\nu},$$

$$f_{\alpha 0} = L f_\alpha - \dot{f}_\alpha = \sum_{r=1}^{d_1} \sum_{\beta \in \mathcal{M}, |\beta| = \rho} c(\beta) L_r f_\beta - \sum_{\gamma^\nu \in \mathcal{M}, |\gamma^\nu| = \rho} c(\gamma^\nu) \dot{f}_{\gamma^\nu}.$$

Hence by using assertion (i) and noticing that

$$\dot{L}_{\gamma^\nu} = \sum_{|\mu|=1} L_{\gamma^{\nu+\mu}}, \quad \dot{f}_{\gamma^\nu} = \sum_{|\mu|=1} f_{\gamma^{\nu+\mu}},$$

we get (5.13) for αr with $r = 0, 1, \dots, d_1$. Thus the induction on the length of α completes the proof. $\square$

For $\beta \in \mathcal{K}$ we write $\bar{L}_\beta$ and $\bar{f}_\beta$, when the time change $A_0(t)$ is done in the coefficients of L_β and in f_β. We set $w_\gamma := \bar{L}_\gamma w + \bar{f}_\gamma$ for $\gamma \in \mathcal{K}$, $|\gamma| \le 1 + l/2$. Notice that Lemma 5.8 has an obvious translation in terms of these functions. Namely, by Lemma 5.8(ii) for every $\alpha \in \mathcal{N}$ such that $\rho := |\alpha| \le 1 + l/2$ there exist constants $c(\gamma) = c(\alpha, \gamma) \in \{0, \pm 1, \pm 2, \dots\}$ defined for all $\gamma \in \mathcal{K}$ with $|\gamma| = \rho$ such that

$$w_\alpha = \sum_{\gamma \in \mathcal{K}, |\gamma| = \rho} c(\gamma) w_\gamma.$$

For every integer $i \ge 1$ we introduce finite sequences $\sigma := (\beta_1, \beta_2, \dots, \beta_i)$ of graded multinumbers $\beta_i \in \mathcal{K}$, and we set $|\sigma| := |\beta_1| + |\beta_2| + \dots + |\beta_i|$. We also introduce the empty sequence e of length $|e| := 0$. The set of all these sequences is denoted by $\mathcal{I}$.

For $\sigma = (\beta_1, \beta_2, \dots, \beta_i)$ with $|\sigma| \le 1 + l/2$ we define

$$S_\sigma := \mathcal{R}L_{\beta_1}\mathcal{R}L_{\beta_2}\cdot\dots\cdot\mathcal{R}L_{\beta_i}, \quad \bar{S}_\sigma := \bar{\mathcal{R}}\bar{L}_{\beta_1}\bar{\mathcal{R}}\bar{L}_{\beta_2}\cdot\dots\cdot\bar{\mathcal{R}}\bar{L}_{\beta_i},$$

and for $|\sigma| = 0$ we set

$$S_e := \mathcal{R}, \quad \bar{S}_e := \bar{\mathcal{R}}.$$

Notice that for any $g \in L_p([0,T], W_p^{2|\sigma|})$

$$\bar{S}_\sigma \bar{g}(t,\cdot) = (S_\sigma g)(A_0(t),\cdot), \tag{5.14}$$

where $\bar{g}(t,\cdot) := g(A_0(t),\cdot)$. This follows from (5.8) by induction on $|\sigma|$. In order to formulate the counterparts of Lemma 4.7 and Proposition 4.8 we use the following sets:

$$A(i) = \{(\sigma,\beta) : \sigma \in \mathcal{I}, \beta \in \mathcal{K}, |\sigma| + |\beta| \le i\},$$

$$B(i,j) = \{(\alpha,\beta) : \alpha \in \mathcal{N}, \beta \in \mathcal{K}, |\alpha| = i, |\beta| \le j\},$$

$$B^*(i,j) = \bigcup_{i_1=1}^{i} B(i_1,j).$$

Remember that if g_ν is a collection of functions indexed by a parameter ν taking values in a set A, then $\sum^*_{\nu\in A} g_\nu$ means any linear combination of g_ν with coefficient independent of the argument of g_ν and of δ.

LEMMA 5.9. *Let $\sigma \in \mathcal{I}$, $\kappa, \mu \ge 0$ be integers, and $\alpha \in \mathcal{N}, \beta \in \mathcal{K}$. Assume that*

$$2(|\sigma| + |\beta| + \kappa) + \mu + 2 \le l.$$

Then

$$\bar{S}_\sigma(B_\alpha w_\beta) = \sum_{i=0}^{\kappa} \delta^i \sum\nolimits^*_{A(|\sigma|+|\beta|+i)} \bar{S}_{\sigma_1} w_{\beta_1}$$

$$+ \sum_{i=1}^{\kappa} \delta^i \sum\nolimits^*_{B(|\alpha|+i, |\sigma|+|\beta|+i-1)} B_{\alpha_1} w_{\beta_1} + O_\mu(\delta^{\kappa+1}).$$

Proof. We can derive this lemma from Lemma 5.7 in the same way that we proved Lemma 4.7. We need only use the sets $\mathcal{K}$ and $\mathcal{I}$ in place of $\mathcal{M}$ and $\mathcal{J}$, and the operators $\bar{\mathcal{R}}$, $\bar{L}_\nu$, $\bar{S}_\sigma$ for $\nu \in \mathcal{K}$, $\sigma \in \mathcal{I}$, in place of $\mathcal{R}$, L_ν, S_σ for $\nu \in \mathcal{M}$, $\sigma \in \mathcal{J}$, respectively. ∎

PROPOSITION 5.10. *Let $k, m \ge 0$ be integers, and*

$$4 + m + 4k \le l.$$

Then for any $j = 0, 1, \dots, k$ we have

$$w = v + \sum_{i=1}^{j} \delta^i \sum\nolimits^*_{A(2i)} \bar{S}_\sigma v_\beta + \sum_{i=j+1}^{k} \delta^i \sum\nolimits^*_{A(i+j+1)} \bar{S}_{\sigma_1} w_{\beta_1}$$

$$+ \sum_{i=1}^{k} \delta^i \sum\nolimits^*_{B^*(i,i+j)} B_{\alpha_1} w_{\beta_1} + O_m(\delta^{k+1}),$$

where $v_\beta := \bar{L}_\beta v + \bar{f}_\beta$.

Proof. The proof of this proposition is a straightforward translation of the proof of the corresponding proposition, Proposition 4.8, in the time independent case. To make this translation we use the sets $\mathcal{K}$ and $\mathcal{I}$ in place of $\mathcal{M}$ and $\mathcal{J}$, and the operators $\bar{\mathcal{R}}$, $\bar{L}_\nu$, $\bar{S}_\sigma$ for $\nu \in \mathcal{K}$, $\sigma \in \mathcal{I}$, in place of $\mathcal{R}$, L_ν, S_σ for $\nu \in \mathcal{M}$, $\sigma \in \mathcal{J}$, respectively. ∎

Now we can finish the proof of Theorem 5.1 as follows. Taking $j = k$ in Proposition 5.10, we get

$$w = v + \sum_{i=1}^{j} \delta^i \sum_{A(2i)} c(\sigma, \beta) \bar{S}_\sigma v_\beta + \sum_{B^*(k,2k)} c(\alpha, \beta, \delta) B_{\alpha_1} w_{\beta_1} + r_\delta, \tag{5.15}$$

where $c(\sigma, \beta)$, $c(\alpha, \beta, \delta)$ are certain constants, r_δ is a function in $C_w([0, d'T], W_p^m)$ for each δ, and

$$\sup_{t \in [0,d'T]} \sup_{n, \delta = T/n} \delta^{-(k+1)} \|r_\delta(t, \cdot)\|_{m,p} \le N.$$

Observe that, in contrast with (4.21), the functions v and $\bar{S}_\sigma v_\beta$ in (5.15) may depend on δ. To proceed further, define $R_n^{(k)}(t, x) := r_\delta(d't, x)$, and

$$u^{(i)} := \sum_{A(2i)} c(\sigma, \beta) S_\sigma u_\beta, \quad i = 1, 2, \dots, k,$$

where $u_\beta := L_\beta u + f_\beta$. Then by virtue of equality (5.14) and the fact that $v(t) = u(A_0(t))$, from (5.15) we get

$$w(t, \cdot) = u(A_0(t), \cdot) + \sum_{i=1}^{j} \delta^i u^{(i)}(A_0(t), \cdot)$$

$$+ \sum_{B^*(k,2k)} c(\alpha, \beta, \delta) B_{\alpha_1}(t) w_{\beta_1}(t, \cdot) + R_n^{(k)} \left(\frac{t}{d'}, \cdot \right).$$

Substituting $d't$ in place of t, we get the required representation (5.1) by taking into account that

$$w(d't) = u_n(t), \quad A_0(d't) = t, \quad B_{\alpha_1}(d't) = 0 \quad \forall t \in T_n.$$

Remark 5.1. Let $1 \le j \le d_1$. Then Theorems 5.1, 5.2, and 5.3 also hold when the operator $\mathbb{S}_\delta^{(r)}(t_{i+1}^n)$ is replaced with $\mathbb{S}_\delta^{(r)}(t_i^n)$ for every $r = 1, 2, \dots, j$ in the definition (5.2) of the splitting-up approximation u_n. To see this we need only repeat the proof of the previous theorem with A_j in place of A_0 in (5.6) and with A_0 and A_j interchanged in (5.5).

Acknowledgment. We express our sincere gratitude to the referees for useful criticism which helped improve the presentation.

REFERENCES

[1] T. Beale and A. Majda, *Rates of convergence of viscous splitting of the Navier–Stokes equations*, Math. Comp., 37 (1981), pp. 243–259.

[2] K. A. Bagrinovsky and S. K. Godunov, *Difference schemes for multidimensional problems*, Dokl. Acad. Nauk USSR, 115 (1957), pp. 431–433 (in Russian).

[3] A. J. Chorin, *Numerical solution of the Navier–Stokes equations*, Math. Comp., 22 (1968), pp. 745–762.

[4] J. Douglas, Jr., and H. H. Rachford, *On the numerical solution of the heat conduction problems in two and three space variables*, Trans. Amer. Math. Soc., 82 (1956), pp. 421–439.

[5] M. S. Espedal and K. H. Karlsen, *Numerical solution of reservoir flow models based on large time step operator splitting algorithms*, in Filtration in Porous Media and Industrial Applications (Cetraro, Italy 1998), M. S. Espedal and A. Mikelić, eds., Lecture Notes in Math. 1734, Springer Verlag, Berlin, 2000, pp. 9–77.

[6] D. Gottlieb, *Strang-type difference schemes for multidimensional problems*, SIAM J. Numer. Anal., 9 (1972), pp. 650–661.

[7] I. Gyöngy and N. Krylov, *On splitting up method and stochastic partial differential equations*, Ann. Probab., 31 (2003), pp. 564–591.

[8] I. Gyöngy and N. Krylov, *On the rate of convergence of splitting-up approximations for SPDEs*, in Progr. Probab. 56, 2003, Birkhäuser-Verlag, Basel, pp. 301–321.

[9] R. H. Hardin and F. D. Tappert, *Application of the split-step Fourier method to the numerical solution of nonlinear and variable coefficient wave equations*, SIAM Rev., 15 (1973), p. 423 (abstract).

[10] N. Krylov and B. Rozovskii, *On the characteristics of degenerate second order parabolic Itô equations*, Trudy seminara imeni Petrovskogo, 8 (1982), pp. 153–168 (in Russian); English translation in J. Soviet Math., 32 (1986), pp. 336–348.

[11] G. I. Marchuk, *Methods of Numerical Mathematics*, Springer–Verlag, New York, Heidelberg, Berlin, 1975.

[12] G. I. Marchuk, *Splitting and alternating direction methods*, in Handbook of Numerical Analysis, P. G. Ciarlet, and J. L. Lions, eds., North-Holland, Amsterdam, 1990, Volume 1, pp. 197–462.

[13] R. I. McLachlan and G. R. W. Quispel, *Splitting methods*, Acta Numer., 11 (2002), pp. 341–434.

[14] A. Murua and J. M. Sanz-Serna, *Order conditions for numerical integrators obtained by composing simpler integrators*, Phil. Trans. Roy. Soc. London A, 357 (1999), pp. 1079–1100.

[15] F. Neri, *Lie Algebras and Canonical Integration*, preprint, Department of Physics, University of Maryland, College Park, MD, 1988.

[16] D. W. Peaceman and H. H. Rachord, Jr., *The numerical solution of parabolic and elliptic differential equations*, J. Soc. Indust. Appl. Math., 3 (1955), pp. 28–41.

[17] W. G. Strang, *Accurate partial difference methods* I. *Linear Cauchy problems*, Arch. Mech., 12 (1963), pp. 392–402.

[18] W. G. Strang, *Accurate partial difference methods* II. *Nonlinear problems*, Numer Math., 13 (1964), pp. 37–46.

[19] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 506–517.

[20] Q. Sheng, *Solving linear partial differential equations by exponential splitting*, IMA J. Numer. Anal., 9 (1989), pp. 199–212.

[21] M. Suzuki, *General theory of higher order decomposition of exponential operators and symplectic integrators*, Phys. Lett. A, 165 (1992), pp. 387–395.

[22] M. Suzuki, *Convergence of exponential product formulas for unbounded operators*, Rev. Math. Phys., 8 (1996), pp. 487–502.

[23] F. D. Tappert, *Numerical solutions of the Korteweg–de Vries equation and its generalizations by the split step Fourier method*, in Nonlinear Wave Motion, A. C. Newell, ed., Lecture Notes in Appl. Math. 15, AMS, Providence, RI, 1977, pp. 215–216.

[24] R. Temam, *Sur la stabilité et la convergence de la méthode de pas fractionnaires*, Ann. Mat. Pura Appl., 79 (1968), pp. 191–379.

[25] H. F. Trotter, *On the products of semigroups of operators*, Proc. Amer. Math. Soc., 10 (1959), pp. 545–551.

[26] Z. Tsuboi and M. Suzuki, *Determining equations for higher-order decompositions of exponential operators*, Int. J. Mod. Phys. B, 25 (1995), pp. 3241–3268.

[27] N. Victoir, *On the Nonexistence of Nonnegative Splitting Method of Degree Greater or Equal to* 3, preprint, Mathematical Institute, Oxford University, Oxford, UK, 2002.

[28] H. Yosida, *Construction of higher order symplectic integrators*, Phys. Lett., 165 (1990), pp. 387–395.

[29] N. N. Yanenko, *The Method of Fractional Steps. The Solution of Problems of Mathematical Physics in Several Variables*, Springer-Verlag, New York, 1971.

SIAM J. Math. Anal.
Vol. 37, No. 4, pp. 1098–1116

PARTIAL REGULARITY FOR A SELECTIVE SMOOTHING FUNCTIONAL FOR IMAGE RESTORATION IN BV SPACE*

YUNMEI CHEN†, M. RAO†, Y. TONEGAWA‡, AND T. WUNDERLI§

Abstract. In this paper we study the partial regularity of a functional on BV space proposed by Chambolle and Lions [*Numer. Math.*, 76 (1997), pp. 167–188] for the purposes of image restoration. The functional is designed to smooth corrupted images using isotropic diffusion via the Laplacian where the gradients of the image are below a certain threshold ϵ and retain edges where gradients are above the threshold using the total variation. Here we prove that if the solution $u \in BV$ of the model minimization problem, defined on an open set Ω, is such that the Lebesgue measure of the set where the gradient of u is below the threshold ϵ is positive, then there exists a nonempty open region E for which $u \in C^{1,\alpha}$ on E and $|\nabla u| < \epsilon$, and $|\nabla u| \geq \epsilon$ on $\Omega \backslash E$ almost everywhere. Thus we indeed have smoothing where $|\nabla u| < \epsilon$.

Key words. bounded variation, selective smooothing, image processing, image restoration, noise removal, partial regularity

AMS subject classifications. 49J40, 35K65

DOI. 10.1137/050623887

1. Introduction.

In the last decade PDE and variational method based diffusion models have grown significantly to tackle the problems of image restoration, reconstruction, and inpainting. The challenging aspect of these problems is to design methods which can filter selectively the noise without losing significant features.

Total variation (TV) based regularization, as first proposed by Rudin, Osher, and Fatemi [17], has proved to be an invaluable tool for preserving edges while reconstructing an image. This method has been studied extensively in [1, 7, 4, 20, 2, 5, 19, 16] and a sequence of papers in the book of [2]. The definition of the total variation seminorm for $u \in L^1(\Omega)$, given by

$$TV(u) = \sup \left\{ \int_\Omega u \,\mathrm{div}(\varphi) dx : \varphi \in C_0^1(\Omega, \mathbf{R}^n), |\varphi| \leq 1 \right\},$$

does not require differentiability or even continuity of u. Thus images with discontinuities are allowed as solutions in the space of $BV(\Omega)$, which is the space of the functions $u \in L^1(\Omega)$ with $TV(u) < \infty$. Moreover, the diffusion resulting from minimizing TV norm is strictly orthogonal to the gradient of the image, and tangential to the edges. This is important for preserving edges while image is smoothed. However, TV-based denoising sometimes causes a *staircasing* effect [6, 7, 8]. The restored image by this regularization can consequently be blocky and even contain false edges.

To overcome this problem and make the filter self-adjustable in order to reap the benefits of isotropic smoothing and TV based regularization, Chambolle and Lions [7]

*Received by the editors February 5, 2005; accepted for publication (in revised form) July 11, 2005; published electronically December 30, 2005.
http://www.siam.org/journals/sima/37-4/62388.html
†Department of Mathematics, University of Florida, Gainesville, FL 32611 (yun@math.ufl.edu, rao@math.ufl.edu).
‡Department of Mathematics, Hokkaido University, Sapparo 060-0810, Japan (tonegawa@math.sci.hokudai.ac.jp).
§Department of Mathematics and Statistics, The American University of Sharjah, Sharjah, United Arab Emirates (twunderli@ausharjah.edu).

proposed minimizing the following energy functional for image restoration:

$$\frac{1}{2a}\int_{|\nabla u|<a}|\nabla u|^2dx+\int_{|\nabla u|\geq a}\left(|\nabla u|-\frac{a}{2}\right)+\frac{1}{2}\int_{\Omega}(u-I)^2dx,$$

where I is an observed noisy image, and we want to recover an image u from I, which is related to I by

$$I=u+noise.$$

Using the above functional we then expect to have isotropic diffusion where the image is more uniform ($|\nabla u|<a$) and feature preservation via TV-based diffusion where the boundaries of features are present (the locations where the image gradients most likely have high magnitude: $|\nabla u|\geq a$). It has been shown numerically in [7] that this model is successful in restoring images where homogeneous regions are separated by distinct edges. The purpose of this paper is to prove this mathematically. Our partial regularity results for $a=1$ (without loss of generality) below indicates that the restored image through this model is smooth on the region with smaller gradients. The edges appear at the points where the gradient is larger.

More precisely, consider the problem

$$\min_{u\in BV(\Omega)\cap L^2(\Omega)}\left\{\int_{\Omega}\varphi(Du)+\frac{1}{2}\int_{\Omega}(u-I)^2dx\right\}, \tag{1.1}$$

where φ is the C^1 convex function defined on $\mathbf{R}^n$

$$\varphi(p)=\begin{cases}\frac{1}{2}|p|^2 & \text{if } |p|<1,\\ |p|-\frac{1}{2} & \text{if } |p|\geq 1,\end{cases}$$

$\Omega\subset\mathbf{R}^n$ is a bounded domain with Lipschitz boundary, and $I\in L^\infty(\Omega)\cap BV(\Omega)$ is given.

For $u\in BV(\Omega)$ the gradient of u is a measure Du; it can be decomposed into its absolutely continuous and singular parts with respect to Lebesgue measure, that is,

$$Du=\nabla u\,dx+D^su.$$

See [10] for a complete discussion. Then we define ([12] or [9])

$$J(u)\equiv\int_{\Omega}\varphi(Du)\equiv\int_{\Omega}\varphi(\nabla u)dx+\int_{\Omega}|D^su|$$

with

$$\int_{\Omega}|D^su|\equiv\int_{\Omega}d|D^su|=|D^su|(\Omega).$$

It is important to note ([21] or [12]) that the functional J can also be written as

$$J(u)=\sup_{\phi\in C_0^1(\Omega,\mathbf{R}^n)}\left\{-\int_{\Omega}\left(\frac{1}{2}|\phi|^2+u\,\mathrm{div}(\phi)\right)dx:|\phi(x)|\leq 1\ \forall x\in\Omega\right\}.$$

Using this, we see that the functional J is lower semicontinuous with respect to convergence in $L^1(\Omega)$. Then by a standard argument we can show that there is a unique solution $u \in BV(\Omega) \cap L^2(\Omega)$ to (1.1). Now we are interested as to whether this solution $u \in BV$ is smooth on the region where $|\nabla u| < 1$. If so, it shows that the denoising governed by (1.1) smoothes out lower gradients while preserving the boundaries of features, which are the discontinuities in an image.

We now state the two main partial regularity results of this paper.

THEOREM 1.1. *If u is the solution to (1.1), then for any given $0 < \mu < 1$ there exist positive constants ϵ_0 and κ_0 depending only on n and μ such that if*

$$\frac{1}{|B_r|}\int_{B_r(a)} |Du - l| \le \epsilon_0$$

holds for some ball $B_r(a) \subset\subset \Omega$ and for some $l \in \mathbf{R}^n$, with

$$rC\left(1 + \|I\|_{L^\infty(\Omega)}\right) < \kappa_0 \quad \text{and} \quad |l| < 1 - 2\mu,$$

for some constant C depending only on n and Ω, then

$$|D^s u|(B_{r/2}(a)) = 0 \quad \text{and} \quad |\nabla u| < 1 - \mu \quad \text{on} \quad B_{r/2}(a)$$

and u solves

$$-\Delta u = I - u \quad \text{on} \quad B_{r/2}(a).$$

Hence $u \in C^{1,\alpha}(B_{r/2}(a))$ for any $\alpha < 1$.

THEOREM 1.2. *Let u be as in Theorem (1.1). If $\mathcal{L}^n(\{|\nabla u| < 1\}) > 0$, then there exists a nonempty open region E on which u is $C^{1,\alpha}$, $|\nabla u| < 1$, and u solves*

$$-\Delta u = I - u \quad \text{on} \quad E.$$

In addition we have $|\nabla u| \ge 1$ a.e. on $\Omega \setminus E$.

It is actually straightforward to show that Theorem 1.2 is a direct consequence of Theorem 1.1. Thus from Theorem 1.2, we do indeed have smoothing where $|\nabla u| < 1$.

Here we should point out that partial regularity results were obtained in [3] for minimizers in $BV(\Omega)$ of functionals of the form $\int_\Omega (F(x, Du) + G(x, u))$, where $F(x, p)$ is a convex function in p with $c_1|p| \le F(x, p) \le c_2(1 + |p|)$ for all $p \in \mathbf{R}^n$, F is locally Hölder continuous in x, and $G(x, z)$ satisfies Hölder continuity conditions in both x and z. In our case, $G(x, z) = 1/2(z - I(x))^2$ with only the stated assumption on I, and therefore their results cannot directly be applied in our case. Moreover, our approach is quite different from theirs and can be applied to more general cases.

The partial regularity results for the flow associated with the minimization problem (1) is also discussed in [15] for more general φ. However, these hold only for $\Omega \subset \mathbf{R}^n$ for $n = 1$ and $n = 2$. We also apply some different techniques to get our results.

2. Proof of Theorems 1.1 and 1.2. First we will show that the solution u to (1.1) is in $L^\infty(\Omega)$. To prove this we could consider the time evolution problem corresponding to (1.1), prove an L^∞ bound for the time-dependent solution $u(x, t)$, and then consider the time-asymptotic limit u, which is the solution to (1.1). We would then conclude that $u \in L^\infty(\Omega)$. The following, however, provides a proof of this without having to consider the time evolution of (1.1).

LEMMA 2.1. *If u is the solution to (1.1), then $u \in L^\infty(\Omega)$. In fact, we have $\|u\|_{L^\infty(\Omega)} \le \|I\|_{L^\infty(\Omega)}$.*

Proof. Let φ_ϵ be defined on $\mathbf{R}^n$ by

$$\varphi_\epsilon(p) = \begin{cases} \dfrac{1}{2}|p|^2 & \text{if } |p| < 1, \\ \dfrac{1}{1+\epsilon}|p|^{1+\epsilon} + \left(\dfrac{1}{2} - \dfrac{1}{1+\epsilon}\right) & \text{if } |p| \geq 1 \end{cases}$$

for $\epsilon > 0$, and consider the following minimization problem:

$$\min_{u \in W^{1,1+\epsilon}(\Omega) \cap L^2(\Omega)} \left\{ \int_\Omega \varphi_\epsilon(\nabla u) + \frac{1}{2} \int_\Omega (u - I)^2 dx \right\}.$$

By standard methods, there is a unique solution u_ϵ to this problem. We follow a standard truncation argument where we fix ϵ and $t \geq 0$ and let $v = \min(u_\epsilon, t)$. Noting that $v \in W^{1,1+\epsilon}(\Omega) \cap L^2(\Omega)$ with

$$\nabla v = \begin{cases} \nabla u_\epsilon & \text{if } u_\epsilon < t, \\ 0 & \text{if } u_\epsilon \geq t, \end{cases}$$

we have

$$\int_\Omega \varphi_\epsilon(\nabla u_\epsilon) + \frac{1}{2} \int_\Omega (u_\epsilon - I)^2\, dx \leq \int_\Omega \varphi_\epsilon(\nabla v) + \frac{1}{2} \int_\Omega (v - I)^2 dx, \tag{2.1}$$

and thus after subtracting

$$\int_{\{u_\epsilon \geq t\}} \varphi_\epsilon(\nabla u_\epsilon)\, dx + \int_{\{u_\epsilon \geq t\}} (u_\epsilon - I)^2 dx \leq \int_{\{u_\epsilon \geq t\}} (t - I)^2 dx.$$

Hence

$$\int_{\{u_\epsilon \geq t\}} (u_\epsilon - I)^2 dx \leq \int_{\{u_\epsilon \geq t\}} (t - I)^2 dx.$$

But setting $t = \|I\|_{L^\infty(\Omega)}$ we see that if ess sup $u_\epsilon > t$, then

$$\int_{\{u_\epsilon \geq t\}} (t - I)^2 dx < \int_{\{u_\epsilon \geq t\}} (u_\epsilon - I)^2 dx,$$

which contradicts the above, hence ess sup $u_\epsilon \leq \|I\|_{L^\infty(\Omega)}$. Applying a similar argument to $v = \max(u_\epsilon, -t)$ for $t = \|I\|_{L^\infty(\Omega)}$ we get ess inf $u_\epsilon \geq -\|I\|_{L^\infty(\Omega)}$ and thus $\|u_\epsilon\|_{L^\infty(\Omega)} \leq \|I\|_{L^\infty(\Omega)}$. Furthermore, letting $v = 0$ in (2.1) we see that u_ϵ is bounded in $W^{1,1+\epsilon}(\Omega) \cap L^2(\Omega) \subset BV(\Omega) \cap L^2(\Omega)$ independent of ϵ. Thus there is a $\tilde{u} \in BV(\Omega) \cap L^2(\Omega)$ and a subsequence of $\{u_\epsilon\}$, still denoted by $\{u_\epsilon\}$, such that $u_\epsilon \to \tilde{u}$ strongly in $L^1(\Omega)$, $u_\epsilon \rightharpoonup \tilde{u}$ weakly in $L^2(\Omega)$, and $u_\epsilon \to \tilde{u}$ almost everywhere (a.e.) in Ω. Letting $\epsilon \to 0$ in (2.1), noting that $\varphi(p) \leq \varphi_\epsilon(p)$ for all p, $\int_\Omega \varphi_\epsilon(\nabla v) \to \int_\Omega \varphi(\nabla v)$, lower semicontinuity of the functional $\int_\Omega \varphi(\nabla u)$ defined on $BV(\Omega)$, and weak lower semicontinuity of the second term on the left-hand side, we get

$$\int_\Omega \varphi(\nabla \tilde{u}) + \frac{1}{2} \int_\Omega (\tilde{u} - I)^2 dx \leq \int_\Omega \varphi(\nabla v) + \frac{1}{2} \int_\Omega (v - I)^2 dx$$

for all $v \in W^{1,1+\epsilon}(\Omega) \cap L^2(\Omega)$. We now note [12] that for any $v \in BV(\Omega) \cap L^2(\Omega)$ there exists a sequence v_n in $C^\infty(\overline{\Omega})$ such that

$$\int_\Omega \varphi(\nabla v_n) dx \to \int_\Omega \varphi(\nabla v)$$

and $v_n \to v$ in $L^1(\Omega)$, and since $v \in L^2(\Omega)$ from the construction of v_n [12] we can also take $v_n \to v$ in $L^2(\Omega)$. Therefore we see that the above holds for all $v \in BV(\Omega) \cap L^2(\Omega)$ as well. Hence $\tilde{u}$ solves (1.1). By uniqueness, $\tilde{u} = u$. By the uniform L^∞ bound for u_ϵ and the convergence of u_ϵ to u a.e. in Ω we have $u \in L^\infty(\Omega)$ with $\|u\|_{L^\infty(\Omega)} \le \|I\|_{L^\infty(\Omega)}$ ▯

Throughout the rest of the paper, we fix $\mu > 0$ and unless otherwise stated, all constants depend at most on n, μ, u, Ω, φ, and possibly I.

We begin with a local lower bound estimate for any BV function u and C^1 function h with gradient strictly less than 1.

LEMMA 2.2. *Let $u \in BV(B_r(a))$ for $B_r(a) \subset\subset \Omega$ and $h \in C^1(\overline{B}_r(a))$ with*

$$\sup_{B_r(a)} |\nabla h| \le 1 - \mu;$$

then

$$\int_{B_r(a)} \varphi(Du) - \int_{B_r(a)} \varphi(\nabla h) dx \ge \mu \int_{B_r(a)} |D^s u| + \int_{B_r(a)} \nabla(u-h) \cdot \nabla h dx$$
$$+ \int_{B_r(a)} D^s u \cdot \nabla h + \frac{\mu^2}{2} \int_{B_r(a) \cap \{|\nabla u| \ge 1\}} |\nabla u| dx$$
$$+ \frac{1}{2} \int_{B_r(a) \cap \{|\nabla u| < 1\}} |\nabla(u-h)|^2 dx.$$

Proof. Where $|\nabla u| \ge 1$, we have

$$\begin{aligned} &\varphi(\nabla u) - \varphi(\nabla h) - \nabla(u-h) \cdot \nabla h \\ &= |\nabla u| - \frac{1}{2} + \frac{1}{2}|\nabla h|^2 - \nabla u \cdot \nabla h \\ &\ge \frac{1}{2}(2|\nabla u| - 1 - 2|\nabla u||\nabla h| + |\nabla h|^2) \\ &= \frac{1}{2}(2|\nabla u| - 1 - |\nabla h|)(1 - |\nabla h|) \ge \frac{\mu^2}{2}|\nabla u|. \end{aligned}$$

Where $|\nabla u| < 1$, we have

$$\varphi(\nabla u) - \varphi(\nabla h) - \nabla(u-h) \cdot \nabla h = \frac{1}{2}|\nabla(u-h)|^2.$$

We now obtain the lemma by using

$$\int_{B_r(a)} |D^s u| \ge \int_{B_r(a)} D^s u \cdot \nabla h + \int_{B_r(a)} |D^s u|(1 - |\nabla h|),$$

the assumption on h, and the above estimates. ▯

We now fix $B_{2r}(a) \subset\subset \Omega$. Let v be a Lipschitz function defined on $B_{2r}(a)$ and assume there exists an $l \in \mathbf{R}^n$ with $|l| \le 1 - 2\mu$, such that $\sup_{B_{2r}(a)} |\nabla v - l| \le \beta^{2\delta}$ for $\delta > 0$ and $0 < \beta < 1$ to be chosen later. Also let $\overline{v}$ be defined by $\overline{v}(x) = v(x) - l \cdot x$. Let η_ϵ be the usual mollifier on $\mathbf{R}^n$ and denote $\overline{v}_\beta = \eta_{r\beta} * \overline{v}$ and $v_\beta = \eta_{r\beta} * v$. We then have the following estimates from [18]:

$$\sup_{B_r(a)} |\nabla v_\beta - l| = \sup_{B_r(a)} |\nabla \overline{v}_\beta| \le \beta^{2\delta}, \tag{2.2}$$

$$\sup_{B_r(a)} |v_\beta - v| = \sup_{B_r(a)} |\overline{v}_\beta - \overline{v}| \leq r\beta \sup_{B_r(a)} |\nabla \overline{v}_\beta| \leq r\beta^{1+2\delta}, \tag{2.3}$$

$$\begin{aligned} r^\delta \sup_{B_r(a)} |x-y|^{-\delta} |\nabla v_\beta(x) - \nabla v_\beta(y)| \\ \leq c_1 r^\delta \sup_{B_r(a)} |\nabla v - l| \sup_{x' \neq y'} |x'-y'|^{-\delta} |\eta_1((r\beta)^{-1}x') - \eta_1((r\beta)^{-1}y')| \\ \leq c_2 \beta^{2\delta}\beta^{-\delta} = c_2\beta^\delta. \end{aligned} \tag{2.4}$$

Now for any $\tilde{r} \in [\frac{r}{2}, r]$ there exists a unique solution [11] $w \in H^1(B_{\tilde{r}}(a)) \cap C^{1,\delta}(\overline{B}_{\tilde{r}}(a))$ with $\delta \in (0,1)$ for the problem

$$-\Delta w = I - w \quad \text{on } B_{\tilde{r}}(a), \quad w = v_\beta \quad \text{on } \partial B_{\tilde{r}}(a). \tag{2.5}$$

LEMMA 2.3. *For $I \in L^\infty(\Omega)$, the solution w to* (2.5) *satisfies*

$$\|w\|_{L^\infty(B_{\tilde{r}}(a))} \leq \|v_\beta\|_{L^\infty(\partial B_{\tilde{r}}(a))} + \|I\|_{L^\infty(\Omega)}. \tag{2.6}$$

$$\sup_{B_{\tilde{r}}(a)} |\nabla w - l| \leq c_3(\beta^\delta + r(\|I\|_{L^\infty(\Omega)} + \|v_\beta\|_{L^\infty(\partial B_{\tilde{r}}(a))})) \quad \textit{for any } l \in R^n. \tag{2.7}$$

$$\begin{aligned} \sup_{x,y \in B_{\tilde{r}/2}(a)} \frac{|\nabla w(x) - \nabla w(y)|}{|x-y|^{1/2}} \leq c_4 \Bigg(\frac{1}{r^{n+1/2}} \int_{(\partial B_{\tilde{r}}(a))} |v_\beta| \, d\mathcal{H}^{n-1} \\ + r^{1/2}(\|I\|_{L^\infty(\Omega)} + \|v_\beta\|_{L^\infty(\partial B_{\tilde{r}}(a))}) \Bigg). \end{aligned} \tag{2.8}$$

Proof. Estimate (2.6) is from Theorem 8.16 in [11]. To prove (2.7) and (2.8), we decompose w as $w = w_1 + w_2$, such that

$$-\Delta w_1 = I - w \quad \text{on } B_{\tilde{r}}(a), \quad w_1 = 0 \quad \text{on } \partial B_{\tilde{r}}(a) \tag{2.9}$$

and

$$-\Delta w_2 = 0 \quad \text{on } B_{\tilde{r}}(a), \quad w = v_\beta \quad \text{on } \partial B_{\tilde{r}}(a). \tag{2.10}$$

Let $\tilde{w}_2 \equiv w_2 - v_\beta$. Then $\tilde{w}_2$ solves

$$-\Delta \tilde{w}_2 = -\mathrm{div}(\nabla v_\beta - l) \quad \text{on } B_{\tilde{r}}(a), \quad \tilde{w}_2 = 0 \quad \text{on } \partial B_{\tilde{r}}(a), \tag{2.11}$$

for any $l \in R^n$. Representing the solution of (2.9) using Green's function, i.e., $w_1(x) = \int_{B_{\tilde{r}}(a)} \Gamma(x-y)(I-w)(y)dy$, where Γ is the fundamental solution of Laplace's equation, it is not difficult to get

$$\|\nabla w_1\|_{L^\infty(B_{\tilde{r}}(a))} \leq cr\|I - w\|_{L^\infty(B_{\tilde{r}}(a))}, \tag{2.12}$$

where c is independent of r.

Moreover, by the Stobolev imbedding theorem, Theorem 9.9 in [11], and (2.6),

$$\begin{aligned} \|\nabla w_1\|_{C^{0,1/2}(B_{\tilde{r}}(a))} \leq c\|w_1\|_{W^{2,2n}(B_{\tilde{r}}(a))} \leq c\|I - w\|_{L^{2n}(B_{\tilde{r}}(a))} \\ \leq cr^{1/2}\|I - w\|_{L^\infty(B_{\tilde{r}}(a))} \leq cr^{1/2}(\|v_\beta\|_{L^\infty(\partial B_{\tilde{r}}(a))} + \|I\|_{L^\infty(\Omega)}). \end{aligned} \tag{2.13}$$

Next we shall estimate w_2. Multiplying both sides of (2.11) by $\tilde{w}_2$, and integrating over $B_{\tilde{r}}(a)$, carrying out a simple computation, and using (2.2), we have for any $l \in R^n$,

$$\int_{B_{\tilde{r}}(a)} |\nabla w_2 - l|^2 dx \le c \int_{B_{\tilde{r}}(a)} |\nabla v_\beta - l|^2 dx \le c r^n \beta^{4\delta}, \tag{2.14}$$

where $c > 0$ is a constant independent of r.

Furthermore, applying Theorem 8.16 and 8.33 (with a rescaling argument) in [11] to (2.11), we get the following estimates:

$$\|\tilde{w}_2\|_{L^\infty(B_{\tilde{r}})} \le c \|\nabla v_\beta - l\|_{L^\infty(B_{\tilde{r}})} \tag{2.15}$$

and

$$r^\delta [D\tilde{w}_2]_{C^{0,\delta}(B_{\tilde{r}})} \le c(\|\tilde{w}_2\|_{L^\infty(B_{\tilde{r}})} + \|\nabla v_\beta - l\|_{L^\infty(B_{\tilde{r}})} + r^\delta [Dv_\beta]_{C^{0,\delta}(B_{\tilde{r}})}), \tag{2.16}$$

where $c > 0$ is a constant independent of r. Inserting (2.15) into (2.16), and using (2.2) and (2.4), it yields

$$\begin{aligned} r^\delta [Dw_2]_{C^{0,\delta}(B_{\tilde{r}})} &\le (r^\delta [D\tilde{w}_2]_{C^{0,\delta}(B_{\tilde{r}})} + r^\delta [Dv_\beta]_{C^{0,\delta}(B_{\tilde{r}})}) \\ &\le c(\|\nabla v_\beta - l\|_{L^\infty(B_{\tilde{r}})} + r^\delta [Dv_\beta]_{C^{0,\delta}(B_{\tilde{r}})}) \le c\beta^\delta. \end{aligned} \tag{2.17}$$

Now we can estimate $\sup_{B_{\tilde{r}}(a)} |\nabla w_2 - l|$. Denoting $|B_{\tilde{r}}(a)|^{-1} \int_{B_{\tilde{r}}(a)} f dx$ by $(f)_{B_{\tilde{r}}(a)}$, and using (2.14) and (2.17), we get

$$\begin{aligned} \sup_{B_{\tilde{r}}(a)} |\nabla w_2 - l| &\le \sup_{B_{\tilde{r}}(a)} \{|\nabla w_2 - (\nabla w_2)_{B_{\tilde{r}}(a)}| + |(\nabla w_2)_{B_{\tilde{r}}(a)} - l|\} \\ &\le r^\delta [Dw_2]_{C^{0,\delta}(B_{\tilde{r}})} + |B_{\tilde{r}}(a)|^{-1/2} \left(\int_{B_{\tilde{r}}(a)} |\nabla w_2 - l|^2 \right)^{1/2} dx \le c\beta^\delta; \end{aligned} \tag{2.18}$$

here we used (2.14) and (2.17) in the last inequality.

We then have from (2.6) and (2.18)

$$\begin{aligned} \sup_{B_{\tilde{r}}(a)} |\nabla w - l| &\le \sup_{B_{\tilde{r}}(a)} |\nabla w_2 - l| + \sup_{B_{\tilde{r}}(a)} |\nabla w_1| \\ &\le c_3(\beta^\delta + r(\|I\|_{L^\infty(B_{\tilde{r}}(a))} + \|v_\beta\|_{L^\infty(\partial B_{\tilde{r}}(a))})). \end{aligned}$$

Hence (2.7) is proved. To prove (2.8) we represent w_2 by the Poisson's formula on the ball $B_{\tilde{r}}(a)$, i.e.,

$$w_2(x) = \frac{\tilde{r}^2 - |x|^2}{n\alpha_n r} \int_{\partial B_{\tilde{r}}(a)} \frac{v_\beta(y)}{|x-y|^n} dS_y, \quad x \in B_{\tilde{r}}(a),$$

where α_n represents the volume of n dimensional unit ball. A direct computation leads to the estimate

$$\sup_{B_{\tilde{r}/2}(a)} |D^2 w_2| \le c r^{-n-1} \int_{\partial B_{\tilde{r}}(a)} |v_\beta(y)| dS_y,$$

where $c > 0$ dependents only on n. Then we have

$$\sup_{x,y\in B_{\tilde{r}/2}(a)} \frac{|\nabla w_2(x) - \nabla w_2(y)|}{|x-y|^{1/2}} \leq \left(\sup_{x,y\in B_{\tilde{r}/2}(a)} |D^2 w_2| \right) |x-y|^{1/2} \tag{2.19}$$

$$\leq \frac{c}{r^{n+1/2}} \int_{\partial B_{\tilde{r}}(a))} |v_\beta| \, d\mathcal{H}^{n-1}.$$

Now (2.8) immediately follows from (2.13) and (2.19). ∎

LEMMA 2.4. *Suppose there is a $v \in C^{0,1}(B_{2r}(a))$ and $l \in \mathbf{R}^n$ with $|l| \leq 1-2\mu$, $\sup_{B_{2r}(a)} |\nabla v - l| \leq \beta^{2\delta}$, and $\sup_{B_{2r}(a)} |v| \leq C_u$, where C_u is a constant depending only on u. Let $v_\beta, \tilde{r}$, and w be as in the previous discussion. Then there exists constants c_5 and c_6 such that if $\beta \leq c_5$ and $r(C_u + \|I\|_{L^\infty(\Omega)}) \leq c_6$, then*

$$\int_{B_{\tilde{r}}(a)} \varphi(Du) - \int_{B_{\tilde{r}}(a)} \varphi(\nabla w)dx \geq \int_{\partial B_{\tilde{r}}(a)} (u - v_\beta) \frac{\partial w}{\partial n} \, d\mathcal{H}^{n-1}$$

$$+ \int_{B_{\tilde{r}}(a)} (u-w)(I-w)dx + \mu \int_{B_{\tilde{r}}(a)} |D^s u| + \frac{\mu^2}{2} \int_{B_{\tilde{r}}(a)\cap\{|\nabla u|\geq 1\}} |\nabla u| dx$$

$$+ \frac{1}{2} \int_{B_{\tilde{r}}(a)\cap\{|\nabla u|<1\}} |\nabla(u-w)|^2 dx$$

$$\geq \int_{\partial B_{\tilde{r}}(a)} (u - v_\beta) \frac{\partial w}{\partial n} \, d\mathcal{H}^{n-1} + \frac{1}{2} \int_{B_{\tilde{r}}(a)} (w-I)^2 dx - \frac{1}{2} \int_{B_{\tilde{r}}(a)} (u-I)^2 dx$$

$$+ \mu \int_{B_{\tilde{r}}(a)} |D^s u| + \frac{\mu^2}{2} \int_{B_{\tilde{r}}(a)\cap\{|\nabla u|\geq 1\}} |\nabla u| dx$$

$$+ \frac{1}{2} \int_{B_{\tilde{r}}(a)\cap\{|\nabla u|<1\}} |\nabla(u-w)|^2 dx.$$

Proof. From (2.7)–(2.8), the definition of v_β, and the assumption on l we see that

$$\sup_{B_{\tilde{r}(a)}} |\nabla w| \leq \sup_{B_{\tilde{r}(a)}} |\nabla w - l| + |l|$$

$$\leq c_3(\beta^\delta + r(\|v\|_{L^\infty(\partial B_{\tilde{r}}(a))} + \|I\|_{L^\infty(\Omega)})) + 1 - 2\mu$$

$$\leq c_3(\beta^\delta + r(C_u + \|I\|_{L^\infty(\Omega)})) + 1 - 2\mu.$$

Later, v will be chosen (see, for instance, [14]) to be a Lipschitz approximation of u so that $\|v\|_{L^\infty(B_{2r}(a))}$ can be bounded by a constant C_u depending only on u. Now choose c_5 and c_6 such that $\beta^\delta \leq c_5$ and

$$r(C_u + \|I\|_{L^\infty(\Omega)}) \leq c_6$$

imply

$$c_3(\beta^\delta + r(C_u + \|I\|_{L^\infty(\Omega)})) \leq \mu.$$

Thus

$$\sup_{B_{\tilde{r}(a)}} |\nabla w| \leq 1 - \mu. \tag{2.20}$$

The conditions of Lemma 2.2 now hold for $h = w$. Substituting in w for h in the inequality in Lemma 2.2, integrating by parts, and using Young's inequality for $(u - w)(I - w) = -(u - I)(w - I) + (I - w)^2$ the lemma is proved. ∎

LEMMA 2.5. *If the function $u \in BV(\Omega)$ is solution to* (1.1), *then*

$$\int_{B_r} \varphi(Du) - \int_{B_r} \varphi(Dw) \leq 1/2 \int_{B_r} (w - I)^2 dx$$
$$-1/2 \int_{B_r} (u - I)^2 dx + \int_{\partial B_r} |Tw - Tu| d\mathcal{H}^{n-1}$$

for any $w \in BV(B_r)$, $B_r \subset\subset \Omega$. Here T denotes the trace operator on BV.

Proof. Let $w \in BV(B_r)$ and define

$$\zeta = \begin{cases} w - u & \text{on } B_r, \\ 0 & \text{in } \Omega \backslash \overline{B}_r. \end{cases}$$

Then since u is a solution we have letting $v = u + \zeta$ in (1.1) and using Theorem 1 of section 5.4 in [10],

$$\int_{\Omega} \varphi(Du) + 1/2 \int_{\Omega} (u - I)^2 dx \leq \int_{B_r} \varphi(Dw) + \int_{\partial B_r} |Tw - Tu| d\mathcal{H}^{n-1}$$
$$+ \int_{\Omega \backslash \overline{B}_r} \varphi(Du) + 1/2 \int_{B_r} (w - I)^2 dx + 1/2 \int_{\Omega \backslash B_r} (u - I)^2 dx.$$

Hence

$$\int_{\overline{B}_r} \varphi(Du) + 1/2 \int_{B_r} (u - I)^2 dx \leq \int_{B_r} \varphi(Dw) + 1/2 \int_{B_r} (w - I)^2 dx$$
$$+ \int_{\partial B_r} |Tw - Tu| d\mathcal{H}^{n-1}. \qquad \square$$

We use the above lemma, Lemma 2.4, and estimates (2.2)–(2.4) to obtain the following inequality for the solution u to (1.1).

LEMMA 2.6. *Let v, l be as in Lemma* 2.4 *with*

$$r(C_u + \|I\|_{L^\infty(\Omega)}) \leq c_6,$$

w as in (2.5), *and u a solution to* (1.1). *Then*

$$\int_{B_{\tilde{r}}(a)} |D^s u| + \int_{B_{\tilde{r}(a)} \cap \{|\nabla u| \geq 1\}} |\nabla u| dx + \int_{B_{\tilde{r}(a)} \cap \{|\nabla u| < 1\}} |\nabla (u - w)|^2 dx$$
$$\leq c_7 \int_{\partial B_{\tilde{r}}(a)} |u - v| \, d\mathcal{H}^{n-1} + c_8 r^n \beta^{1+2\delta},$$

where u and v on $\partial B_{\tilde{r}}(a)$ is understood in the sense of trace.

Proof. By the previous lemma with w from (2.5) and Lemma 2.4 we have

$$\int_{\partial B_{\tilde r}(a)} |u - v_\beta|\, d\mathcal{H}^{n-1} \geq \int_{B_{\tilde r}(a)} \varphi(Du) + \frac{1}{2}\int_{B_{\tilde r}(a)} (u - I)^2 dx$$

$$- \int_{B_{\tilde r}(a)} \varphi(\nabla w) dx - \frac{1}{2}\int_{B_{\tilde r}(a)} (w - I)^2 dx$$

$$\geq \int_{\partial B_{\tilde r}(a)} (u - v_\beta)\frac{\partial w}{\partial n}\, d\mathcal{H}^{n-1} + \mu \int_{B_{\tilde r}(a)} |D^s u| + \frac{\mu^2}{2}\int_{B_{\tilde r(a)}\cap\{|\nabla u|\geq 1\}} |\nabla u| dx$$

$$+ \frac{1}{2}\int_{B_{\tilde r(a)}\cap\{|\nabla u|<1\}} |\nabla(u - w)|^2 dx.$$

The lemma is thus proved by using (2.20) and the estimate for $|v - v_\beta|$ from (2.3). ◻

The following first variational formula is from Hardt and Kinderlehrer [12]: if u is a solution to (1.1), then

$$\int_\Omega \sigma \cdot \nabla\zeta dx + \int_\Omega \sigma \cdot \xi |D^s u| = -\int_\Omega (u - I)\zeta dx, \tag{2.21}$$

where ζ is any function in $BV_0(\Omega)$ with $D^s\zeta << |D^s u|$, ξ is the Radon–Nikodym derivative of $D^s\zeta$ with respect to $|D^s u|$, and $\sigma \in L^1(\Omega)$ is the stress tensor defined by

$$\sigma(u) = \begin{cases} \varphi_P(\nabla u) & \text{in } \Omega_a, \\ D^s u/|D^s u| & \text{in } \Omega_s. \end{cases}$$

Here $D^s u/|D^s u|$ denotes the Radon–Nikodym derivative of $D^s u$ with respect to $|D^s u|$ and $\Omega = \Omega_a \cup \Omega_s$ is the decomposition of Ω with respect to the mutually singular measures $\mathcal{L}^n$ and $|D^s u|$. Clearly $|\sigma(u)| \leq 1$. Note that $\sigma(u)$ depends only on u. In the sequel we will write σ instead of $\sigma(u)$ and write the left-hand side of (2.21) as

$$\int_\Omega \sigma \cdot D\zeta.$$

We may also note that if

$$\int_\Omega \sigma \cdot D\zeta = -\int_\Omega (u - I)\zeta dx$$

holds for arbitrary $\zeta \in BV(\Omega)$ for some u where σ is defined as above, then u solves (1.1). In fact, for arbitrary $v \in BV(\Omega)$ we take $\zeta = v - u$, noting that by convexity of φ we have $\varphi(\nabla v) - \varphi(\nabla u) \geq \nabla(v - u) \cdot \varphi_P(\nabla u)$ on Ω_a and that on Ω_s we have

$$\int_{\Omega_s} |D^s v| - \int_{\Omega_s} |D^s u| \geq \int_{\Omega_s} D^s(v - u) \cdot \frac{D^s u}{|D^s u|}.$$

The proof of the lemma below is based on [13], with some necessary modifications.

LEMMA 2.7. *Suppose u is a solution to our minimization problem, $B_{2r}(a) \subset\subset \Omega$, $v \in C^{0,1}(B_{2r}(a))$ with $\sup_{B_{2r}(a)} |\nabla v| \leq 1 - \mu$, and*

$$\mathcal{L}^n(\{u \neq v\} \cap B_\rho(a)) \leq \frac{1}{2}|B_\rho| \quad \textit{for all} \quad r \leq \rho \leq 2r;$$

then there exists positive constants c_9 and c_{10} such that if

$$\mathcal{L}^n(\{u \neq v\} \cap B_{2r}(a)) \leq c_9 r^n,$$

then

$$\|u - v\|_{L^\infty(B_r(a))} \leq c_{10} \left(\mathcal{L}^n(\{u \neq v\} \cap B_{2r}(a))\right)^{\frac{1}{n}}.$$

Proof. First we note that the function φ satisfies $|p| - \lambda \leq \varphi(p) \leq |p|$ for all $p \in \mathbf{R}^n$, some $\lambda > 0$. By convexity of φ we have $\varphi(p) \leq \varphi_P(p) \cdot p + \varphi(0)$ for all $p \in \mathbf{R}^n$. Hence we have

$$|Du| = |\nabla u|dx + |D^s u| \leq \varphi(\nabla u)dx + |D^s u| + \lambda dx$$

$$\leq \varphi_P(\nabla u) \cdot \nabla u dx + |D^s u| + (\lambda + \varphi(0))dx = \sigma \cdot Du + \lambda dx.$$

Let $\theta : \mathbf{R} \to \mathbf{R}$ be a bounded, increasing, piecewise differentiable function with $\theta'(t) \leq 1$ for almost all t. Let $0 < \rho < h$ and

$$\eta(x) = \begin{cases} 1 & \text{in } B_\rho(a), \\ (h-\rho)^{-1}(h - |x-a|) & \text{in } B_h(a) \backslash B_\rho(a), \\ 0 & \text{in } \Omega \backslash B_h(a). \end{cases}$$

Now apply the first variational formula to $\zeta = \eta\theta(u - v)$ to get

$$\int_{B_h(a)} \eta\sigma \cdot D[\theta(u-v)] = (h-\rho)^{-1} \int_{B_h(a)\backslash B_\rho(a)} \sigma \cdot \frac{x-a}{|x-a|}\theta(u-v)dx$$

$$- \int_{B_h(a)} \eta\theta(u-v)(u-I)dx. \tag{2.22}$$

To obtain a lower bound for $\eta\sigma \cdot D[\theta(u - v)]$ we use the above properties of φ. We have $D[\theta(u - v)] = \theta'(u - v)D(u - v)$ and hence by noting the bound of $|\nabla v|$

$$\int_{B_\rho(a)} |D[\theta(u-v)]| \leq \int_{B_\rho(a)} \theta'(u-v)|Du| + \int_{B_\rho(a)} \theta'(u-v)$$

$$\leq \int_{B_\rho(a)} \theta'(u-v)\varphi(Du) + \int_{B_\rho(a)} (\lambda+1)\theta'(u-v)$$

$$\leq \int_{B_\rho(a)} \theta'(u-v)\sigma \cdot Du + \int_{B_\rho(a)} (\lambda+1)\theta'(u-v)$$

$$= \int_{B_\rho(a)} \theta'(u-v)\sigma \cdot D(u-v) + \int_{B_\rho(a)} \theta'(u-v)\sigma \cdot Dv$$

$$+ \int_{B_\rho(a)} (\lambda+1)\theta'(u-v) \leq \int_{B_h(a)} \eta\sigma \cdot D[\theta(u-v)] + \int_{B_h(a)} C_\lambda \theta'(u-v) \tag{2.23}$$

for some constant C_λ depending only on λ. Inserting (2.22) into (2.23), and noting the L^∞ bound for u, we get

$$\int_{B_\rho(a)} |D[\theta(u-v)]|$$

$$\leq (h-\rho)^{-1} \int_{B_h(a)\setminus B_\rho(a)} |\theta(u-v)| dx + C_\lambda |\operatorname{supp} \eta\theta(u-v)|$$

$$+ 2\|I\|_{L^\infty(\Omega)} \int_{B_h(a)} |\theta(u-v)| dx.$$

For $0 < k < s$ we choose θ as

$$\theta(t) = \begin{cases} 0 & \text{for } t \leq k, \\ t-k & \text{for } k < t < s, \\ s-k & \text{for } t \geq s. \end{cases}$$

Now let $A(k,h) \equiv B_h \cap \{u - v > k\}$. Clearly $\operatorname{supp}[\eta\theta(u-v)] \subset A(k,h)$. Thus

$$\int_{B_\rho(a)} |D[\theta(u-v)]|$$

$$\leq ((h-\rho)^{-1} + 2\|I\|_{L^\infty(\Omega)}) \int_{B_h(a)} |\theta(u-v)| dx + C_\lambda |A(k,h)|.$$

By assumption, $|A(0,\rho)| \leq \frac{1}{2}|B_\rho(a)|$ for $r \leq \rho \leq 2r$. Thus we see that

$$\frac{\mathcal{L}^n\{\{\theta(u-v)=0\} \cap B_\rho(a)\}}{|B_\rho(a)|} \geq \frac{1}{2}.$$

We can then apply the isoperimetric inequality for $s > k > 0$ to get

$$(s-k)|A(s,\rho)|^{\frac{n-1}{n}} \leq \left(\int_{B_\rho(a)} |\theta(u-v)|^{\frac{n}{n-1}} dx \right)^{\frac{n-1}{n}}$$

$$\leq c_{11} \int_{B_\rho(a)} |D[\theta(u-v)]|$$

$$\leq c_{12}((h-\rho)^{-1} + \|I\|_{L^\infty(\Omega)}) \int_{B_h(a)} |\theta(u-v)| dx + c_{13}|A(k,h)|.$$

So since $h \leq 2r$ we get

$$(s-k)|A(s,\rho)|^{\frac{n-1}{n}} \leq c_{14}(h-p)^{-1} \int_{B_h(a)} |\theta(u-v)| dx + c_{14}|A(k,h)|.$$

And since

$$\int_{B_h(a)} |\theta(u-v)| dx \leq (s-k)|A(k,h)|,$$

we arrive at

$$|A(s,\rho)|^{\frac{n-1}{n}} \le c_{14}((h-p)^{-1} + (s-k)^{-1})|A(k,h)|$$

for every $r \le \rho < h \le 2r$ and $s > k > 0$. We now apply Lemma 2.1 in [13] to obtain the upper bound.

The lower bound for $u - v$ is obtained by using a similar argument for $0 < k < s < \infty$,

$$\tilde{\theta}(t) = \begin{cases} 0 & \text{for } t \ge -k, \\ -t-k & \text{for } -s < t < -k, \\ s-k & \text{for } t \le -s, \end{cases}$$

and $\tilde{A}(k,h) \equiv B_h \cap \{u - v < -k\}$. The lemma then follows by again applying Lemma 2.1 in [13]. ∎

Now define the energy function

$$\Phi(r,l,x) = \frac{1}{|B_r|}\left\{\int_{B_r(x)\cap\{|\nabla u|\ge 1\}} |\nabla u| dx + \int_{B_r(x)\cap\{|\nabla u|<1\}} |\nabla u - l|^2 dx + \int_{B_r(x)} |D^s u|\right\}.$$

The following theorem provides a decay estimate for Φ.

Theorem 2.8. *If u solves (1.1) with $B_r(a) \subset\subset \Omega$, $l_1 \in \mathbf{R}^n$ with $|l_1| \le 1-\mu$, then there exist positive constants ω, ϵ, κ, c_{37}, c_{38}, and c_{39} such that*

$$\Phi(4r, l_1, a) \le \epsilon$$

and

$$r \le \kappa$$

implies

$$\Phi(\omega r, l_2, a) \le \frac{1}{2}\Phi(4r, l_1, a) + c_{37} r,$$

where

$$|l_1 - l_2| \le c_{38}\Phi(4r, l_1, a)^{\frac{1}{2}} + c_{39} r.$$

Proof. For fixed $\lambda > 0$, define

$$R^\lambda \equiv \{x \in B_{2r}(a) \,|\, \Phi(\rho, l_1, x) \le \lambda \text{ for all } 0 < \rho \le 2r\}.$$

By Vitali's covering theorem, there exist disjoint balls $\{B_{r_i}(x_i)\}_{i=1}^\infty$ such that

$$B_{2r}(a)\backslash R^\lambda \subset \cup_{i=1}^\infty B_{5r_i}(x_i)$$

and $\Phi(r_i, l_1, x_i) \ge \lambda$. Then we have

$$\mathcal{L}^n(B_{2r}(a)\backslash R^\lambda) \le 5^n \sum_{i=1}^\infty |B_{r_i}(x_i)| \le \frac{5^n}{\lambda}|B_{4r}(a)|\Phi(4r, l_1, a).$$

Let $g(x) = u(x) - l_1 \cdot x$. By Poincarè's inequality we have for $x \in R^\lambda$ and $0 < \rho \le 2r$

$$\frac{1}{|B_\rho|}\int_{B_\rho(x)} |g(y) - \overline{g}_{x,\rho}|dy \le \frac{c_{15}}{\rho^{n-1}}\int_{B_\rho(x)} |Dg|$$

$$\le \frac{c_{15}}{\rho^{n-1}}\left\{2\int_{B_\rho(x)\cap\{|\nabla u|\ge 1\}} |\nabla u|dx + \int_{B_\rho(x)} |D^s u|\right.$$

$$\left. + |B_\rho|^{1/2}\left(\int_{B_\rho(x)\cap\{|\nabla u|<1\}} |\nabla u - l_1|^2 dx\right)^{1/2}\right\}$$

$$\le c_{16}\rho\Phi(\rho, l_1, x)^{1/2} \le c_{16}\lambda^{1/2}\rho,$$

where $\overline{g}_{x,\rho} = \frac{1}{|B_\rho|}\int_{B_\rho(x)} g(y)dy$. Then

$$|\overline{g}_{x,\rho/2^{k+1}} - \overline{g}_{x,\rho/2^k}| \le \frac{1}{|B_{\rho/2^{k+1}}|}\int_{B_{\rho/2^{k+1}}(x)} |g(y) - \overline{g}_{x,\rho/2^k}|dy$$

$$\le 2^n \frac{1}{|B_{\rho/2^k}|}\int_{B_{\rho/2^k}(x)} |g(y) - \overline{g}_{x,\rho/2^k}|dy \le c_{17}\rho\lambda^{1/2}/2^k.$$

Since $g(x) = \lim_{\rho\to 0}\overline{g}_{x,\rho}$ for $\mathcal{L}^n$ a.e. $x \in R^\lambda$,

$$|g(x) - \overline{g}_{x,\rho}| \le \sum_{k=1}^{\infty} |\overline{g}_{x,\rho/2^{k+1}} - \overline{g}_{x,\rho/2^k}| \le c_{17}\rho\lambda^{1/2}.$$

For x, $y \in R^\lambda$ with $|x-y| \le 2r$, set $\rho = |x-y|$. Then

$$|\overline{g}_{x,\rho} - \overline{g}_{y,\rho}| \le \frac{1}{|B_\rho(x)\cap B_\rho(y)|}\int_{B_\rho(x)\cap B_\rho(y)} |\overline{g}_{x,\rho} - g(z)| + |g(z) - \overline{g}_{y,\rho}|dz$$

$$\le c_{18}\frac{1}{B_\rho}\left(\int_{B_\rho(x)} |g(z) - \overline{g}_{x,\rho}|dz + \int_{B_\rho(y)} |g(z) - \overline{g}_{y,\rho}|dz\right) \le c_{19}\lambda^{1/2}\rho.$$

So by combining the above, we have

$$|g(x) - g(y)| \le c_{20}\lambda^{1/2}\rho = c_{20}\lambda^{1/2}|x-y|$$

for $\mathcal{L}^n$ a.e. x, $y \in R^\lambda \subset B_{2r}(a)$. Let $\lambda = c_{20}^{-2}\beta^{4\delta}$, so that

$$|u(x) - l_1 \cdot x - u(y) + l_1 \cdot y| = |g(x) - g(y)| \le \beta^{2\delta}|x-y|,$$

and let v be a Lipschitz function defined on $B_{2r}(a)$ such that

$$v = u \text{ on } R^\lambda, \text{ and } \sup_{B_{2r}(a)} |\nabla v - l_1| \le \beta^{2\delta}. \tag{2.24}$$

Such a v exists by a standard extension for a Lipschitz function. Also note that for this choice of v we have $\sup_{B_{2r}(a)} |v| \le C_u$. With the choice of λ, and by choosing

$$\beta = \Phi(4r, l_1, a) \text{ and } \delta = \frac{1}{8(n+1)},$$

we can estimate the size of the nonzero set of $u-v$ as

$$\mathcal{L}^n(B_{2r}(a)\cap\{u\neq v\})\le c_{21}r^n\beta^{-4\delta}\Phi(4r,l_1,a)\le c_{21}r^n\Phi(4r,l_1,a)^{1-4\delta}.$$

We made the choice of δ so that $(1-4\delta)\cdot\frac{n+1}{n}=1+\frac{1}{2n}>1$. Now choose $\tilde r\in[\frac12 r,r]$ so that both

$$\int_{\partial B_{\tilde r(a)}}|u-v|d\mathcal{H}^{n-1}\le\frac{3}{r}\int_{B_{\tilde r(a)}}|u-v|dx$$

and

$$\int_{\partial B_{\tilde r(a)}}|u-\overline{u}_{a,r}-l_1\cdot(x-a)|d\mathcal{H}^{n-1}\le\frac{3}{r}\int_{B_{\tilde r(a)}}|u-\overline{u}_{a,r}-l_1\cdot(x-a)|dx$$

are satisfied. By the choice of $\tilde r$,

$$\int_{\partial B_{\tilde r(a)}}|u-v|d\mathcal{H}^{n-1}\le\frac{3}{r}\|u-v\|_{L^\infty(B_r(a))}\cdot\mathcal{L}^n\{B_r(a)\cap\{u\neq v\}\}.$$

Choose $r(C_u+\|I\|_{L^\infty(\Omega)})\le c_6$. By Lemma 2.7, for $\Phi(4r,l_1,a)\le c_{22}$, we have

$$\frac{1}{r}\|u-v\|_{L^\infty(B_r(a))}\le c_{10}\frac{1}{r}(\mathcal{L}^n(B_{2r}(a)\cap\{u\neq v\}))^{1/n}.$$

Thus

$$\frac{1}{r^n}\int_{\partial B_{\tilde r(a)}}|u-v|d\mathcal{H}^{n-1}\le c_{23}\Phi(4r,l_1,a)^{1+\frac{1}{2n}}.$$

We now apply Lemma 2.6 to the above, using the estimate for the boundary integral of $u-v$, to obtain

$$\begin{aligned}\int_{B_{r\omega}(a)}|D^s u|+\int_{B_{r\omega(a)}\cap\{|\nabla u|\ge1\}}|\nabla u|dx+\int_{B_{r\omega(a)}\cap\{|\nabla u|<1\}}|\nabla(u-w)|^2dx\\ \le c_{24}r^n\left(\Phi(4r,l_1,a)^{1+\frac{1}{2n}}+\Phi(4r,l_1,a)^{1+\frac{1}{4(n+2)}}\right)\end{aligned}\tag{2.25}$$

for any $\omega\le1/2$. Let $l_2\equiv\nabla\omega(a)$. By using the gradient estimate, (2.7)–(2.8), for ω, the choice of $\tilde r$, the definition of v_β, the above bound for v, and Poincarè's inequality,

$$\begin{aligned}|l_1-l_2|&\le\frac{1}{|B_{\tilde r}|}\int_{\partial B_{\tilde r(a)}}|v_\beta-\overline{u}_{a,r}-l_1\cdot(x-a)|d\mathcal{H}^{n-1}+c_{25}r(\|I\|_{L^\infty(\Omega)}+C_u)\\ &\le\frac{1}{|B_{\tilde r}|}\int_{\partial B_{\tilde r(a)}}|v_\beta-u|+|u-\overline{u}_{a,r}-l_1\cdot(x-a)|d\mathcal{H}^{n-1}+c_{25}r(\|I\|_{L^\infty(\Omega)}+C_u)\\ &\le c_{26}\Phi(4r,l_1,a)+\frac{c_{27}}{r^n}\int_{B_r(a)}|Du-l_1|+c_{25}r(\|I\|_{L^\infty(\Omega)}+C_u).\end{aligned}$$

By the Hölder inequality, we obtain $|l_1-l_2|\le c_{28}\Phi(4r,l_1,a)^{1/2}+c_{25}r(\|I\|_{L^\infty(\Omega)}+C_u)$. The last term on the left side of inequality (2.25) satisfies

$$\int_{B_{r\omega}(a)\cap\{|\nabla u|<1\}}|\nabla(u-w)|^2dx\ge\int_{B_{r\omega}(a)\cap\{|\nabla u|<1\}}\frac12|\nabla u-l_2|^2-|\nabla w-l_2|^2dx.$$

Thus by (2.25) and the above inequality,

$$|B_{r\omega}|\Phi(r\omega, l_2, a) \le c_{29} r^n \Phi(4r, l_1, a)^{1+\frac{1}{4(n+2)}} + c_{30} \int_{B_{r\omega}} |\nabla w - l_2|^2 dx. \tag{2.26}$$

To estimate the last term, we again use the estimates for the gradient of w. Note that

$$\begin{aligned}
&\sup_{x,y\in B_{r/4}(a)} \frac{|\nabla w(x) - \nabla w(y)|}{|x-y|^{1/2}} \\
&\le c_4 \frac{1}{r^{n+1/2}} \int_{\partial B_{\tilde{r}(a)}} |v_\beta - \overline{u}_{a,r} - l_1 \cdot (x-a)| d\mathcal{H}^{n-1} \\
&+ c_4 r^{1/2}(\|I\|_{L^\infty(\Omega)} + C_u).
\end{aligned}$$

Therefore, similar to the estimate for $|l_1 - l_2|$, we have

$$\begin{aligned}
&\sup_{x,y\in B_{r/4}(a)} \frac{|\nabla w(x) - \nabla w(y)|}{|x-y|^{1/2}} \\
&\le c_{31}(r^{-1/2}\Phi(4r, l_1, a)^{1/2} + r^{1/2}(\|I\|_{L^\infty(\Omega)} + C_u)).
\end{aligned}$$

Using this we then have

$$\begin{aligned}
&\int_{B_{r\omega}(a)} |\nabla w - l_2|^2 dx \le c_{32}(r\omega)^n\{\omega\Phi(4r, l_1, a) \\
&+ r\Phi(4r, l_1, a)^{1/2}(\|I\|_{L^\infty(\Omega)} + C_u) + r^2(\|I\|_{L^\infty(\Omega)} + C_u)^2\} \\
&\le c_{33}(r\omega)^n\{\omega\Phi(4r, l_1, a) + r(\|I\|_{L^\infty(\Omega)} + C_u)\}.
\end{aligned}$$

Hence by combining the above with (2.26) we arrive at

$$\begin{aligned}
\Phi(r\omega, l_2, a) \le{}& c_{34}\omega^{-n}\Phi(4r, l_1, a)^{1+\frac{1}{4(n+1)}} + c_{35}\omega\Phi(4r, l_1, a) \\
&+ c_{36} r(\|I\|_{L^\infty(\Omega)} + C_u).
\end{aligned}$$

Choose $\omega < 1/4$ so small so that $c_{35}\omega < 1/4$, and again restrict $\Phi(4r, l_1, a)$ so that $c_{34}\omega^{-n}\Phi(4r, l_1, a)^{1+\frac{1}{4(n+1)}} < 1/4$. This now proves the theorem. ◻

We now prove Theorem 1.1 using an iteration argument (see, for example, [14] or [3]).

Proof. Assume that $\frac{1}{|B_r|}\int_{B_r(a)} |Du - l_1| \le \epsilon_0$ for some $l_1 \in \mathbf{R}^n$ with $|l_1| \le 1 - 4\mu$ and for any r with $r \le \kappa$. For each $x \in B_{r/2}(a)$ we have

$$\Phi(r/2, l_1, x) \le 2^n \Phi(r, l_1, a) \le c_{40} \frac{1}{|B_r|} \int_{B_r(a)} |Du - l_1| \le c_{40}\epsilon_0.$$

We will use Theorem 2.8 iteratively. Choose ϵ_0 so small so that $c_{40}\epsilon_0 \le \epsilon$ and restrict

r so that $c_{37}r \leq r/2$. Assume $|l_{j-1}| < 1-2\mu$ and

$$\Phi\left(\left(\frac{\omega}{4}\right)^{j-1}\frac{r}{2}, l_j, x\right) \leq \left(\frac{1}{2}\right)^{j-1}\Phi\left(\frac{r}{2}, l_1, x\right)$$
$$+\sum_{i=1}^{j-1}\left(\frac{1}{2}\right)^{j-1}\omega^{j-i-1}c_{41}r \text{ for } j = 2, \dots, k.$$

We need to show $\Phi\big(\left(\frac{\omega}{4}\right)^{k-1}\frac{r}{2}, l_k, x\big) \leq \epsilon$ and $|l_k| < 1-2\mu$ in order to continue the inductive step. Since $\omega < 1/2$,

$$\sum_{i=1}^{k-1}\left(\frac{1}{2}\right)^{i-1}\omega^{k-i-1} \leq \left(\frac{1}{2}\right)^{k-2}(k-1) \leq c_{42}\left(\frac{1}{2}\right)^{k/2}$$

for all k. By further restricting r, we have

$$\Phi\left(\left(\frac{\omega}{4}\right)^{k-1}\frac{r}{2}, l_k, x\right) \leq \epsilon.$$

Note that

$$|l_k| \leq \sum_{j=1}^{k-1}|l_{j+1}-l_j| + |l_1|$$
$$\leq \sum_{j=1}^{k-1}\left\{c_{38}\Phi\left(\left(\frac{\omega}{4}\right)^{j-1}\frac{r}{2}, l_j, x\right)^{1/2} + c_{39}\left(\frac{\omega}{4}\right)^{j-1}r\right\} + 1 - 4\mu$$
$$\leq c_{38}\sum_{j=1}^{k-1}\left\{\left(\frac{1}{2}\right)^{(j-1)/2}\Phi(\frac{r}{2}, l_1, x)^{1/2} + c_{41}^{1/2}\left(\frac{1}{2}\right)^{j/4}c_{37}^{1/2}r^{1/2}\right\}$$
$$+c_{39}r\sum_{j=1}^{k-1}\left(\frac{\omega}{4}\right)^{j-1} + 1 - 4\mu$$
$$\leq c_{42}\Phi\left(\frac{r}{2}, l_1, x\right)^{1/2} + c_{42}r^{1/2} + 1 - 4\mu.$$

So by restricting ϵ_0 and r again, we see that $|l_k| < 1-2\mu$. Thus we may continue the iterative step indefinitely, giving

$$\lim_{k\to\infty}\left(\Phi\left(\frac{\omega}{4}\right)^k\frac{r}{2}, l_{k+1}, x\right) = 0 \text{ for all } x \in B_{r/2}(a).$$

Thus

$$\lim_{\rho\to 0}\frac{1}{|B_\rho|}\left(\int_{B_\rho(x)}|D^s u| + \int_{B_\rho(x)\cap\{|\nabla u|\geq 1\}}|\nabla u|dx\right) = 0$$

for all $x \in B_{r/2}(a)$. We then have (see, for instance, [10]) $|D^s u|\left(B_{r/2}(a)\right) = 0$ with $|\nabla u| \leq 1-\mu < 1$ a.e. on $B_{r/2}(a)$. By (2.21), u also satisfies the stated equation. ∎

Using Theorem 1.1, we can now easily prove Theorem 1.2.

Proof. Assume that u is a minimizer of (1.1) and that $\tilde{E} = \{|\nabla u| < 1\}$ has positive Lebesgue measure. From standard measure theory (see, for example, [10]),

$$\lim_{r\to 0} \frac{1}{|B_r|} \int_{B_r(x)} |D^s u| = 0 \tag{2.27}$$

for $\mathcal{L}^n$-a.e. $x \in \tilde{E}$. Also, since $|\nabla u| \in L^1(\Omega)$,

$$\lim_{r\to 0} \frac{1}{|B_r|} \int_{B_r(x)} |\nabla u(y) - \nabla u(x)| dx = 0 \tag{2.28}$$

for $\mathcal{L}^n$-a.e. $x \in \tilde{E}$ by Lebesgue's differentiation theorem. Now let E be the set of all points of $\tilde{E}$ for which both (2.27) and (2.28) hold. Clearly $\mathcal{L}^n(\tilde{E}\backslash E) = 0$, $|\nabla u| < 1$ on E, and both (2.27) and (2.28) hold at each point of E. For each fixed $x \in E$, there exists some $\mu_x > 0$ such that

$$|\nabla u(x)| < 1 - 2\mu_x.$$

Then (2.27) and (2.28) combined with Theorem 1.1 show that there exists an r_x such that

$$|D^s u|(B_{r_x}(x)) = 0 \text{ and } |\nabla u| < 1 - \mu_x \text{ on } B_{r_x}(x)$$

and $u \in C^{1,\alpha}(B_{r_x}(x))$, giving $B_{r_x}(x) \subset E$ in particular. Thus E is an open set in Ω with the required properties. □

REFERENCES

[1] R. Acar and C. R. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems*, Inverse Problems, 10 (1994), p. 1217.

[2] F. Andreu-Vaillo, V. Caselles, and J. M. Mazon, *Parabolic Quasilinear Equations Minimizing Linear Growth Functionals*, Springer, New York, 2004.

[3] G. Anzellotti and M. Giaquinta, *Convex functionals and partial regularity*, Arch. Ration. Mech. Anal. 102 (1988), pp. 243–272.

[4] G. Aubert and L. Vese, *A variational method in image recovery*, SIAM J. Numer. Anal., 34 (1997), pp. 1948–1979.

[5] P. Blomgren and T. Chan, *Color TV: Total variation methods for restoration of vector-valued images*, IEEE Trans. Image Process., 7 (1998), pp. 304–309.

[6] P. Blomgren, T. Chan, P. Mulet, and C. K. Wong, *Total variation image restoration: Numerical methods and extensions*, in Proceedings of the 1997 IEEE International Conference on Image Processing, Vol. 3, pp. 384–387.

[7] A. Chambolle and P.L. Lions, *Image recovery via total variation minimization and related problems*, Numer. Math. 76 (1997), pp. 167–188.

[8] D. Dobson and F. Santosa, *Recovery of blocky images from noisy and blurred data*, SIAM J. Appl. Math., 56 (1996), pp. 1181–1198.

[9] F. Demengel and R. Temam, *Convex functions of a measure and applications*, Indiana Univ. Math. J., 33 (1984), pp. 673–709.

[10] L. Evans and R. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[11] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order, 2nd ed.,* Springer-Verlag, New York, 1983.

[12] R. Hardt and D. Kinderlehrer, *Elastic plastic deformation,* Appl. Math. Optim. 10 (1983), pp. 203–246.

[13] R. Hardt and D. Kinderlehrer, *Variational problems with linear growth,* Partial Differential Equations and the Calculus of Variations: Essays in Honor of Ennio De Giorgi, Vol. 2, Birkhäuser, Boston, 1989, pp. 633–659.

[14] R. Hardt and Y. Tonegawa, *Partial regularity for evolution problems with discontinuity,* Manuscripta Math., 90 (1996), pp. 85–103.
[15] R. Hardt and X. Zhou, *An evolution problem for linear growth functionals*, Commun. Partial Differential Equations, 19 (1994), pp. 1879–1907.
[16] L. Rudin and S. Osher, *Total variation based image restoration with free local constraints*, in Proceedings of IEEE ICIP-94, Vol. 1, Austin, TX, 1994, pp. 31–35.
[17] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D 60, (1992), pp. 259–268.
[18] R. Schoen and L. Simon, *A new proof of the regularity theorem for rectifiable currents which minimize parametric elliptic functionals*, Indiana Univ. Math. J., 31 (1982), pp. 415–434.
[19] D. Strong and T. Chan, *Edge-preserving and scale-dependent properties of total variation regularization*, Inverse Problems, 19 (2003), pp. S165–S187.
[20] L. Vese, *A study in the BV space of a denoising-deblurring variational problem*, Appl. Math. Optim., 44 (2001), pp. 131–161.
[21] X. Zhou, *An evolution problem for plastic antiplanar shear*, Appl. Math. Optim., 25 (1992), pp. 263–285.

SIAM J. MATH. ANAL.
Vol. 37, No. 4, pp. 1117–1130

EXISTENCE AND UNIQUENESS OF SOLUTIONS WITH LOW REGULARITY FOR A CLASS OF NONLINEAR DISPERSIVE EQUATIONS*

OCTAVIAN G. MUSTAFA†

Abstract. Working in Lagrangian coordinates, we prove the existence and uniqueness of solutions for a class of periodic nonlinear dispersive equations with continuously differentiable initial data. This lowers the regularity requirements available for the Cauchy problem by means of the semigroup approach for quasi-linear hyperbolic equations of evolution or by the viscosity method.

Key words. nonlinear dispersive equation, Cauchy problem

AMS subject classifications. 35Q35, 35Q72

DOI. 10.1137/040612397

1. Introduction.

This paper is devoted to the Cauchy problem

$$u_t + uu_x + \partial_x(1-\partial_x^2)^{-1}(\alpha u^2 + \beta u_x^2) = 0, \qquad x \in \mathbb{R},\ t \geq 0, \tag{1}$$

with

$$u(t, x+1) = u(t, x),\ x \in \mathbb{R},\ t \geq 0, \qquad u(0, x) = u_0(x),\ x \in \mathbb{R}, \tag{2}$$

where α, $\beta \in \mathbb{R}$ are given constants. Nonlinear periodic dispersive equations of this type, obtained for particular values of the constants α and β, has recently attracted a lot of attention in mathematical physics. Setting $\alpha = 1$ and $\beta = \frac{1}{2}$ we obtain the periodic Camassa–Holm equation modeling shallow water waves [1, 4]. This equation, originally derived as a bi-Hamiltonian system with infinitely many conservation laws [18], is also a reexpression of geodesic flow on the diffeomorphism group of the circle [23, 9, 10]. It is a completely integrable infinite-dimensional Hamiltonian system for a large class of initial data [11, 3]. Setting $\alpha = \frac{3}{2}$ and $\beta = 0$ in (1), we obtain the Degasperis–Procesi equation [16], another model for shallow water waves (cf. [17]). Finally, setting $\alpha = 3\beta - \frac{1}{2}$ with $\beta \in \mathbb{R}$, equation (1) particularizes to a model for the propagation of waves through cylindrical hyperelastic rods [15], rewritten as in [14] (see relation (2.2) therein, rescaled with $x \mapsto \frac{x}{\gamma}$). Despite their similar form, as expressed by (1), these three models have very distinct features: only the Camassa–Holm equation and the Degasperis–Procesi equation have an integrable structure (with entirely different, nonequivalent, isospectral problems) and a geometric interpretation in terms of geodesic flow is available only for the Camassa–Holm equation.

Due to its relevance to mathematical physics, the study of the Cauchy problem for (1) is therefore of interest. For initial data $u_0 \in H^s(\mathbb{S})$ (Sobolev spaces of periodic functions), it is known that there exists some $T = T(u_0) > 0$ such that (1) has a unique solution

$$u \in C([0,T); H^s(\mathbb{S})) \cap C^1([0,T); H^{s-1}(\mathbb{S})).$$

*Received by the editors July 28, 2004; accepted for publication (in revised form) August 1, 2005; published electronically December 30, 2005.
http://www.siam.org/journals/sima/37-4/61239.html
†Department of Mathematics, University of Craiova, Al. I Cuza 13, Craiova, Romania (octaviangenghiz@yahoo.com).

This fact was established in [2, 6] for $s \geq 3$ using Kato's semigroup approach for quasi-linear hyperbolic equations. For the validity of the statement for $s > \frac{3}{2}$ by implementing the same approach see [24]. In [21], the viscosity method was used to prove the existence and uniqueness of a solution $u \in C([0,T); H^s(\mathbb{S}))$ for $s > \frac{3}{2}$. For the Degasperis–Procesi equation and for the hyperelastic rod equation, an approach similar to the one devised in [2, 6] works, yielding a unique solution $u \in C([0,T); H^s(\mathbb{S})) \cap C^1([0,T); H^{s-1}(\mathbb{S}))$ for some $T = T(u_0) > 0$, provided $u_0 \in H^s(\mathbb{S})$ with $s > \frac{3}{2}$ (see [14, 27]). Let us note that in general $T < \infty$ for all three models: singularities can develop by u_x becoming unbounded in finite time [5, 7, 8, 14, 22, 27, 28]. Attempts to lower the regularity of the initial data from $u_0 \in H^s(\mathbb{S})$ with $s > \frac{3}{2}$ have been successful in the case of the Camassa–Holm equation by using the special geometric structure of the equation and imposing additional restrictions: $u_0 \in H^1(\mathbb{S})$ with $u_0 - \partial_x^2 u_0$ a positive or negative Radon measure (see [6, 12]). Further results in this direction have been obtained in [25, 26]. The motivation for seeking solutions with lower regularity is not purely academic: it is motivated by the fact that the Camassa–Holm equation, the Degasperis–Procesi equation, and the hyperelastic rod equation have peaked traveling wave solutions [1, 15, 17] and these special solutions play an important role in the dynamics of these models (see [13, 19, 20]).

In this paper, we use Lagrangian coordinates to improve the previously mentioned results by lowering the regularity requirement on the initial data for (1) to the class of continuously differentiable periodic functions. In other words, rather than viewing (1) as describing an evolution at every fixed spatial point x, we track the path of each particle using a diffeomorphism that describes at each instant the location of the particles, the initial state corresponding to the identity diffeomorphism of the real line. Such an approach was successful in proving that the periodic Camassa–Holm equation satisfies the least action principle (see [9, 10]). In contrast to [9, 10], where smooth—$C^\infty(\mathbb{S})$—initial data were of interest, we are concerned here with initial data of class C^1. We will prove that for any $u_0 \in C^1(\mathbb{S})$ there is some $T = T(u_0) > 0$ and a unique solution

$$u \in C([0,T); C^1(\mathbb{S})) \cap C^1([0,T); C(\mathbb{S}))$$

to the Cauchy problem (1). To compare this with the previously described results, note that any function $u_0 \in H^s(\mathbb{S})$ is continuously differentiable if $s > \frac{3}{2}$ but the inclusion is strict. It is also worthwhile to point out that an implementation of Kato's semigroup approach for spaces of continuously differentiable functions, rather than Sobolev spaces, leads to considerable technical difficulties. It is therefore not clear that the same conclusion as ours can be reached by this method.

2. Main result and its proof. Let us denote first by $C_1^1(\mathbb{R})$ the linear space over the real field of all continuously differentiable functions $f : \mathbb{R} \to \mathbb{R}$ that are periodic with period 1. If endowed with the Chebyshev norm

$$\|f\| = \sup_{x\in\mathbb{R}} |f(x)| + \sup_{x\in\mathbb{R}} |f'(x)|\,,$$

$C_1^1(\mathbb{R})$ becomes a Banach space. Let us denote now by $X(\mathbb{R})$ the set of all functions $g \in C^1(\mathbb{R},\mathbb{R})$ such that the derivative g' is periodic of period 1 and $g(1) = g(0) + 1$. Since

$$g(x+1) - g(x) = \int_x^{x+1} g'(s)ds = \int_0^1 g'(s)ds = 1, \qquad x \in \mathbb{R},$$

yields that the difference of two elements of $X(\mathbb{R})$ is a continuously differentiable function which is periodic of period 1, it is obvious that $X(\mathbb{R})$ does not have a linear space structure with respect to the usual sum of functions as an internal operation. However, by introducing the distance

$$d(g_1, g_2) = |g_1(0) - g_2(0)| + \sup_{x\in\mathbb{R}} |g_1'(x) - g_2'(x)|,$$

$X(\mathbb{R})$ becomes a complete metric space.

Certain features of these spaces are established by the next lemmas.

LEMMA 1. *Let $r \in (0,1)$ be fixed. The set U_r given by*

$$U_r = \{g \in X(\mathbb{R}) \mid d(Id, g) < r,\ g'(x) > 0 \text{ for } x \in \mathbb{R}\}$$

is open in $(X(\mathbb{R}), d)$.

Proof. Since $|g(0)| + |1 - g'(x)| < r$ for all real x and g' has the intermediate value property (Darboux property), the inequality

$$|g'(x)| \geq 1 - r$$

yields either $g'(x) \geq 1 - r$ or $g'(x) \leq r - 1$ for all $x \in \mathbb{R}$. Consequently, $g(\mathbb{R}) = \mathbb{R}$ and, due to its monotonicity and smoothness, g is a diffeomorphism of the real line. Further, for $g \in U_r$ fixed, let us consider $\varepsilon \in (0, \frac{1}{2}\min(r - d(Id, g), 1 - r))$. Then, for all $z \in B(g, \varepsilon)$, we have

$$\begin{aligned} |z(0)| + |1 - z'(x)| &\leq |g(0) - z(0)| + |g'(x) - z'(x)| \\ + |g(0)| + |1 - g'(x)| &< \varepsilon + (r - 2\varepsilon) \\ &< r. \end{aligned}$$

Further, since $g'(x) - \varepsilon < z'(x)$ and $g'(x) \geq 1 - r$, we obtain that $z'(t) > \frac{1}{2}(1 - r) > 0$ for all $x \in \mathbb{R}$ yielding $B(g, \varepsilon) \subset U_r$. □

LEMMA 2. *Let $r \in (0,1)$ be fixed. For $\varphi \in U_r$ and $v \in C_1^1(\mathbb{R})$, we have*

$$\varphi^{-1} \in X(\mathbb{R}) \qquad v \circ \varphi^{-1} \in C_1(\mathbb{R}).$$

Here, $C_1(\mathbb{R})$ stands for the Banach space (with Chebyshev norm) of all real-valued continuous functions which are periodic with period 1.

Proof. Since φ is onto to $\mathbb{R}$ and $\varphi(x + 1) = \varphi(x) + 1$, we deduce that

$$\varphi^{-1}(y + 1) = \varphi^{-1}(y) + 1$$

and

$$\begin{aligned} (v \circ \varphi^{-1})(y + 1) &= (v \circ \varphi^{-1})(\varphi(x + 1)) = v(x + 1) = v(x) \\ &= (v \circ \varphi^{-1})(\varphi(x)) = (v \circ \varphi^{-1})(y), \end{aligned}$$

where $y = \varphi(x)$. □

The following lemma allows us to introduce an integral representation that is essential for establishing the main result.

LEMMA 3. *Let $V \in C_1(\mathbb{R})$. There exists a unique $y \in C_1^2(\mathbb{R})$ such that*

$$y - y'' = V \text{ in } \mathbb{R}.$$

Here, $C_1^2(\mathbb{R}) = C_1(\mathbb{R}) \cap C^2(\mathbb{R}, \mathbb{R})$.

Proof. Let us fix $k \in \mathbb{R}$ and consider the boundary value problem

$$\left\{ \begin{array}{c} y'' - y = -V, \ x \in [k, k+1], \\ y(k+1) = y(k), \\ y'(k+1) = y'(k). \end{array} \right. \tag{3}$$

By introducing the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

the second order differential equation in (3) can be written as a first order differential system

$$\left\{ \begin{array}{c} z' = Az + w(x), \ x \in [k, k+1] \\ z(k+1) = z(k), \end{array} \right.$$

where

$$z(x) = \begin{pmatrix} y(x) \\ y'(x) \end{pmatrix}, \qquad w(x) = \begin{pmatrix} 0 \\ -V(x) \end{pmatrix}.$$

Using the variation of constants formula, we obtain

$$z(x) = X(x-k)z_k + \int_k^x X(x-s)w(s)ds, \tag{4}$$

where

$$X(x) = \begin{pmatrix} \operatorname{ch} x & \operatorname{sh} x \\ \operatorname{sh} x & \operatorname{ch} x \end{pmatrix}, \qquad z_k = z(k).$$

By imposing that $z_{k+1} = z_k$, formula (4) implies that

$$z_k = [I_2 - X(1)]^{-1} \int_k^{k+1} X(k+1-s)w(s)ds.$$

Formula (4) can now be recast as

$$z(x) = \int_k^{k+1} \left\{ X(x-k)[I_2 - X(1)]^{-1} X(k+1-s) + h(x,s) \right\} w(s)ds,$$

where

$$h(x,s) = \left\{ \begin{array}{c} X(x-s), \ k \le s \le x, \\ 0, \ x < s \le k+1. \end{array} \right.$$

We have used the well-known identity

$$X(x)[X(s)]^{-1} = X(x-s).$$

Straightforward computations that rely on transformations of sums of functions sh, ch into products and vice versa lead us to

$$\begin{cases} y(x) = \displaystyle\int_k^{k+1} G(x-s)V(s)ds, \\ y'(x) = \displaystyle\int_k^{k+1} G'(x-s)V(s)ds, \end{cases} \tag{5}$$

where

$$G(q) = \begin{cases} \dfrac{\text{ch } \left(q - \frac{1}{2}\right)}{2\text{sh } \frac{1}{2}}, \ q \geq 0, \\ \dfrac{\text{ch } \left(q + \frac{1}{2}\right)}{2\text{sh } \frac{1}{2}}, \ q < 0. \end{cases}$$

For $x \in \mathbb{R}$ and $k = [x]$, the periodicity of V allows us to deduce that

$$\begin{cases} y(x) = \displaystyle\int_0^{1} G(x-[x]-s)V(s)ds, \\ y'(x) = \displaystyle\int_0^{1} G'(x-[x]-s)V(s)ds. \end{cases} \tag{6}$$

Here, $[x]$ represents the integer part of x. The proof is complete. □

Let us introduce now the differential system

$$\begin{cases} \partial_t \varphi = v, \\ \partial_t v = P(\varphi, v), \end{cases} \tag{7}$$

where $P(\varphi, v) = -\partial_x(1-\partial_x^2)^{-1}V \circ \varphi$, $V = \alpha(v \circ \varphi^{-1})^2 + \beta(\partial_x(v \circ \varphi^{-1}))^2$ with α, $\beta \in \mathbb{R}$ and $\varphi(t) \in X(\mathbb{R})$, $v(t) \in C_1^1(\mathbb{R})$. According to Lemmas 2 and 3, the system is well-defined. We need also an initial datum:

$$\varphi(0) = Id, \qquad v(0) = v_0 \in C_1^1(\mathbb{R}). \tag{8}$$

Proposition 1. *Let $r \in (0,1)$ and $K > 0$ be fixed. Then, the operator P : $X(\mathbb{R}) \times C_1^1(\mathbb{R}) \to C_1^1(\mathbb{R})$ is Lipschitzian in $U_r \times \overline{B}(v_0, K)$.*

Proof. For $y = (1-\partial_x^2)^{-1}V$, we deduce that $P(\varphi, v) = y' \circ \varphi$, where y' is given by (5). To evaluate the norm of $P(\varphi, v)$, let us notice first that

$$\begin{aligned} P(\varphi, v) + \partial_x P(\varphi, v) &= y' \circ \varphi + (y'' \circ \varphi) \cdot \varphi' \\ &= y' \circ \varphi + ((y - V) \circ \varphi) \cdot \varphi'. \end{aligned}$$

Via the change of variables $s = \varphi(q)$ and the formula $\partial_x(v \circ \varphi^{-1}) = \frac{(\partial_x v) \circ \varphi^{-1}}{(\partial_x \varphi) \circ \varphi^{-1}}$, we obtain

$$\begin{aligned}
y'(\varphi(x)) &= \int_{\varphi(0)}^{\varphi(1)} G'(\varphi(x)-s)V(s)ds = \int_{\varphi(0)}^{\varphi(x)} \frac{\mathrm{sh}\ \left(\varphi(x)-s-\frac{1}{2}\right)}{2\mathrm{sh}\ \frac{1}{2}} V(s)ds \\
&\quad + \int_{\varphi(x)}^{\varphi(1)} \frac{\mathrm{sh}\ \left(\varphi(x)-s+\frac{1}{2}\right)}{2\mathrm{sh}\ \frac{1}{2}} V(s)ds \\
&= \int_0^x \frac{\mathrm{sh}\ \left(\varphi(x)-\varphi(q)-\frac{1}{2}\right)}{2\mathrm{sh}\ \frac{1}{2}} \left[\alpha v^2(q) + \beta\left(\frac{v'(q)}{\varphi'(q)}\right)^2\right] \varphi'(q)dq \\
&\quad + \int_x^1 \frac{\mathrm{sh}\ \left(\varphi(x)-\varphi(q)+\frac{1}{2}\right)}{2\mathrm{sh}\ \frac{1}{2}} \left[\alpha v^2(q) + \beta\left(\frac{v'(q)}{\varphi'(q)}\right)^2\right] \varphi'(q)dq \\
&= P_1(\varphi,v)(x) + P_2(\varphi,v)(x), \qquad x \in [0,1].
\end{aligned}$$

Now, for $\varphi_i \in U_r$, $v_i \in \overline{B}(v_0,K)$, we have

$$\begin{aligned}
&|P_1(\varphi_1,v_1)(x) - P_1(\varphi_2,v_2)(x)| \\
&\leq \int_0^1 |F(a_1,b_1,c_1,d_1,e_1)(q) - F(a_2,b_2,c_2,d_2,e_2)(q)|\, dq,
\end{aligned}$$

where

$$\begin{aligned}
&a_i(q) = \varphi_i(x), \quad b_i(q) = \varphi_i(q), \\
&c_i(q) = v_i(q), \quad\ d_i(q) = \frac{v_i'(q)}{\varphi_i'(q)}, \\
&e_i(q) = \varphi_i'(q)
\end{aligned}$$

for $i = \overline{1,2}$, and $F : \mathbb{R}^5 \to \mathbb{R}$ is given by

$$F(a,b,c,d,e) = \frac{\mathrm{sh}\ \left(a-b-\frac{1}{2}\right)}{2\mathrm{sh}\ \frac{1}{2}} \left(\alpha c^2 + \beta d^2\right) e,$$

with a, b, c, d, $e \in \mathbb{R}$. The quantities $a_i(q)$, $b_i(q)$, $c_i(q)$, $d_i(q)$, $e_i(q)$ are uniformly bounded for $q \in [0,1]$. In fact, since

$$\varphi(q) = \varphi(0) + q - \int_0^q (1-\varphi'(s))ds,$$

we have

$$\begin{aligned}
|a(q)|\,,\ \ |b(q)| &\leq |\varphi(0)| + 1 + \sup_{s\in[0,1]} |1-\varphi'(s)| \\
&< 1 + r, \\
|c(q)| &\leq \|v\|_{C_1^1(\mathbb{R})} \leq \|v_0\|_{C_1^1(\mathbb{R})} + K, \\
|d(q)| &\leq \frac{1}{1-r}\, \|v\|_{C_1^1(\mathbb{R})} \leq \frac{1}{1-r}\left(\|v_0\|_{C_1^1(\mathbb{R})} + K\right)
\end{aligned}$$

and

$$0 < e(q) \leq 1 + \sup_{s\in[0,1]} |1-\varphi'(s)| < 1 + r$$

for all $\varphi \in U_r$, $v \in \overline{B}(v_0, K)$. According to the mean value theorem

$$\begin{aligned}
&|F(a_1,b_1,c_1,d_1,e_1)(q) - F(a_2,b_2,c_2,d_2,e_2)(q)| \\
&\le \sup_{w\in[w_1,w_2]} \|\nabla F(w)\| \left(|a_1(q)-a_2(q)| + |b_1(q)-b_2(q)|\right. \\
&\qquad\qquad \left. + |c_1(q)-c_2(q)| + |d_1(q)-d_2(q)| + |e_1(q)-e_2(q)|\right) \\
&\le \sup_{w\in\mathbb{R}^5,\ \|w\|\le 5C} \|\nabla F(w)\| \left(|a_1(q)-a_2(q)| + |b_1(q)-b_2(q)|\right. \\
&\qquad\qquad \left. + |c_1(q)-c_2(q)| + |d_1(q)-d_2(q)| + |e_1(q)-e_2(q)|\right),
\end{aligned}$$

where

$$w_i = (a_i, b_i, c_i, d_i, e_i)(q) \in \mathbb{R}^5, \qquad i = \overline{1,2},$$

and $C = \max(1+r, \frac{1}{1-r}(\|v_0\|_{C_1^1(\mathbb{R})} + K))$. Before returning to the estimate of $P(\varphi, v)$, we establish that

$$\begin{aligned}
&|a_1(q) - a_2(q)|\,, \ \ |b_1(q) - b_2(q)| \\
&\le |\varphi_1(0) - \varphi_2(0)| + \int_0^{\max(x,q)} |\varphi_1'(s) - \varphi_2'(s)|\, ds \\
&\le d(\varphi_1, \varphi_2)
\end{aligned}$$

and

$$\begin{aligned}
|d_1(q) - d_2(q)| &\le \frac{|v_1'(q) - v_2'(q)|}{\varphi_1'(q)} + |v_2'(q)| \frac{|\varphi_1'(q) - \varphi_2'(q)|}{\varphi_1'(q)\varphi_2'(q)} \\
&\le \frac{1}{1-r} \|v_1 - v_2\|_{C_1^1(\mathbb{R})} + \left(\|v_0\|_{C_1^1(\mathbb{R})} + K\right) \\
&\quad \times \frac{1}{(1-r)^2} d(\varphi_1, \varphi_2) \\
&\le \frac{C}{1-r} D\left((\varphi_1, v_1), (\varphi_2, v_2)\right).
\end{aligned}$$

Here, D denotes the metric in $X(\mathbb{R}) \times C_1^1(\mathbb{R})$, namely,

$$D\left((\varphi_1, v_1), (\varphi_2, v_2)\right) = d(\varphi_1, \varphi_2) + \|v_1 - v_2\|_{C_1^1(\mathbb{R})}\,.$$

It is obvious that $(X(\mathbb{R}) \times C_1^1(\mathbb{R}), D)$ is a complete metric space. Finally,

$$\begin{aligned}
&|P_1(\varphi_1, v_1)(x) - P_1(\varphi_2, v_2)(x)| \\
&\le C(F) \left[3d(\varphi_1, \varphi_2) + \|v_1 - v_2\|_{C_1^1(\mathbb{R})} + \frac{C}{1-r} D\left((\varphi_1, v_1), (\varphi_2, v_2)\right)\right] \\
&\le \left(3 + \frac{C}{1-r}\right) C(F) \cdot D\left((\varphi_1, v_1), (\varphi_2, v_2)\right),
\end{aligned}$$

where $C(F) = \sup_{w\in\mathbb{R}^5,\ \|w\|\le 5C} \|\nabla F(w)\|$. Similar computations, performed for

$P_2(\varphi, v)$ and $\partial_x P(\varphi, v)$, respectively, where

$$\begin{aligned}\partial_x P(\varphi,v)(x) &= (y-V)(\varphi(x))\cdot\varphi'(x)\\ &= [O_1(\varphi,v)(x)+O_2(\varphi,v)(x)]\,\varphi'(x)\\ &\quad -\varphi'(x)\left[\alpha v^2(x)+\beta\left(\frac{v'(x)}{\varphi'(x)}\right)^2\right]\end{aligned}$$

and

$$O_1(\varphi,v)(x) = \int_0^x \frac{\text{ch }\left(\varphi(x)-\varphi(q)-\frac{1}{2}\right)}{2\text{sh }\frac{1}{2}}\left[\alpha v^2(q)+\beta\left(\frac{v'(q)}{\varphi'(q)}\right)^2\right]\varphi'(q)dq,$$

$$O_2(\varphi,v)(x) = \int_x^1 \frac{\text{ch }\left(\varphi(x)-\varphi(q)+\frac{1}{2}\right)}{2\text{sh }\frac{1}{2}}\left[\alpha v^2(q)+\beta\left(\frac{v'(q)}{\varphi'(q)}\right)^2\right]\varphi'(q)dq$$

allow us to complete the proof. □

PROPOSITION 2. *Let $r \in (0,1)$ and $K > 0$ be fixed and denote by c_0 the Lipschitz coefficient of P from Proposition 1. Fix also*

$$0 < T < \min\left(\frac{r}{\|v_0\|_{C_1^1(\mathbb{R})}+K}, \frac{K}{c_0(r+K)+\|P(Id,v_0)\|_{C_1^1(\mathbb{R})}}\right).$$

Then, the initial value problem

$$\begin{cases}\partial_t v = P(\varphi_v, v),\\ v(0) = v_0,\end{cases}$$

where $\varphi_v(t,x) = x + \int_0^t v(s,x)ds$, has a unique solution in $C^1([0,T], C_1^1(\mathbb{R}))$.

Proof. Denote by M the closed ball of radius K and center v_0 in $C([0,T], C_1^1(\mathbb{R}))$ and introduce the operator $\mathcal{T} : C([0,T], C_1^1(\mathbb{R})) \to C([0,T], C_1^1(\mathbb{R}))$ given by

$$(\mathcal{T}v)(t) = v_0 + \int_0^t P(\varphi_v, v)(s)ds, \qquad t \in [0,T].$$

For $v \in M$, we have $v(t) \in C_1^1(\mathbb{R})$ yielding that $\varphi_v(t) \in X(\mathbb{R})$. Furthermore,

$$\begin{aligned}d(Id,\varphi_v(t)) &= |\varphi_v(t,0)| + \sup_{x\in\mathbb{R}}\left|1-\frac{d}{dx}\varphi_v(t,x)\right| \le t\sup_{s\in[0,t]}\|v(s)\|_{C_1^1(\mathbb{R})}\\ &\le T\left(\|v_0\|_{C_1^1(\mathbb{R})}+K\right) < r,\end{aligned}$$

which implies that $\varphi_v(t) \in U_r$ for $t \in [0,T]$. Now,

$$\begin{aligned}&\|(\mathcal{T}v)(t)-v_0\|_{C_1^1(\mathbb{R})}\\ &\le \int_0^T \|P(\varphi_v,v)(s)\|_{C_1^1(\mathbb{R})}\,ds\\ &\le \int_0^T \left(c_0 D((\varphi_v(s),v(s)),(Id,v_0)) + \|P(Id,v_0)\|_{C_1^1(\mathbb{R})}\right)ds\\ &= \int_0^T \left(c_0\left(d(Id,\varphi_v(s)) + \|v_0-v(s)\|_{C_1^1(\mathbb{R})}\right) + \|P(Id,v_0)\|_{C_1^1(\mathbb{R})}\right)ds\\ &\le T\left(c_0(r+K)+\|P(Id,v_0)\|_{C_1^1(\mathbb{R})}\right) < K\end{aligned}$$

and so $\mathcal{T}(M) \subseteq M$. Finally, according to Proposition 1, we have

$$\begin{aligned}
&\|(\mathcal{T}v_1)(t) - (\mathcal{T}v_2)(t)\|_{C_1^1(\mathbb{R})} \\
&\le \int_0^t \|P(\varphi_{v_1}, v_1)(s) - P(\varphi_{v_2}, v_2)(s)\|_{C_1^1(\mathbb{R})}\, ds \\
&\le \int_0^t c_0(1+T) \sup_{q\in[0,s]} \|v_1(q) - v_2(q)\|_{C_1^1(\mathbb{R})}\, ds \\
&\le c_0(1+T) d_k(v_1, v_2) \cdot \frac{e^{kt}-1}{k}
\end{aligned}$$

and, respectively,

$$d_k\left(\mathcal{T}v_1, \mathcal{T}v_2\right) \le \frac{c_0(1+T)}{k} d_k(v_1, v_2),$$

where $k > c_0(1+T)$ is fixed and

$$d_k(v, w) = \sup_{s\in[0,T]} \left(e^{-ks} \sup_{q\in[0,s]} \|v_1(q) - v_2(q)\|_{C_1^1(\mathbb{R})} \right), \qquad v\ ,\ w \in C_1^1(\mathbb{R}).$$

The operator $\mathcal{T}$ being a contraction, the conclusion follows by application of the Banach contraction principle. □

THEOREM 1. *Let v be the solution obtained at Proposition 2. Then, the function $u : [0,T] \times \mathbb{R} \to \mathbb{R}$ given by*

$$u(t,x) = v\left(t, \varphi_v^{-1}(t,x)\right)$$

belongs to $C^1([0,T], C_1(\mathbb{R})) \cap C([0,T], C_1^1(\mathbb{R}))$, is periodic of period 1 with respect to the spatial variable, and satisfies the nonlinear dispersive equation

$$u_t + uu_x + \partial_x(1-\partial_x^2)^{-1}\left(\alpha u^2 + \beta u_x^2\right) = 0, \qquad x \in \mathbb{R},\ t \ge 0,$$

together with the initial datum $u(0) = v_0$.

Proof. The conclusion will be reached in several steps.

Step 1. *$\varphi^{-1}(\cdot, x) \in C^1([0,T], \mathbb{R})$, where $\varphi = \varphi_v$.* Consider $t_1,\ t_2 \in [0,T]$. Then,

$$\begin{aligned}
&\left|\varphi^{-1}(t_1, \varphi(t_2,x)) - \varphi^{-1}(t_2, \varphi(t_2,x))\right| \\
&= \left|\varphi^{-1}(t_1, \varphi(t_2,x)) - \varphi^{-1}(t_1, \varphi(t_1,x))\right| \\
&\le \left| \int_{\varphi(t_1,x)}^{\varphi(t_2,x)} \left|\partial_s \varphi^{-1}(t_1,s)\right| ds \right| \le \frac{1}{1-r} \left|\varphi(t_2,x) - \varphi(t_1,x)\right| \\
&\le \frac{1}{1-r} \left| \int_{t_1}^{t_2} \left|\partial_s \varphi(s,x)\right| ds \right| \le \frac{1}{1-r} \left| \int_{t_1}^{t_2} \|v(s)\|_{C_1^1(\mathbb{R})}\, ds \right| \\
&\le \frac{1}{1-r} \left(\|v_0\|_{C_1^1(\mathbb{R})} + K \right) |t_1 - t_2| .
\end{aligned}$$

Since the estimates are independent of x, by replacing x with $\varphi^{-1}(t_2, x)$, we get

$$\left|\varphi^{-1}(t_1,x) - \varphi^{-1}(t_2,x)\right| \le C\, |t_1 - t_2| . \tag{9}$$

As a Lipschitzian funtion, $\varphi^{-1}(\cdot, x)$ is a.e. differentiable. On the other hand, by differentiating formally with respect to t the identity

$$\varphi^{-1}(t, \varphi(t, x)) = x,$$

we obtain

$$\partial_t \varphi^{-1}(t, \varphi(t, x)) = -\frac{\partial_t \varphi(t, x)}{\partial_x \varphi(t, x)}$$

and, consequently,

$$\partial_t \varphi^{-1}(t, x) = -\frac{\partial_t \varphi(t, \varphi^{-1}(t, x))}{\partial_x \varphi(t, \varphi^{-1}(t, x))} \qquad \text{almost everywhere in } (0, T). \tag{10}$$

The right-hand member of (10) being a continuous function, we have

$$\varphi^{-1}(t, x) = x - \int_0^t \frac{\partial_t \varphi(s, \varphi^{-1}(s, x))}{\partial_x \varphi(s, \varphi^{-1}(s, x))} ds, \qquad t \in [0, T]. \tag{11}$$

The latter formula proves our claim, namely, that φ^{-1} is continuously differentiable with respect to t.

Step 2. $u \in C([0, T], C_1(\mathbb{R}))$ *is Lipschitzian.* We have the following estimates:

$$\begin{aligned} &|u(t_1, x) - u(t_2, x)| \\ &\le \left|v(t_1, \varphi^{-1}(t_1, x)) - v(t_1, \varphi^{-1}(t_2, x))\right| \\ &\quad + \left|v(t_1, \varphi^{-1}(t_2, x)) - v(t_2, \varphi^{-1}(t_2, x))\right| \\ &= I_1(t_1, t_2) + I_2(t_1, t_2), \end{aligned}$$

and

$$\begin{aligned} &I_1(t_1, t_2) \\ &\le \sup_{q \in \mathbb{R}} |\partial_x v(t_1, q)| \cdot \left|\varphi^{-1}(t_1, x) - \varphi^{-1}(t_2, x)\right| \\ &\le \|v(t_1)\|_{C_1^1(\mathbb{R})} C |t_1 - t_2| \le \left(\|v_0\|_{C_1^1(\mathbb{R})} + K\right) C |t_1 - t_2| \\ &\le C^2 |t_1 - t_2|, \end{aligned}$$

$$\begin{aligned} I_2(t_1, t_2) &\le \sup_{t \in [0, T]} \left|\partial_t v(t, \varphi^{-1}(t_2, x))\right| \cdot |t_1 - t_2| \\ &= \sup_{t \in [0, T]} |\partial_x y(t, x)| \cdot |t_1 - t_2| \le c_1 \cdot |t_1 - t_2|, \end{aligned}$$

where, according to (6), we take

$$\begin{aligned} c_1 &\ge \max(|\alpha|, |\beta|) C^2 \int_0^1 |\partial_x G(x - [x] - s)| \, ds \\ &\ge \left(|\alpha| \sup_{q \in \mathbb{R}} |v(t, q)|^2 + \frac{|\beta|}{(1 - r)^2} \sup_{q \in \ R} |\partial_q v(t, q)|^2\right) \\ &\quad \times \int_0^1 |\partial_x G(x - [x] - s)| \, ds. \end{aligned}$$

Step 3. *The family* $\left(\frac{\partial_t\varphi(\cdot,\varphi^{-1}(\cdot,x))}{\partial_x\varphi(\cdot,\varphi^{-1}(\cdot,x))}\right)_{x\in\mathbb{R}}$ *is equicontinuous in* $[0,T]$, that is, the usual ε - δ estimates of continuity are independent of x. Consider $t_1,\ t_2 \in [0,T]$. Then,

$$\begin{aligned}
&\left|\frac{\partial_t\varphi(t_1,\varphi^{-1}(t_1,x))}{\partial_x\varphi(t_1,\varphi^{-1}(t_1,x))} - \frac{\partial_t\varphi(t_2,\varphi^{-1}(t_2,x))}{\partial_x\varphi(t_2,\varphi^{-1}(t_2,x))}\right| \\
&= \left|\frac{v(t_1,\varphi^{-1}(t_1,x))}{\partial_x\varphi(t_1,\varphi^{-1}(t_1,x))} - \frac{v(t_2,\varphi^{-1}(t_2,x))}{\partial_x\varphi(t_2,\varphi^{-1}(t_2,x))}\right| \\
&\leq \frac{\left|v(t_1,\varphi^{-1}(t_1,x) - v(t_2,\varphi^{-1}(t_2,x))\right|}{\partial_x\varphi(t_1,\varphi^{-1}(t_1,x))} \\
&\quad + \left|v(t_2,\varphi^{-1}(t_2,x))\right| \frac{\left|\partial_x\varphi(t_1,\varphi^{-1}(t_1,x)) - \partial_x\varphi(t_2,\varphi^{-1}(t_2,x))\right|}{\partial_x\varphi(t_1,\varphi^{-1}(t_1,x))\partial_x\varphi(t_2,\varphi^{-1}(t_2,x))} \\
&= J_1(t_1,t_2) + J_2(t_1,t_2)
\end{aligned}$$

and

$$J_1(t_1,t_2) \leq \frac{1}{1-r}\left|u(t_1,x) - u(t_2,x)\right| \leq \frac{1}{1-r}\left(C^2 + c_1\right)\left|t_1 - t_2\right|,$$

$$\begin{aligned}
&J_2(t_1,t_2) \\
&\leq \|v(t_2)\|_{C_1^1(\mathbb{R})} \frac{1}{(1-r)^2}\left|\partial_x\varphi(t_1,\varphi^{-1}(t_1,x)) - \partial_x\varphi(t_2,\varphi^{-1}(t_2,x))\right| \\
&\leq \frac{C}{1-r}\left|\partial_x\varphi(t_1,\varphi^{-1}(t_1,x)) - \partial_x\varphi(t_2,\varphi^{-1}(t_2,x))\right|.
\end{aligned}$$

Since the function $\partial_x\varphi$ is periodic with period 1, its restriction to $[0,T]\times\mathbb{R}$ is uniformly continuous, that is, for every $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that

$$|\partial_x\varphi(t_1,x_1) - \partial_x\varphi(t_2,x_2)| < \varepsilon$$

for all $t_1,\ t_2 \in [0,T]$ with $|t_1 - t_2| < \delta$ and x_1 , $x_2 \in \mathbb{R}$ with $|x_1 - x_2| < \delta$. Let us denote by $\delta_1 = \delta_1(\varepsilon) > 0$ the quantity $\min\left(\delta, \frac{\delta}{C}\right)$. According to (9), for $t_1,\ t_2 \in [0,T]$ with $|t_1 - t_2| < \delta_1$ we get

$$\left|\partial_x\varphi(t_1,\varphi^{-1}(t_1,x)) - \partial_x\varphi(t_2,\varphi^{-1}(t_2,x))\right| < \varepsilon, \quad x \in \mathbb{R}. \tag{12}$$

All the preceding estimates are independent of x, thus proving our claim.

Step 4. $\varphi^{-1} \in C^1([0,T], C(\mathbb{R},\mathbb{R}))$. Consider $t,\ t+h \in [0,T]$ with $h \neq 0$. Then, formula (11) and the computations in Step 3 yield

$$\begin{aligned}
&\left|\frac{\varphi^{-1}(t+h,x) - \varphi^{-1}(t,x)}{h} - \frac{\partial_t\varphi(t,\varphi^{-1}(t,x))}{\partial_x\varphi(t,\varphi^{-1}(t,x))}\right| \\
&\leq \left|\frac{1}{h}\int_t^{t+h}\left|\frac{\partial_t\varphi(s,\varphi^{-1}(s,x))}{\partial_x\varphi(s,\varphi^{-1}(s,x))} - \frac{\partial_t\varphi(t,\varphi^{-1}(t,x))}{\partial_x\varphi(t,\varphi^{-1}(t,x))}\right| ds\right| \\
&\leq \frac{\varepsilon}{2} + \frac{1}{1-r}\left(C^2 + c_1\right)|h|
\end{aligned}$$

for $|h| < \delta_1\left(\frac{\varepsilon}{2}\right)$. By taking $|h| < \delta_2(\varepsilon) = \min\left(\frac{\varepsilon(1-r)}{2(C^2+c_1)}, \delta_1\left(\frac{\varepsilon}{2}\right)\right)$, we obtain that

$$\left|\frac{\varphi^{-1}(t+h,x)-\varphi^{-1}(t,x)}{h} - \frac{\partial_t\varphi(t,\varphi^{-1}(t,x))}{\partial_x\varphi(t,\varphi^{-1}(t,x))}\right| < \varepsilon, \qquad x \in \mathbb{R}, \tag{13}$$

which proves our claim.

Step 5. $u \in C^1([0,T], C_1(\mathbb{R}))$. Consider t, $t+h \in [0,T]$ with $h \neq 0$. The following estimates hold

$$\begin{aligned}
&\left|\frac{v(t+h,\varphi^{-1}(t+h,x)) - v(t,\varphi^{-1}(t,x))}{h}\right.\\
&\left.-\partial_t v(t,\varphi^{-1}(t,x)) - \partial_x v(t,\varphi^{-1}(t,x)) \cdot \frac{1}{h}\int_{\varphi^{-1}(t,x)}^{\varphi^{-1}(t+h,x)} ds\right|\\
\le &\left|\frac{v(t+h,\varphi^{-1}(t+h,x)) - v(t+h,\varphi^{-1}(t,x))}{h}\right.\\
&\left.-\frac{1}{h}\int_{\varphi^{-1}(t,x)}^{\varphi^{-1}(t+h,x)} \partial_x v(t,\varphi^{-1}(t,x))ds\right|\\
&+\left|\frac{v(t+h,\varphi^{-1}(t,x)) - v(t,\varphi^{-1}(t,x))}{h} - \partial_t v(t,\varphi^{-1}(t,x))\right|\\
\le &\left|\frac{1}{h}\int_{\varphi^{-1}(t,x)}^{\varphi^{-1}(t+h,x)} \left|\partial_x v(t+h,s) - \partial_x v(t,\varphi^{-1}(t,x))\right| ds\right|\\
&+\sup_{q\in\mathbb{R}}\left|\frac{v(t+h,q)-v(t,q)}{h} - \partial_t v(t,q)\right|\\
= &\, E(h) + o(1) \qquad \text{as } h \to 0,
\end{aligned}$$

since $v \in C^1([0,T], C_1^1(\mathbb{R}))$ according to Proposition 2. Similar to the situation described at Step 3, the restriction of function $\partial_x v$ to $[0,T]\times\mathbb{R}$ is uniformly continuous and so

$$\left|\partial_x v(t+h,s) - \partial_x v(t,\varphi^{-1}(t,x))\right| < \varepsilon \tag{14}$$

for $|h|$ small enough provided that $\left|s - \varphi^{-1}(t,x)\right| \le \left|\varphi^{-1}(t+h,x) - \varphi^{-1}(t,x)\right| \le C\,|h|$ is also small. Since

$$E(h) \le \varepsilon\left|\frac{\varphi^{-1}(t+h,x)-\varphi^{-1}(t,x)}{h}\right| \le C\varepsilon, \qquad x \in \mathbb{R},$$

and taking into account (13), our claim is established.

Step 6. $u \in C([0,T], C_1^1(\mathbb{R}))$. We shall confine ourselves to the issue of $\partial_x u \in C([0,T], C_1(\mathbb{R}))$ since $u \in C([0,T], C_1(\mathbb{R}))$ according to Step 2. Consider t_1, $t_2 \in$

$[0, T]$. Then,

$$
\begin{aligned}
&|\partial_x u(t_1, x) - \partial_x u(t_2, x)| \\
\leq\ & \frac{\left|\partial_x v(t_1, \varphi^{-1}(t_1, x)) - \partial_x v(t_2, \varphi^{-1}(t_2, x))\right|}{\partial_x \varphi(t_1, \varphi^{-1}(t_1, x))} \\
&+ \left|\partial_x v(t_2, \varphi^{-1}(t_2, x))\right| \frac{\left|\partial_x \varphi(t_1, \varphi^{-1}(t_1, x)) - \partial_x \varphi(t_2, \varphi^{-1}(t_2, x))\right|}{\partial_x \varphi(t_1, \varphi^{-1}(t_1, x)) \partial_x \varphi(t_2, \varphi^{-1}(t_2, x))} \\
\leq\ & \frac{1}{1-r} \left|\partial_x v(t_1, \varphi^{-1}(t_1, x)) - \partial_x v(t_2, \varphi^{-1}(t_2, x))\right| \\
&+ \frac{C}{1-r} \left|\partial_x \varphi(t_1, \varphi^{-1}(t_1, x)) - \partial_x \varphi(t_2, \varphi^{-1}(t_2, x))\right|.
\end{aligned}
$$

A simple inspection of formulas (12), (14) establishes our claim.

The proof is complete. □

Acknowledgment. The author is deeply indebted to the referees for their constructive comments.

REFERENCES

[1] R. Camassa and D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.

[2] A. Constantin, *On the Cauchy problem for the periodic Camassa-Holm equation*, J. Differential Equations, 141 (1997), pp. 218–235.

[3] A. Constantin, *On the inverse spectral problem for the Camassa-Holm equation*, J. Funct. Anal. 155 (1998), pp. 352–363.

[4] A. Constantin, *A Lagrangian approximation to the water-wave problem*, Appl. Math. Lett., 14 (2001), pp. 789–795.

[5] A. Constantin and J. Escher, *Wave breaking for nonlinear nonlocal shallow water equations*, Acta Math., 181 (1998), pp. 229–243.

[6] A. Constantin and J. Escher, *Global weak solutions for a shallow water equation*, Indiana Univ. Math. J., 47 (1998), pp. 1527–1545.

[7] A. Constantin and J. Escher, *Well-posedness, global existence, and blowup phenomena for a periodic quasi-linear hyperbolic equation*, Comm. Pure Appl. Math., 51 (1998), pp. 475–504.

[8] A. Constantin and J. Escher, *On the blow-up rate and the blow-up set of breaking waves for a shallow water equation*, Math. Z., 233 (2000), pp. 75–91.

[9] A. Constantin and B. Kolev, *On the geometric approach to the motion of inertial mechanical systems*, J. Phys. A, 35 (2002), pp. R51–R79.

[10] A. Constantin and B. Kolev, *Geodesic flow on the diffeomorphism group of the circle*, Comment. Math. Helv., 78 (2003), pp. 787–804.

[11] A. Constantin and H. P. McKean, *A shallow water equation on the circle*, Comm. Pure Appl. Math., 52 (1999), pp. 949–982.

[12] A. Constantin and L. Molinet, *Global weak solutions for a shallow water equation*, Comm. Math. Phys., 211 (2000), pp. 45–61.

[13] A. Constantin and W. Strauss, *Stability of peakons*, Comm. Pure Appl. Math., 53 (2000), pp. 603–610.

[14] A. Constantin and W. Strauss, *Stability of a class of solitary waves in compressible elastic rods*, Phys. Lett. A, 270 (2000), pp. 140–148.

[15] H. H. Dai and Y. Huo, *Solitary shock waves and other travelling waves in a general compressible hyperelastic rod*, Proc. Roy. Soc. London, 456 (2000), pp. 331–363.

[16] A. Degasperis and M. Procesi, *Asymptotic integrability*, in Symmetry and Perturbation Theory (Rome, 1998), World Science Publishing, River Edge, NJ, 1999, pp. 23–37.

[17] H. Dullin, G. Gottwald, and D. Holm, *Camassa-Holm, Korteweg-de Vries-*5 *and other asymptotically equivalent equations for shallow water waves*, Fluid Dynam. Res., 33 (2003), pp. 73–95.

[18] A. S. Fokas and B. Fuchssteiner, *Symplectic structures, their Bäcklund transformation and hereditary symmetries*, Phys. D, 4 (1981), pp. 47–66.

[19] J. Lenells, *Stability of periodic peakons*, Int. Math. Res. Not., 10 (2004), pp. 485–499.

[20] J. Lenells, *A variational approach to the stability of periodic peakons*, J. Nonlinear Math. Phys., 11 (2004), pp. 151–163.

[21] Y. Li and P. Olver, *Well-posedness and blow-up solutions for an integrable nonlinearly dispersive model wave equation*, J. Differential Equations, 162 (2000), pp. 27–63.

[22] H. P. McKean, *Breakdown of a shallow water equation*, Asian J. Math., 2 (1998), pp. 867–874.

[23] G. Misiołek, *A shallow water equation as a geodesic flow on the Bott-Virasoro group*, J. Geom. Phys., 24 (1998), pp. 203–208.

[24] G. Rodríguez-Blanco, *On the Cauchy problem for the Camassa-Holm equation*, Nonlinear Anal., 46 (2001), pp. 309–327.

[25] Z. Xin and P. Zhang, *On the weak solution to a shallow water equation*, Comm. Pure Appl. Math., 53 (2000), pp. 1411–1433.

[26] Z. Xin and P. Zhang, *On the uniqueness and large time behavior of the weak solutions to a shallow water equation*, Comm. Partial Differential Equations, 27 (2002), pp. 1815–1844.

[27] Y. Zhou, *Blow-up phenomenon for the integrable Degasperis-Procesi equation*, Phys. Lett. A, 328 (2004), pp. 157–162.

[28] Y. Zhou, *Blow-up of solutions to a nonlinear dispersive rod equation*, Calc. Var. Partial Differential Equations, to appear.

SIAM J. MATH. ANAL.
Vol. 37, No. 4, pp. 1131–1160

RENORMALIZED ENERGY AND FORCES ON DISLOCATIONS*

PAOLO CERMELLI† AND GIOVANNI LEONI‡

Abstract. In this work we discuss, from a variational viewpoint, the equilibrium problem for a finite number of Volterra dislocations in a plane domain. For a given set of singularities at fixed locations, we characterize elastic equilibrium as the limit of the minimizers of a family of energy functionals, obtained by a finite-core regularization of the elastic-energy functional. We give a sharp asymptotic estimate of the minimum energy as the core radius tends to zero, which allows one to eliminate this internal length scale from the problem. The energy content of a set of dislocations is fully characterized by the regular part of the asymptotic expansion, the so-called renormalized energy, which contains all information regarding self- and mutual interactions between the defects. Thus our result may be considered as the analogue for dislocations of the classical result of Bethuel, Brezis and Hélein for Ginzburg–Landau vortices. We view the renormalized energy as the basic tool for the study of the discrete-to-continuum limit in plasticity of crystals, i.e., the passage from models of isolated defects to theories of continuous distributions of dislocations. The renormalized energy is a function of the defect positions only: we prove that its derivative with respect to the position of a given dislocation is the resultant of the Eshelby stress on that dislocation, which can be identified in turn with the classical Peach–Köhler force.

Key words. dislocations, variational methods, variational techniques for singularities, forces on defects

AMS subject classifications. 35J50, 74G65, 74G70

DOI. 10.1137/040621636

1. Introduction.

Dislocations are common defects in crystals and influence their behavior in multiple ways. For instance, isolated dislocations generate concentrations of stress which affect the chemical and electronic properties of solids, while the collective motion of large sets of dislocations represents the basic mechanism for plastic slip in ductile solids (cf., e.g., [3], [16], and [23]).

Hence, it is of considerable interest to study the behavior of both isolated and large sets of dislocations.

However, the study of isolated dislocations and of large clusters of defects requires widely different approaches. Problems involving isolated defects involve scales which are typically of the order of the interatomic distances in the crystal, while the characteristic scales involved in the collective behavior of large clusters of dislocations, typically in plasticity, are much larger. A typical example is self-organization of stored dislocations in cell patterns [25]: the characteristic distance between the cell walls is macroscopic, many orders of magnitude larger than the interatomic distances.

Such problems are better studied in terms of dislocation densities, rather than of isolated dislocations, and require the introduction, in the expression for the macroscopic energy of the solid, of terms which depend on the gradients of the (plastic)

*Received by the editors December 28, 2004; accepted for publication (in revised form) August 15, 2005; published electronically December 30, 2005. This research was partially supported by the Center for Nonlinear Analysis (NSF grant DMS–9803791), Carnegie Mellon University, Pittsburgh, PA, and the Italian MIUR ("Modelli matematici in Scienza dei Materiali" 2002).
http://www.siam.org/journals/sima/37-4/62163.html

†Dipartimento di Matematica, Università di Torino, Via Carlo Alberto 10, I–10123 Torino, Italy (paolo.cermelli@unito.it).

‡Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (giovanni@andrew.cmu.edu). The research of this author was partially supported by the National Science Foundation under Grant DMS–0405423.

strain [1], [15], [18], [19], [20]. These terms are necessarily phenomenological—for instance, energies are assumed to be quadratic in the plastic strain gradients, but such simple choices often lead to unphysical behavior, as shown in [8] for interfacial dislocations in epitaxial films.

At a still larger scale lives classical plasticity: plastic strain gradients are ignored, no internal length scale is introduced, and dislocations are only implicitly taken into account. Classical models cannot describe the self-organization of defects in regular patterns.

Hence, a major open problem in the theory of defects in solids is to correlate the microscopic (isolated defects) and the macroscopic (gradient theories) approaches. Specifically, it would be useful to develop a theoretical framework which allows to characterize the constitutive relations of the continuum models, using the information gained by ab initio models of finite sets of dislocations.

The goal for this paper may be viewed as the first stage of this project: we give a variational formulation of the equilibrium problem for a finite number of dislocations in a plane domain and characterize the energy content of a body with isolated defects in terms of a regular function of the defect configuration, the so-called renormalized energy.

Precisely, consider a finite number of dislocations in an elastic solid: since the stress field induced by a dislocation is short ranged, it is reasonable to work in the approximation of linear elasticity, which may be assumed to be valid sufficiently far from the defect. (This topic has been studied extensively in the literature, and explicit solutions are known in special cases [28], [23], [29].) We restrict attention to plane isotropic elasticity.[1] Let Ω be a regular domain in $\mathbb{R}^2$: in linear elasticity, a displacement of Ω is a regular vector field on Ω, with gradient $\nabla \; = \;$. The equilibrium equations have the form $\operatorname{Div} C[\;(\;)] = 0$ with C a linear operator from $\mathbb{R}^{2\times 2}$ into itself and $(\;) = \frac{1}{2}(\nabla \; + (\nabla \;)^\top)$ the infinitesimal strain tensor.

In this framework, Volterra dislocations may be viewed as singularities of the field . Precisely, fix a finite set of points $\{\;_1, \dots, \;_N\}$ in Ω and a set of vectors $\{\;_1, \dots, \;_N\}$ with $\;_i \in \mathbb{R}^2$: we say that a tensor field on $\Omega \setminus \{\;_1, \dots, \;_N\}$ corresponds to a system of dislocations located at $\{\;_1, \dots, \;_N\}$ with Burgers vectors $\{\;_1, \dots, \;_N\}$ if[2]

$$\begin{cases} \operatorname{Curl} \; = \sum_{i=1}^N \;_i \, \delta_{\;_i} \\ \operatorname{Div} C[\;(\;)] = 0 \end{cases} \quad \text{in } \Omega \tag{1.1}$$

in the sense of distributions, where $(\;) = \frac{1}{2}(\; + \;^\top)$ is the strain associated to .

Solutions of (1.1) are not unique even modulo an infinitesimal rigid motion and, moreover, no variational principle may be associated to (1.1), since the elastic energy of a system of Volterra dislocations is not finite.

Hence, it is necessary to regularize the theory by removing a core $B_\varepsilon(\;_i)$ of radius ε around each dislocation; letting $\Omega_\varepsilon = \Omega \setminus (\cup_{i=1}^N B_\varepsilon(\;_i))$, we solve the family

[1]Point defects in plane elasticity may be effectively used to model straight edge dislocations orthogonal to the plane of strain.

[2]The Curl of a two-dimensional tensor field $\boldsymbol{H}$ is the vector field whose Cartesian components are $(\operatorname{Curl} \boldsymbol{H})_i = (\partial_1 H_{i2} - \partial_2 H_{i1})$.

of minimization problems

$$\min_{\boldsymbol{H}\in H(\ _1,\ldots,\ _N;\Omega_\varepsilon)} \int_{\Omega_\varepsilon} W(\ \ (\ \ ))\, da, \tag{1.2}$$

where $W(\ \ ) = \frac{1}{2}\ \ \cdot C[\ \ ]$ is the elastic energy density,

$$H(\ _1,\ldots,\ _N;\Omega_\varepsilon) = \left\{\ \ \in H(\mathrm{Curl}\,0;\Omega_\varepsilon)\ :\ \int_{\partial B_\varepsilon(\ _i)}\ \ \ ds =\ _i,\quad i=1,\ldots,N\right\},$$

and is the unit tangent vector[3] to $\partial B_\varepsilon(\ _i)$.

Our first result shows that the solutions $\ _\varepsilon$ of (1.2) converge strongly in $L^2_{\mathrm{loc}}(\Omega\setminus\cup_{i=1}^N\{\ _i\};\mathbb{R}^{2\times2})$, as $\varepsilon\to0$, to a solution $\ _0$ of (1.1). This solution is unique modulo a rigid motion. More precisely, we show that

$$\ _\varepsilon \to\ \ _0 = \sum_{i=1}^N\ \ _i + \nabla\ \ _0, \tag{1.3}$$

where $\ _i$ are distributional solutions of (cf. Proposition 3.1),

$$\begin{cases}\mathrm{Curl}\ \ \ = \ _i\,\delta_{\ _i}\\ \mathrm{Div}\,C[\ \ (\ \ )] = 0\end{cases}\qquad \text{in } \mathbb{R}^2,$$

and $\ _0\in H^1(\Omega;\mathbb{R}^2)$ is a regular displacement field which is a minimizer of the functional

$$I_0(\ \ ) := \int_\Omega W(\ \ (\ \ ))\,da + \sum_{i=1}^N \int_{\partial\Omega}\ \ \cdot C[\ \ (\ \ _i)]\ \ \ ds \tag{1.4}$$

on $H^1(\Omega;\mathbb{R}^2)$.

The field $\ _0$ is independent of the internal length scale ε, but its energy is not finite: we obtain a sharp asymptotic estimate as $\varepsilon\to0$ for the minimum energy in (1.2) of the form

(1.5)
$$\int_{\Omega_\varepsilon} W(\ \ (\ \ _\varepsilon))\,da = \sum_{i=1}^N \frac{\mu(\lambda+\mu)}{4\pi(\lambda+2\mu)}|\ _i|^2 \ln\frac{1}{\varepsilon} + F(\ _1,\ldots,\ _N) + O(\varepsilon) + \mathrm{Const.},$$

where λ, μ are the Lamé moduli, and

$$F(\ _1,\ldots,\ _N) = F_{\mathrm{self}}(\ _1,\ldots,\ _N) + F_{\mathrm{int}}(\ _1,\ldots,\ _N) + F_{\mathrm{elastic}}(\ _1,\ldots,\ _N) \tag{1.6}$$

[3]We choose $\boldsymbol{t} = \boldsymbol{n}^\perp$ to be a counterclockwise $\pi/2$-rotation of the outward unit normal $\boldsymbol{n}$ to $\partial B_\varepsilon(\boldsymbol{x}_i)$.

is the *renormalized energy*, with

(1.7)
$$\begin{cases} F_{\mathrm{self}}(\ _1, \dots, \ _N) = \displaystyle\sum_{i=1}^{N} \int_{\Omega \setminus B_R(\ _i)} W(\ (\ _i))\, da + \sum_{i=1}^{N} \frac{\mu(\lambda+\mu)}{4\pi(\lambda+2\mu)} |\ _i|^2 \ln R, \\ F_{\mathrm{int}}(\ _1, \dots, \ _N) = \displaystyle\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \int_{\Omega} C[\ (\ _i)] \cdot \ (\ _j)\, da, \\ F_{\mathrm{elastic}}(\ _1, \dots, \ _N) = \displaystyle\int_{\Omega} W(\ (\ _0))\, da + \sum_{i=1}^{N} \int_{\partial\Omega} \ _0 \cdot C[\ (\ _i)] \ \ ds, \end{cases}$$

and

$$0 < R < \frac{1}{4} \min\{|\ -\ | : \ \neq \ , (\ ,\ ) \in S \times (S \cup \partial\Omega)\},$$

where $S = \{\ _1, \dots, \ _N\}$. It can be shown that F_{self} is independent of R.

It is important to remark that while for special domains the asymptotic formula

$$\int_{\Omega_\varepsilon} W(\ (\ _\varepsilon))\, da \sim \sum_{i=1}^{N} \frac{\mu(\lambda+\mu)}{4\pi(\lambda+2\mu)} |\ _i|^2 \ln \frac{1}{\varepsilon} \tag{1.8}$$

is classical (see, e.g., [23], [29]) and can be obtained by solving explicitly the Euler equation (1.1) (see [28]), for general domains there are various formal arguments in support of (1.8) but we are not aware of any rigorous derivation prior to ours.

More importantly, the introduction of the renormalized energy in this context appears to be new, and thus our result may be considered as the analogue for dislocations of the classical result of Bethuel, Brezis, and Hélein (see Chapter 2 in [5]; see also [4]) for Ginzburg–Landau vortices. We refer to the monograph [5] for more details about the Ginzburg–Landau functional. (See also [2], [6], [22] and the references contained therein for more recent results.)

Note that the renormalized energy is independent of the core radius and is a function of the defect position which fully characterizes the energy content of a dislocated body. Hence, it provides a basis for the study of the behavior of finite sets of dislocations.

As an example application of these ideas, we prove that the interaction energy F_{int} in $(1.7)_2$ diverges logarithmically with the relative distance between the defects:

$$F_{\mathrm{int}}(\ _1, \dots, \ _N) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\mu(\lambda+\mu)}{\pi(\lambda+2\mu)} \ _i \cdot \ _j \ln \frac{1}{|\ _i - \ _j|} + O(1)$$

as $|\ _i - \ _j| \to 0$.

When more than one dislocation is present, or an external stress is applied to the dislocated body, defects interact between themselves and with the applied field, by means of the so-called Peach–Köhler force [30]. Since the renormalized energy contains all the information about defect interactions, a natural question is whether it is somehow related to the Peach–Köhler force on dislocations. Indeed, the asymptotic analysis of a regularized Ginzburg–Landau equation, intended to model disclinations in liquid crystals, shows that on a long time scale defects move according to a simple evolution equation, which has the form velocity = force on the defect [21], [27], [26], where the force on the defect is defined as the derivative of the renormalized energy with respect to the defect position.

In this line of thought, we prove the fundamental relation

$$\nabla_{\ \ k} F = -\int_{\partial B_R(\ \ k)} \Big\{ W(\ \ (\ \ _0))\mathbf{1} - \ \ _0^{\top} C[\ \ (\ \ _0)] \Big\} \ \ ds \tag{1.9}$$

for $R < \min_i \frac{1}{2}\mathrm{d}(\ \ _i, \partial\Omega)$, where the integrand $\ \ = W(\ \ (\ \ _0))\mathbf{1} - \ \ _0^{\top} C[\ \ (\ \ _0)]$ is called the Eshelby stress. This object, also known as configurational stress, is usually introduced in continuum mechanics in conjunction with an additional balance law, the configurational balance, when defective structures such as interfaces, cracks, or inclusions are present [18]. The configurational balance governs the evolution of the defect, and the resultant of the Eshelby stress in (1.9) may be identified with the force acting on a defect. In the theory of elastic dislocations, the force on a defect is defined by means of the Peach–Köhler force, and indeed it can be shown that the resultant of the configurational stress coincides with the Peach–Köhler force on a dislocation [7].

Hence, (1.9) shows that the derivative of the renormalized energy coincides with the force on a dislocation.

The idea that the force on a defect is the derivative of the minimum energy with respect to changes of the defect position is the basis of Eshelby's treatment of defects [12], [13], [11]. However, when dislocations are present the energy is not finite, so that Eshelby's approach fails without modifications. Our result may be viewed as the generalization of Eshelby's notion of force on a defect, when bad singularities are associated to the defect itself.

2. The variational problem. Let $\Omega \subset \mathbb{R}^2$ be a simply connected bounded open domain with smooth boundary $\partial\Omega$, with outward unit normal $\ \ $. In the absence of defect, we denote by $\ \ : \Omega \to \mathbb{R}^2$ the *displacement field*, with *displacement gradient* $\nabla\ \ $ and *strain tensor* $\ \ (\ \ ) = \frac{1}{2}\left(\nabla\ \ + (\nabla\ \ )^{\top}\right)$. We write

$$\ \ = C[\ \ ]$$

for the (symmetric) Cauchy stress, with $C : \mathrm{Sym} \to \mathrm{Sym}$ the *elasticity tensor*, a symmetric[4] linear map on the space Sym of symmetric 2×2 tensors. For isotropic materials, the stress has the form

$$C[\ \ ] = \lambda(\mathrm{tr}\ \ )\mathbf{1} + 2\mu\ \ \tag{2.1}$$

with λ, μ the Lamé moduli. The associated *energy functional* is

$$J(\ \ ) = \int_{\Omega} W(\ \ (\ \ ))\, da, \tag{2.2}$$

which is defined on $H^1(\Omega;\mathbb{R}^2)$. Here $W(\ \ ) = \frac{1}{2}\ \ \cdot C[\ \ ]$ is the *strain energy density*, and we assume that the elasticity tensor C is positive definite.[5] In plane elasticity and for isotropic materials this is equivalent to requiring

$$\mu > 0 \quad \text{and} \quad \lambda + \mu > 0. \tag{2.3}$$

[4]That is, $\boldsymbol{E} \cdot C[\boldsymbol{F}] = \boldsymbol{F} \cdot C[\boldsymbol{E}]$ for any $\boldsymbol{E}, \boldsymbol{F} \in \mathrm{Sym}$, where $\cdot$ is the inner product of 2×2 tensors.

[5]This implies that there exist constants $c_1, c_2 > 0$ such that $c_1|\boldsymbol{E}|^2 \le W(\boldsymbol{E}) \le c_2|\boldsymbol{E}|^2$ for any $\boldsymbol{E} \in \mathrm{Sym}$.

In this paper we shall deal with isotropic materials only.

When defects such as dislocations are present, the displacement field is not single-valued, and the equilibrium problem must be formulated in terms of a 2×2 tensor field , defined away from the defects, and such that Curl $= 0$. The field plays the role of displacement gradient but is not necessarily the gradient of a displacement field globally defined on Ω: we will continue to use the denomination *strain tensor* associated to for the symmetric part of , writing

$$(\ ) = \frac{1}{2}\Big(\quad + \quad ^{\top} \Big). \tag{2.4}$$

More precisely, we are interested in situations in which the field has a finite number of singularities in Ω: to this purpose, let $\{ \ _i\}_{i=1,\dots,N}$ be a finite sequence of points in Ω, and for $\varepsilon > 0$ let

$$\Omega_\varepsilon = \Omega \setminus \left(\bigcup_{i=1}^{N} B_\varepsilon(\ _i) \right),$$

and consider the space

$$H(\mathrm{Curl};\Omega_\varepsilon) := \big\{ \quad \in L^2(\Omega_\varepsilon;\mathbb{R}^{2\times 2}) \,:\, \mathrm{Curl} \quad \in L^2(\Omega_\varepsilon;\mathbb{R}^{2\times 2}) \big\}. \tag{2.5}$$

Following [10] we set

$$H(\mathrm{Curl}\,0;\Omega_\varepsilon) := \{ \quad \in H(\mathrm{Curl};\Omega_\varepsilon) \,:\, \mathrm{Curl} \quad = \mathbf{0} \}.$$

We say that $\in H(\mathrm{Curl}\,0;\Omega_\varepsilon)$ corresponds to a *system of dislocations* located at $\ _i$, with *Burgers vectors* $\ _i$ and *cores* $B_\varepsilon(\ _i)$, if

$$\int_{\partial B_\varepsilon(\ _i)} \quad ds = \ _i, \quad i = 1,\dots,N, \tag{2.6}$$

with the unit tangent to $\partial B_\varepsilon(\ _i)$, obtained by rotating counterclockwise by $\pi/2$ the outward unit normal to $\partial B_\varepsilon(\ _i)$. Here we have used the fact that for each $i = 1,\dots,N$ the trace map

$$\mapsto$$

defined on $C^\infty(\overline{\Omega_\varepsilon};\mathbb{R}^{2\times 2})$ extends by continuity to a continuous linear mapping, still denoted , from $H(\mathrm{Curl};\Omega_\varepsilon)$ to $H^{-\frac{1}{2}}(\partial B_\varepsilon(\ _i);\mathbb{R}^2)$. (See, e.g., Theorem 2 in [10].) With an abuse of notation for every $\varphi \in H^{\frac{1}{2}}(\partial B_\varepsilon(\ _i))$ we continue to denote by $\int_{\partial B_\varepsilon(\ _i)} \varphi \quad ds$ the value of the linear mapping applied to φ. We shall denote by $H(\ _1,\dots,\ _N;\Omega_\varepsilon)$ the closed subspace of $H(\mathrm{Curl}\,0;\Omega_\varepsilon)$ of tensor fields corresponding to systems of dislocations with Burgers vectors $\ _i$, i.e.,

$$H(\ _1,\dots,\ _N;\Omega_\varepsilon) := \left\{ \quad \in H(\mathrm{Curl}\,0;\Omega_\varepsilon) \,:\, \int_{\partial B_\varepsilon(\ _i)} \quad ds = \ _i, \quad i = 1,\dots,N \right\}. \tag{2.7}$$

The *strain energy functional* is defined as in the absence of defects (cf. (2.2)),

$$J_\varepsilon(\ ) = \int_{\Omega_\varepsilon} W(\ (\ ))\, da, \tag{2.8}$$

and the associated minimization problem is

$(M_{\boldsymbol{H},\varepsilon})$: minimize the strain energy functional over all systems of dislocations located at given points $(\ _1, \dots, \ _N)$, and with given Burgers vectors $(\ _1, \dots, \ _N)$, i.e., find the solutions of

$$\min_{\boldsymbol{H} \in H(\ _1,\dots, \ _N;\Omega_\varepsilon)} J_\varepsilon(\quad). \tag{2.9}$$

PROPOSITION 2.1. *Assume that the elasticity tensor C satisfies condition* (2.3). *Then*

$$\ _\varepsilon \in H(\ _1, \dots, \ _N; \Omega_\varepsilon)$$

is a minimizer of (2.9) *if and only if* $\ _\varepsilon$ *is a weak solution of the Neumann boundary problem*

$$\begin{cases} \operatorname{Div} C[\ (\ _\varepsilon)] = \mathbf{0} & \text{in } \Omega_\varepsilon, \\ C[\ (\ _\varepsilon)] \ = 0 & \text{on } \partial\Omega_\varepsilon = \partial\Omega \cup \left(\cup_{i=1}^N \partial B_\varepsilon(\ _i)\right). \end{cases} \tag{2.10}$$

Moreover, $\ _\varepsilon$ *is unique modulo an infinitesimal rigid-body motion.*

Proof. Since J_ε is quadratic it follows from standard arguments in the calculus of variations that $\ _\varepsilon$ is a minimizer if and only if it satisfies the weak Euler equation

$$\int_{\Omega_\varepsilon} C[\ _\varepsilon] \cdot \ (\ ) \, da = 0 \qquad \text{for all } \ \in H^1(\Omega_\varepsilon; \mathbb{R}^2). \tag{2.11}$$

Indeed, note that for every $\ , \tilde{\ } \in H(\ _1, \dots, \ _N; \Omega_\varepsilon)$ there exists $\ \in H^1(\Omega_\varepsilon; \mathbb{R}^2)$ such that $\tilde{\ } = \ + \nabla \ $: moreover, for t a real parameter,

$$J_\varepsilon(\ + t\nabla \ ) - J_\varepsilon(\ ) = t \int_{\Omega_\varepsilon} C[\ ] \cdot \ (\ ) \, da + t^2 J_\varepsilon(\nabla \ ),$$

and this proves the assertion.

To prove uniqueness let $\ _\varepsilon$ and $\ '_\varepsilon$ be two solutions of (2.11); then $\ '_\varepsilon = \ _\varepsilon +$ with $\ $ a constant skew-symmetric tensor. Indeed, since $\ _\varepsilon$ and $\ '_\varepsilon$ both belong to $H(\ _1, \dots, \ _N; \Omega_\varepsilon)$, then $\ '_\varepsilon = \ _\varepsilon + \nabla \ $ for some $\ \in H^1(\Omega_\varepsilon; \mathbb{R}^2)$, which satisfies the equation

$$\int_{\Omega_\varepsilon} C[\ (\ )] \cdot \ (\ ) \, da = 0 \qquad \text{for all } \ \in H^1(\Omega_\varepsilon; \mathbb{R}^2);$$

choosing $\ = \ $ and using the strong ellipticity of C, this implies that $\ (\ ) = 0$, and, in turn, that $\ (\ ) = \ + \ $ with $\ $ and $\ $ a constant vector and a constant skew-symmetric tensor, respectively, which proves the assertion. ∎

Remark 2.2. Uniqueness of the solution of (2.9) is guaranteed, for instance, by assuming that the total infinitesimal rotation of the body vanishes, i.e.,

$$\int_{\Omega_\varepsilon} (\ _\varepsilon - \ _\varepsilon^\top) \, da = 0. \tag{2.12}$$

3. Existence for a single dislocation in a ball. In this section we consider the special case where

$$\Omega = B_R(\ _0)$$

and we have a single dislocation located at $\ _0$ and with Burger vector $\ $. We are interested in the asymptotic behavior as $\varepsilon \to 0^+$ and $R \to \infty$ of the solutions of the minimization problem

$$\min_{\boldsymbol{H} \in H(\ ;B_R(\ _0)\setminus B_\varepsilon(\ _0))} \int_{B_R(\ _0)\setminus B_\varepsilon(\ _0)} W(\ (\ ))\, da. \tag{3.1}$$

Proposition 3.1. *Assume that the elasticity tensor C satisfies condition* (2.3) *and let* $\ _{,\varepsilon,R}$ *be the unique solution of* (3.1) *such that*

$$\int_{B_R(\ _0)\setminus B_\varepsilon(\ _0)} (\ _{,\varepsilon,R} - \ ^{\top}_{,\varepsilon,R})\, da = 0. \tag{3.2}$$

Then $\{\ _{,\varepsilon,R}\}$ *converges uniformly on compact subsets of* $\mathbb{R}^2 \setminus \{\ _0\}$ *as* $\varepsilon \to 0$ *and* $R \to \infty$ *to the function*[6]

$$(\ ;\ _0) := \frac{1}{2\pi|\ -\ _0|^2}\ \otimes(\ -\ _0)^{\perp} + \nabla\ (\ -\ _0) \tag{3.3}$$

with

$$(\ ) = -\frac{\mu \log|\ |}{2\pi(\lambda+2\mu)}\ ^{\perp} - \frac{\lambda+\mu}{4\pi(\lambda+2\mu)|\ |^2}\left\{(\ \cdot\ ^{\perp})\ + (\ \cdot\ )\ ^{\perp}\right\}, \tag{3.4}$$

which is a solution in the distributional sense of the system

$$\begin{cases} \mathrm{Curl}\ \ = \delta_{\ _0} \\ \mathrm{Div}\, C[\ (\ )] = 0 \end{cases} \quad \text{in } \mathbb{R}^2.$$

Proof. By Proposition 2.1 and Remark 2.2 the functions $\ _{,\varepsilon,R}$ are given by the solutions of (2.9) in $B_R(\ _0) \setminus B_\varepsilon(\ _0)$ satisfying (3.2). In the isotropic case, these are explicitly known [28]:

$$\ _{,\varepsilon,R}(\ ;\ _0) = \ (\ ;\ _0) + \nabla\ _{,\varepsilon,R}(\ -\ _0), \tag{3.5}$$

where

$$\ _{,\varepsilon,R}(\ ) = \frac{(\lambda+\mu)|\ |^2}{2\pi(\lambda+2\mu)(\varepsilon^2+R^2)}\left\{-\ ^{\perp} - \frac{\lambda+3\mu}{2(\lambda+\mu)|\ |^2}\left((\ \cdot\ ^{\perp})\ + (\ \cdot\ )\ ^{\perp}\right)\right\}$$
$$+ \frac{(\lambda+\mu)\,\varepsilon^2 R^2}{2\pi(\lambda+2\mu)(\varepsilon^2+R^2)|\ |^4}\left\{(\ \cdot\ ^{\perp})\ + (\ \cdot\ )\ ^{\perp}\right\}.$$

A straightforward calculation shows that $\ _{,\varepsilon,R}$ satisfies the constraint (3.2) and the Euler equations (2.10). Uniform convergence to $\ $ is immediate.

It is easy to see that $\frac{1}{2\pi|\ -\ _0|^2}\ \otimes(\ -\ _0)^{\perp}$ satisfies

$$\mathrm{Curl}\ \ = \delta_{\ _0} \tag{3.6}$$

[6]Given a vector $\boldsymbol{v}$, we denote by $\boldsymbol{v}^{\perp}$ the vector perpendicular to $\boldsymbol{v}$ obtained by rotating $\boldsymbol{v}$ counterclockwise by $\pi/2$.

in the sense of distributions, and it is clear that all other solutions have the form $\frac{1}{2\pi|\ \ -\ \ _0|^2} \ \otimes (\ \ -\ \ _0)^\perp + \nabla\ \ ,$ a vector field in $W^{1,1}_{\mathrm{loc}}(\mathbb{R}^2;\mathbb{R}^2)$. A straightforward calculation shows that choosing as in (3.4) we obtain that also satisfies

$$\operatorname{Div} C[\ \ (\ \ )] = 0, \tag{3.7}$$

and the proof is complete. ∎

Remark 3.2.

(i) The field may be regarded as the deformation induced by a dislocation with Burgers vector in the whole plane. By introducing polar coordinates (ϱ, ϑ) centered at $_0$, with associated basis $(\ _\varrho,\ _\vartheta)$, we may write

$$= \frac{1}{2\pi\varrho} \ \otimes\ \ _\vartheta + \nabla \tag{3.8}$$

with

$$= -\frac{\mu\log\varrho}{2\pi(\lambda+2\mu)}\ \ ^\perp - \frac{\lambda+\mu}{4\pi(\lambda+2\mu)}\left\{(\ \cdot\ _\vartheta)\ _\varrho + (\ \cdot\ _\varrho)\ _\vartheta\right\}. \tag{3.9}$$

The complete expression for in polar coordinates is

$$\begin{aligned} = \frac{1}{2\pi\varrho(\lambda+2\mu)} [&\mu(\ \cdot\ _\vartheta)\ _\varrho\otimes\ _\varrho + (2\lambda+3\mu)(\ \cdot\ _\varrho)\ _\varrho\otimes\ _\vartheta \\ &-\mu(\ \cdot\ _\varrho)\ _\vartheta\otimes\ _\varrho + \mu(\ \cdot\ _\vartheta)\ _\vartheta\otimes\ _\vartheta], \end{aligned}$$

and the corresponding stress tensor is

$$= \frac{\mu(\lambda+\mu)}{\pi\varrho(\lambda+2\mu)}\{(\ \cdot\ _\vartheta)\ _\varrho\otimes\ _\varrho + (\ \cdot\ _\varrho)(\ _\varrho\otimes\ _\vartheta + \ _\vartheta\otimes\ _\varrho) + (\ \cdot\ _\vartheta)\ _\vartheta\otimes\ _\vartheta\}.$$

Note that is homogeneous of degree -1 in ϱ so that we may write

$$(\varrho,\vartheta;\ _0) = \frac{1}{\varrho}\ \ (\vartheta), \tag{3.10}$$

where is independent of ϱ and $_0$.

(ii) In what follows we shall use extensively the family of tensor fields

$$_{,\varepsilon}(\ ;\ _0) := \ \ (\ ;\ _0) + \nabla\ \ _{,\varepsilon}(\ \ -\ \ _0) \tag{3.11}$$

with

$$_{,\varepsilon}(\ \ ) = \lim_{R\to\infty}\ \ _{,\varepsilon,R}(\ \ ) = \frac{(\lambda+\mu)\,\varepsilon^2}{2\pi(\lambda+2\mu)|\ \ |^4}\left\{(\ \cdot\ ^\perp)\ \ + (\ \cdot\ \ )\ ^\perp\right\}, \tag{3.12}$$

which have the property that $\operatorname{Div} C[\ \ (\ \ _{,\varepsilon})] = 0$ on $\mathbb{R}^2 \setminus B_\varepsilon(\ _0)$, and $C[\ \ (\ \ _{,\varepsilon})]\ \ = 0$ on $\partial B_\varepsilon(\ _0)$. Notice also that $_{,\varepsilon} \to 0$ uniformly on compacta in $\mathbb{R}^2 \setminus \{\ _0\}$.

4. Existence for systems of dislocations in a bounded domain. In this section we study the asymptotic behavior as $\varepsilon \to 0^+$ of the solutions of the minimization problem

$$\min_{\boldsymbol{H} \in H(\ _1, \ldots, \ _N; \Omega_\varepsilon)} \int_{\Omega_\varepsilon} W(\ (\ )) \, da,$$

where, we recall,

$$H(\ _1, \ldots, \ _N; \Omega_\varepsilon) = \left\{ \quad \in H(\operatorname{Curl} 0; \Omega_\varepsilon) \, : \, \int_{\partial B_\varepsilon(\ _i)} \quad ds = \ _i, \quad i = 1, \ldots, N \right\},$$

and where $\ _1, \ldots, \ _N$ and $\ _1, \ldots, \ _N$ are given sets of points in Ω and of Burgers vectors, respectively. The main result of this section is the following theorem.

THEOREM 4.1. *Assume that the elasticity tensor C satisfies condition* (2.3). *Then the minimization problem*

$$\min_{\boldsymbol{H} \in H(\ _1, \ldots, \ _N; \Omega_\varepsilon)} \int_{\Omega_\varepsilon} W(\ (\ )) \, da \tag{4.1}$$

admits a unique solution, modulo an infinitesimal rigid-body motion, $\ _\varepsilon$ which converges as $\varepsilon \to 0$, strongly in $L^2_{\mathrm{loc}}(\Omega \setminus \cup_{i=1}^N \{ \ _i\}; \mathbb{R}^{2\times 2})$, to a solution, in the distributional sense, of the system

$$\begin{cases} \operatorname{Curl} \quad = \sum_{i=1}^N \ _i \, \delta_{\ _i} \\ \operatorname{Div} C[\ (\ )] = 0 \end{cases} \qquad \text{in } \Omega. \tag{4.2}$$

The proof of the previous theorem is divided in several lemmas. We begin by recalling that any tensor field $\ \in H(\ _1, \ldots, \ _N; \Omega_\varepsilon)$ can be written as the sum of a given tensor field in $H(\ _1, \ldots, \ _N; \Omega_\varepsilon)$ and the gradient of a vector field. In particular, we may choose

$$= \sum_{i=1}^N \ _{i,\varepsilon} + \nabla \tag{4.3}$$

with $\ \in H^1(\Omega_\varepsilon; \mathbb{R}^2)$ and (cf. (3.5))

$$\ _{i,\varepsilon}(\ ) := \ _{i,\varepsilon}(\ ; \ _i). \tag{4.4}$$

Notice that $\ _{i,\varepsilon}$ satisfies the Euler equations (2.10) on $\mathbb{R}^2 \setminus B_\varepsilon(\ _i)$, i.e.,

$$\begin{cases} \operatorname{Div} C[\ (\ _{i,\varepsilon})] = 0 & \text{in } \mathbb{R}^2 \setminus B_\varepsilon(\ _i), \\ C[\ (\ _{i,\varepsilon})] \quad = \mathbf{0} & \text{on } \partial B_\varepsilon(\ _i). \end{cases}$$

Also, $\int_{\partial B_\varepsilon(\ _i)} \ _{i,\varepsilon} \ ds = \ _i$.

Inserting (4.3) into the energy functional (2.8) and applying the divergence theorem we obtain

$$J_\varepsilon(\ ) = \sum_{i=1}^N J_\varepsilon(\ _{i,\varepsilon}) + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \int_{\Omega_\varepsilon} C[\ (\ _{i,\varepsilon})] \cdot \ _{j,\varepsilon} \, da + I_\varepsilon(\ ) \tag{4.5}$$

with

$$\text{(4.6)}\quad I_\varepsilon(\ ) := \int_{\Omega_\varepsilon} W(\ (\ ))\, da + \sum_{i=1}^{N} \int_{\partial\Omega} \ \cdot\ _{i,\varepsilon}\ \ ds - \sum_{i=1}^{N}\sum_{j\neq i} \int_{\partial B_\varepsilon(\ _i)} \ \cdot\ _{j,\varepsilon}\ \ ds$$

and where $\ _{i,\varepsilon} := C[\ (\ _{i,\varepsilon})]$.

Hence, granted the decomposition (4.5), for ε fixed, the minimization problem (4.1) is equivalent to the problem

$(M\ _{,\varepsilon})$: minimize the functional I_ε over all displacement fields $\ \in H^1(\Omega_\varepsilon;\mathbb{R}^2)$, i.e., find the solutions of

$$\text{(4.7)}\qquad \min_{\ \in H^1(\Omega_\varepsilon;\mathbb{R}^2)} I_\varepsilon(\ ).$$

In view of (4.3), (4.5), and the invariance of the functional J_ε with respect to infinitesimal rigid-body motions, it is clear that to minimize the functional I_ε over all displacement fields $\ \in H^1(\Omega_\varepsilon;\mathbb{R}^2)$ is equivalent to minimize I_ε over all $\ \in H^1(\Omega_\varepsilon;\mathbb{R}^2)$ such that

$$\text{(4.8)}\qquad \int_B \ \ da = \mathbf{0}, \qquad \int_{\Omega_\varepsilon} (\nabla\ \ - (\nabla\ \ )^\top)\, da = 0$$

for a fixed ball $B \subset \Omega_\varepsilon$. Conditions (4.8) guarantee the coerciveness of the functional I_ε and in turn the existence of minimizers. Indeed we have the following lemma.

LEMMA 4.2. *Assume that the elasticity tensor C satisfies condition* (2.3). *Then there exist two positive constants c_1 and c_2 independent of ε such that*

$$\text{(4.9)}\qquad I_\varepsilon(\ ) \geq c_1 \|\ \|^2_{H^1(\Omega_\varepsilon;\mathbb{R}^2)} - c_2 \|\ \|_{H^1(\Omega_\varepsilon;\mathbb{R}^2)}$$

for every $\ _\varepsilon \in H^1(\Omega_\varepsilon;\mathbb{R}^2)$ satisfying the constraint (4.8). *Moreover, for every ε the minimization problem*

$$\min_{\ \in H^1(\Omega_\varepsilon;\mathbb{R}^2)} I_\varepsilon(\ )$$

admits a unique solution $\ _\varepsilon \in H^1(\Omega_\varepsilon;\mathbb{R}^2)$ satisfying (4.8) *and such that*

$$\text{(4.10)}\qquad \|\ _\varepsilon\|_{H^1(\Omega_\varepsilon;\mathbb{R}^2)} \leq M$$

for some positive constant M independent of ε.

Proof. By the positive definiteness of the elasticity tensor C, for $\ \in H^1(\Omega_\varepsilon;\mathbb{R}^2)$,

$$\text{(4.11)}\qquad I_\varepsilon(\ ) \geq K_0 \int_{\Omega_\varepsilon} |\ (\ )|^2\, da - \sum_{i=1}^{N} \sup_{\partial\Omega} |\ _{i,\varepsilon}| \int_{\partial\Omega} |\ |\, ds$$

$$\text{(4.12)}\qquad - \sum_{i=1}^{N}\sum_{j\neq i} \sup_{\partial B_\varepsilon(\ _i)} |\ _{j,\varepsilon}| \int_{\partial B_\varepsilon(\ _i)} |\ |\, ds.$$

By Korn's inequality (see Proposition A.5) there exists a constant c_3, independent of ε, such that

$$\text{(4.13)}\qquad \int_{\Omega_\varepsilon} |\ (\ _\varepsilon)|^2\, da \geq c_3 \|\ _\varepsilon\|^2_{H^1(\Omega_\varepsilon;\mathbb{R}^2)}.$$

Now, by Proposition A.6

$$(4.14) \qquad \int_{\partial\Omega} |\ _{\varepsilon}|\, ds \leq c_4 \|\ _{\varepsilon}\|_{H^1(\Omega_\varepsilon;\mathbb{R}^2)}, \quad \text{and} \quad \int_{\partial B_\varepsilon(\ _i)} |\ _{\varepsilon}|\, ds \leq c_4 \|\ _{\varepsilon}\|_{H^1(\Omega_\varepsilon;\mathbb{R}^2)}$$

with c_4 independent of ε. Moreover,

$$(4.15) \qquad \sup_{\partial B_\varepsilon(\ _i)} |\ _{j,\varepsilon}| \leq c_5, \qquad j \neq i,$$

with c_5 independent of ε.

Combining (4.12), (4.13), (4.14), and (4.15) yields (4.9). In turn, since the functional I_ε is convex and $I_\varepsilon(\mathbf{0}) = 0$ the remaining of the proof follows immediately. $\square$

We now study the asymptotic behavior of the minimizers $\ _\varepsilon$.

Lemma 4.3. *Assume that the elasticity tensor C satisfies condition* (2.3). *Let $\ _\varepsilon \in H^1(\Omega_\varepsilon;\mathbb{R}^2)$ be the unique solution of*

$$\min_{\ \in H^1(\Omega_\varepsilon;\mathbb{R}^2)} I_\varepsilon(\ )$$

satisfying (4.8). *Then as $\varepsilon \to 0$ the sequence $\{\ _\varepsilon\}$ converges strongly in $H^1_{\mathrm{loc}}(\Omega \setminus \cup_{i=1}^N \{\ _i\};\mathbb{R}^2)$ to a solution $\ _0$ of the minimization problem*

$$(4.16) \qquad \min_{\ \in H^1(\Omega;\mathbb{R}^2)} I_0(\ ).$$

Here

$$(4.17) \qquad I_0(\ ) := \int_\Omega W(\ (\ ))\, da + \sum_{i=1}^N \int_{\partial\Omega} \ \cdot\ _i\ \, ds,$$

where $\ _i := C[\ (\ _i)]$ and

$$(4.18) \qquad \ _i(\ ) := \ _i(\ ;\ _i)$$

is the fundamental solution defined in (3.3). *Moreover,*

$$(4.19) \qquad I_\varepsilon(\ _\varepsilon) \to I_0(\ _0).$$

Proof. By Proposition A.7, we can extend $\ _\varepsilon$ to Ω in such a way that

$$\|\ _\varepsilon\|_{H^1(\Omega;\mathbb{R}^2)} \leq cM,$$

where M is the constant given by (4.10). Hence there exists a subsequence of $\{\ _\varepsilon\}$ not relabeled, such that

$$\ _\varepsilon \rightharpoonup\ _0 \qquad \text{in } H^1(\Omega;\mathbb{R}^2)$$

for some $\ _0 \in H^1(\Omega;\mathbb{R}^2)$. By Hölder's inequality,

(4.20)

$$\left| \int_{\partial B_\varepsilon(\ _i)} \ _\varepsilon \cdot\ _{j,\varepsilon}\ \, ds \right|^2 \leq \int_{\partial B_\varepsilon(\ _i)} |\ _\varepsilon|^2 ds \int_{\partial B_\varepsilon(\ _i)} |\ _{j,\varepsilon}|^2 ds \leq \varepsilon c \sup_{B_\varepsilon(\ _i)} |\ _{j,\varepsilon}|^2 M^2,$$

which vanishes as $\varepsilon \to 0$, and where we have used Proposition A.6, (4.10) and the fact that ${}_{j,\varepsilon} \to {}_{j}$ uniformly on $B_\varepsilon(\ {}_i)$. Hence

$$\lim_{\varepsilon\to 0}\int_{\partial B_\varepsilon(\ {}_i)} {}_\varepsilon \cdot {}_{j,\varepsilon}\ ds = 0. \tag{4.21}$$

Fix now $\varepsilon_0 > 0$. For $\varepsilon < \varepsilon_0$, by (4.6),

$$I_\varepsilon(\ {}_\varepsilon) \geq \int_{\Omega_{\varepsilon_0}} W(\ (\ {}_\varepsilon))\, da + \sum_{i=1}^{N}\int_{\partial\Omega} {}_\varepsilon \cdot {}_{i,\varepsilon}\ ds - \sum_{i=1}^{N}\sum_{j\neq i}\int_{\partial B_\varepsilon(\ {}_i)} {}_\varepsilon \cdot {}_{j,\varepsilon}\ ds.$$

Letting $\varepsilon \to 0^+$, by standard lower semicontinuity results and (4.21), we obtain that

$$\liminf_{\varepsilon\to 0^+} I_\varepsilon(\ {}_\varepsilon) \geq \int_{\Omega_{\varepsilon_0}} W(\ (\ {}_0))\, da + \sum_{i=1}^{N}\int_{\partial\Omega} {}_0 \cdot {}_i\ ds.$$

Letting $\varepsilon_0 \to 0^+$ we get

$$\liminf_{\varepsilon\to 0^+} I_\varepsilon(\ {}_\varepsilon) \geq I_0(\ {}_0).$$

On the other hand, since

$$I_\varepsilon(\ {}_\varepsilon) \leq I_\varepsilon(\ {}_0),$$

we also have that

$$\limsup_{\varepsilon\to 0^+} I_\varepsilon(\ {}_\varepsilon) \leq I_0(\ {}_0).$$

Hence

$$\lim_{\varepsilon\to 0^+} I_\varepsilon(\ {}_\varepsilon) = I_0(\ {}_0). \tag{4.22}$$

To prove strong convergence, notice that (4.19) implies that

$$\lim_{\varepsilon\to 0^+}\int_{\Omega_\varepsilon} W(\ (\ {}_\varepsilon))\, da = \int_{\Omega} W(\ (\ {}_0))\, da,$$

from which we conclude that, as in Evans [14],

$$\lim_{\varepsilon\to 0^+}\int_{\Omega_\varepsilon} |\ (\ {}_\varepsilon) - \ (\ )|^2\, da = 0,$$

and strong convergence of ${}_\varepsilon$ in $H^1_{\mathrm{loc}}(\Omega;\mathbb{R}^2)$ follows from Korn's inequality.

Now we claim that ${}_0$ minimizes I_0. Indeed for any $\ \in H^1(\Omega;\mathbb{R}^2)$ we have that

$$I_\varepsilon(\ ) \geq I_\varepsilon(\ {}_\varepsilon),$$

so that letting $\varepsilon \to 0^+$ and using (4.19),

$$I_0(\ ) \geq I_0(\ {}_0).$$

Next we claim that ${}_0$ satisfies

$$\int_B {}_0 da = \mathbf{0}, \qquad \int_\Omega (\nabla {}_0 - (\nabla {}_0)^\top)\, da = 0, \tag{4.23}$$

The first constraint follows immediately from (4.8), since $\{ {}_\varepsilon\}$ converges strongly in $H^1_{\rm loc}(\Omega \setminus \cup_{i=1}^N \{ {}_i\}; \mathbb{R}^2)$ to ${}_0$. To prove the second constraint, let ${}_\varepsilon := \frac{1}{2}(\nabla {}_\varepsilon - (\nabla {}_\varepsilon)^\top)$ and ${}_0 = \frac{1}{2}(\nabla {}_0 - (\nabla {}_0)^\top)$, and notice that ${}_\varepsilon \to {}_0$ strongly in $L^2_{\rm loc}(\Omega \setminus \cup_{i=1}^N \{ {}_i\}; \mathbb{R}^{2\times 2})$. Now, denote by $\tilde{}_\varepsilon$ the extension of ${}_\varepsilon$ to zero on Ω. Then, since $\|\nabla {}_\varepsilon\|_{L^2(\Omega_\varepsilon;\mathbb{R}^{2\times 2})}$ is bounded independently of ε, it follows that the sequence $\|\tilde{}_\varepsilon\|_{L^2(\Omega;\mathbb{R}^{2\times 2})}$ is bounded, so that $\tilde{}_\varepsilon \rightharpoonup \tilde{}_0$ in $L^2(\Omega;\mathbb{R}^{2\times 2})$ for some $\tilde{}_0 \in L^2(\Omega;\mathbb{R}^{2\times 2})$. Hence, also $\tilde{}_\varepsilon \rightharpoonup \tilde{}_0$ weakly in $L^2_{\rm loc}(\Omega \setminus \{ {}_0\}; \mathbb{R}^{2\times 2})$, and $\tilde{}_0 = {}_0$. By weak convergence,

$$0 = \int_{\Omega_\varepsilon} {}_\varepsilon da = \int_\Omega \tilde{}_\varepsilon da \to \int_\Omega {}_0 da,$$

and the claim follows. Since the minimization problem

$$\min_{\in H^1(\Omega;\mathbb{R}^2)} I_0(\ ),$$

admits a unique solution modulo an infinitesimal rigid-body displacement, we conclude that *all* sequences ${}_\varepsilon$ converge strongly to ${}_0$. □

We are now ready to conclude the proof of Theorem 4.1.

Proof of Theorem 4.1. Let ${}_\varepsilon \in H^1(\Omega_\varepsilon;\mathbb{R}^2)$ be the unique solution of

$$\min_{\in H^1(\Omega_\varepsilon;\mathbb{R}^2)} I_\varepsilon(\ )$$

satisfying (4.8). It suffices to define

$${}_\varepsilon := \sum_{i=1}^N {}_{i,\varepsilon} + \nabla {}_\varepsilon.$$

Since, by Proposition 3.1, ${}_{i,\varepsilon} \to {}_i$ uniformly on compact subsets of $\mathbb{R}^2 \setminus \{ {}_i\}$ the proof is concluded. □

5. The renormalized energy. In this section we prove a sharp estimate for the minimum energy

$$\min_{\boldsymbol{H}\in H({}_1,\dots, {}_N;\Omega_\varepsilon)} \int_{\Omega_\varepsilon} W(\ (\ ))\, da,$$

as the core radius $\varepsilon \to 0$, and compute the renormalized energy which, being a function of the defect position only, allows to study the equilibrium configurations of the defects and the force acting on them.

Let Ω a bounded domain with the cone property as before, let $S = \{ {}_1, \dots, {}_N\}$ be a system of dislocations in Ω, and let

$$\bar{R} := \frac{1}{4}\min\{|\ -\ | : \ \neq\ , (\ ,\ ) \in S \times (S \cup \partial\Omega)\}. \tag{5.1}$$

The main result of this section is the following theorem.

THEOREM 5.1. *Assume that the elasticity tensor C satisfies condition* (2.3)*. Let* $_{\varepsilon} \in H(\ _1, \ldots, \ _N; \Omega_\varepsilon)$ *be a solution of*

$$\min_{\boldsymbol{H} \in H(\ _1, \ldots, \ _N; \Omega_\varepsilon)} \int_{\Omega_\varepsilon} W(\ (\ ))\, da$$

for a system of dislocations with Burgers vectors $_i$*. Then*

$$\int_{\Omega_\varepsilon} W(\ (\ _\varepsilon))\, da = \sum_{i=1}^{N} \frac{\mu(\lambda+\mu)}{4\pi(\lambda+2\mu)} |\ _i|^2 \ln \frac{1}{\varepsilon} + F(\ _1, \ldots, \ _N) + c + O(\varepsilon), \tag{5.2}$$

where

$$F(\ _1, \ldots, \ _N) = F_{\text{self}}(\ _1, \ldots, \ _N) + F_{\text{int}}(\ _1, \ldots, \ _N) + F_{\text{elastic}}(\ _1, \ldots, \ _N) \tag{5.3}$$

is the renormalized energy with

(5.4)

$$\begin{cases} F_{\text{self}}(\ _1, \ldots, \ _N) = \displaystyle\sum_{i=1}^{N} \int_{\Omega \setminus B_R(\ _i)} W(\ (\ _i))\, da + \sum_{i=1}^{N} \frac{\mu(\lambda+\mu)}{4\pi(\lambda+2\mu)} |\ _i|^2 \ln R, \\ F_{\text{int}}(\ _1, \ldots, \ _N) = \displaystyle\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \int_{\Omega} C[\ (\ _i)] \cdot \ (\ _j)\, da, \\ F_{\text{elastic}}(\ _1, \ldots, \ _N) = \displaystyle\int_{\Omega} W(\ (\ _0))\, da + \sum_{i=1}^{N} \int_{\partial\Omega} \ _0 \cdot \ _i \ \ ds, \end{cases}$$

where $_i(\ ) := \ _i(\ ;\ _i)$ *is the fundamental solution defined in* (3.3)*, the function* $_0$ *is defined in Lemma* 4.3*, c is a constant independent of* $_1, \ldots, \ _N$*, and* $0 < R < \bar{R}$ *is arbitrary. Moreover,* F_{self} *is independent of R.*

Proof. Consider the fundamental solution $_i(\ ) := \ _i(\ ;\ _i)$ defined in (3.3): by (3.10) we may write

$$_i(\varrho_i, \vartheta_i) = \frac{1}{\varrho_i} \ _i(\vartheta_i), \tag{5.5}$$

where (ϱ_i, ϑ_i) are polar coordinates centered at $_i$, and $_i(\vartheta_i)$ is independent of ϱ_i and is defined by (3.10), with replaced by $_i$. A straightforward computation using (3.8) and (3.9) yields

$$a_i := \int_0^{2\pi} W(\ (\ _i(\vartheta)))\, d\vartheta = \frac{\mu(\lambda+\mu)}{4\pi(\lambda+2\mu)} |\ _i|^2. \tag{5.6}$$

By (4.5) we can write the minimum energy in the form

$$J_\varepsilon(\ _\varepsilon) = I_\varepsilon(\ _\varepsilon) + \sum_{i=1}^{N} J_\varepsilon(\ _{i,\varepsilon}) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \int_{\Omega_\varepsilon} C[\ (\ _{i,\varepsilon})] \cdot \ _{j,\varepsilon}\, da,. \tag{5.7}$$

where $I_\varepsilon(\ _\varepsilon)$ is the functional defined by (4.6).

Notice first that the representation of the elastic contribution F_{elastic} in $(5.4)_3$ follows immediately from (4.19).

We now compute the self-energy contribution F_{self}: fix $R < \bar{R}$ and write

$$J_\varepsilon(\ _{i,\varepsilon}) = \int_{\Omega_\varepsilon \setminus B_R(\ _i)} W(\ (\ _{i,\varepsilon}))\, da + \int_{C_{\varepsilon,R}(\ _i)} W(\ (\ _{i,\varepsilon}))\, da \tag{5.8}$$

with $C_{\varepsilon,R}(\ _i) = B_R(\ _i) \setminus B_\varepsilon(\ _i)$.

Now, as $\varepsilon \to 0$, by uniform-on-compacta convergence of $\ _{i,\varepsilon}$ on $\mathbb{R}^2 \setminus \{\ _i\}$,

$$\int_{\Omega_\varepsilon \setminus B_R(\ _i)} W(\ (\ _{i,\varepsilon}))\, da \to \int_{\Omega \setminus B_R(\ _i)} W(\ (\ _i))\, da. \tag{5.9}$$

Moreover, writing as in (3.11) $\ _{i,\varepsilon} = \ _i + \nabla\ _\varepsilon$, with $\ _\varepsilon \to 0$ uniformly on compacta in $\mathbb{R}^2 \setminus \{\ _i\}$, we have

$$\begin{aligned}\int_{C_{\varepsilon,R}(\ _i)} W(\ (\ _{i,\varepsilon}))\, da = \int_{C_{\varepsilon,R}(\ _i)} W(\ (\ _i))\, da + \int_{C_{\varepsilon,R}(\ _i)} C[\ (\ _i)] \cdot \nabla\ _\varepsilon\, da \\ + \int_{C_{\varepsilon,R}(\ _i)} W(\ (\ _\varepsilon))\, da,\end{aligned}$$

and, by (5.5), the first integral on the right-hand side of this identity gives

$$\int_{C_{\varepsilon,R}(\ _i)} W(\ (\ _i))\, da = a_i(\ln R - \ln \varepsilon), \tag{5.10}$$

while the second and third integral may be written as

$$\begin{aligned}&\int_{\partial B_R(\ _i)} \ _\varepsilon \cdot C\left[\ (\ _i) + \frac{1}{2}\ (\ _\varepsilon)\right]\ ds - \int_{\partial B_\varepsilon(\ _i)} \ _\varepsilon \cdot C\left[\ (\ _i) + \frac{1}{2}\ (\ _\varepsilon)\right]\ ds \\ &= \int_{\partial B_R(\ _i)} \ _\varepsilon \cdot C\left[\ (\ _i) + \frac{1}{2}\ (\ _\varepsilon)\right]\ ds - \frac{1}{2}\int_{\partial B_\varepsilon(\ _i)} \ _\varepsilon \cdot C[\ (\ _i)]\ ds \end{aligned} \tag{5.11}$$

where we have used the fact that $C[\ (\ _\varepsilon)]\ = -C[\ (\ _i)]\ $ on $\partial B_\varepsilon(\ _i)$, since $C[\ (\ _{i,\varepsilon})]\ = \mathbf{0}$ on $\partial B_\varepsilon(\ _i)$. The first integral on the right-hand side of the above expression vanishes as $\varepsilon \to 0$, while by (3.12) we may write

$$\ _\varepsilon(\varrho_i, \vartheta_i) = \frac{\varepsilon^2}{\varrho_i^2}\ ^-(\vartheta_i),$$

which, in conjunction with (5.5), shows that

$$-\frac{1}{2}\int_{\partial B_\varepsilon(\ _i)} \ _\varepsilon \cdot C[\ (\ _i)]\ ds = -\frac{1}{2}\int_0^{2\pi} \ ^-(\vartheta) \cdot C[\ (\ _i(\vartheta))]\ d\vartheta = c$$

with c a constant independent of $(\ _1, \dots, \ _N)$. To summarize, (5.11) converges, as $\varepsilon \to 0$, to a constant c independent of R and $\ _i$. Note that this is the constant which appears in (5.2).

Notice that F_{self} is independent of R, since, for $R' < \bar{R}$, say, $R' < R$,

$$\begin{aligned}&\int_{\Omega \setminus B_{R'}(\ _i)} W(\ (\ _i))\, da + a_i \ln R' \\ &\quad = \int_{\Omega \setminus B_R(\ _i)} W(\ (\ _i))\, da + \int_{C_{R',R}(\ _i)} W(\ (\ _i))\, da + a_i \ln R'\end{aligned}$$

$$= \int_{\Omega \setminus B_R(\;_i)} W(\;(\;_i))\, da + a_i \ln \frac{R}{R'} + a_i \ln R'$$
$$= \int_{\Omega \setminus B_R(\;_i)} W(\;(\;_i))\, da + a_i \ln R.$$

We finally compute the contribution of the interaction term F_{int} to the renormalized energy and prove that

$$\int_{\Omega_\varepsilon} C[\;(\;_{i,\varepsilon})] \cdot \;(\;_{j,\varepsilon})\, da = \int_\Omega C[\;(\;_i)] \cdot \;(\;_j)\, da + O(\varepsilon). \tag{5.12}$$

To see this, let as before $\;_{i,\varepsilon} = \;_i + \nabla \;_{i,\varepsilon}$ and $\;_{j,\varepsilon} = \;_j + \nabla \;_{j,\varepsilon}$ (cf. (3.11)), so that

$$\int_{\Omega_\varepsilon} C[\;(\;_{i,\varepsilon})] \cdot \;(\;_{j,\varepsilon})\, da = \int_{\Omega_\varepsilon} C[\;(\;_i)] \cdot \;(\;_j)\, da + \int_{\Omega_\varepsilon} C[\;(\;_i)] \cdot \nabla \;_{j,\varepsilon}\, da$$
$$+ \int_{\Omega_\varepsilon} C[\;(\;_j)] \cdot \nabla \;_{i,\varepsilon}\, da + \int_{\Omega_\varepsilon} C[\;(\;_{i,\varepsilon})] \cdot \nabla \;_{j,\varepsilon}\, da.$$

It can be easily proved that

$$\int_{\Omega_\varepsilon} C[\;(\;_i)] \cdot \;(\;_j)\, da \to \int_\Omega C[\;(\;_i)] \cdot \;(\;_j)\, da,$$

while, applying the divergence theorem, the last three integrals become

$$\int_{\partial\Omega} (\;_{j,\varepsilon} \cdot C[\;(\;_i)]\; + \;_{i,\varepsilon} \cdot C[\;(\;_j)]\; + \;_{j,\varepsilon} \cdot C[\;(\;_{i,\varepsilon})]\;)\, da$$
$$-\sum_{k=1}^N \int_{\partial B_\varepsilon(\;_k)} (\;_{j,\varepsilon} \cdot C[\;(\;_i)]\; + \;_{i,\varepsilon} \cdot C[\;(\;_j)]\; + \;_{j,\varepsilon} \cdot C[\;(\;_{i,\varepsilon})]\;)\, da. \tag{5.13}$$

Now, recall that $\;_{i,\varepsilon} \to 0$ uniformly on compacta in $\mathbb{R}^2 \setminus \{\;_i\}$ (see Remark 3.2(ii)), so that the integrals over $\partial\Omega$ and $\partial B_\varepsilon(\;_k)$ with $k \neq i, j$ vanish in the limit as $\varepsilon \to 0$, and (5.13) becomes

$$-\int_{\partial B_\varepsilon(\;_i)} (\;_{j,\varepsilon} \cdot C[\;(\;_i)]\; + \;_{i,\varepsilon} \cdot C[\;(\;_j)]\; + \;_{j,\varepsilon} \cdot C[\;(\;_{i,\varepsilon})]\;)\, da$$
$$-\int_{\partial B_\varepsilon(\;_j)} (\;_{j,\varepsilon} \cdot C[\;(\;_i)]\; + \;_{i,\varepsilon} \cdot C[\;(\;_j)]\; + \;_{j,\varepsilon} \cdot C[\;(\;_{i,\varepsilon})]\;)\, da, \tag{5.14}$$

which tends to 0 as $\varepsilon \to 0$. Consider in fact the first term: then

$$\left| \int_{\partial B_\varepsilon(\;_i)} \;_{j,\varepsilon} \cdot C[\;(\;_i)]\;\, da \right| \leq C \sup_{B_\varepsilon(\;_i)} |\;_{j,\varepsilon}| \int_{\partial B_\varepsilon(\;_i)} |C[\;(\;_i)]\;|\, da$$
$$\leq \sup_{B_\varepsilon(\;_i)} |\;_{j,\varepsilon}| \int_{\partial B_\varepsilon(\;_i)} |\;_i|\, da \to 0$$

by uniform convergence of $\;_{j,\varepsilon}$. Consider now the second term in (5.14): as before, we may write

$$\;_{i,\varepsilon}(\varrho_i, \vartheta_i) = \frac{\varepsilon^2}{\varrho_i^2}\;\bar{\;}_i(\vartheta_i)$$

so that, since $_j$ is continuous at $_i$,

$$\left|\int_{\partial B_\varepsilon(\ _i)} \ _{i,\varepsilon}\cdot C[\ (\ _j)]\ \ ds\right| \le \sup_{B_\varepsilon(\ _i)} |C[\ (\ _j)]| \int_{\partial B_\varepsilon(\ _i)} |\ ^-{}_i|\, ds \to 0$$

as $\varepsilon \to 0$, since $\ ^-{}_i$ is bounded. The remaining terms in (5.14) can be treated analogously, and this completes the proof of (5.2). ◻

PROPOSITION 5.2. *The interaction energy F_{int} in $(5.4)_2$ diverges logarithmically with the relative distance between the defects:*

$$F_{\text{int}}(\ _1,\dots,\ _N) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{\mu(\lambda+\mu)}{\pi(\lambda+2\mu)}\ _i\cdot\ _j \ln\frac{1}{|\ _i-\ _j|} + O(1), \tag{5.15}$$

as $|\ _i - \ _j| \to 0$.

Proof. Recall that $_i \in L^1(\Omega;\mathbb{R}^{2\times 2})$ for each i, and

$$F_{\text{int}}(\ _1,\dots,\ _N) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\int_\Omega C[\ (\ _i)]\cdot\ (\ _j)\, da,$$

let $_i,\ _j \in \Omega$, and γ a line segment parallel to $_j -\ _i$ connecting $_j$ to $\partial\Omega$ (cf. Figure 5.1), so that $\gamma = \{\ \in\Omega :\ =\ _j + s(\ _j -\ _i)\}$, with $s\in[0,\bar{s}]$. Moreover, let $\ = (\frac{\ _j-\ _i}{|\ _j-\ _i|})^\perp$ be the unit vector orthogonal to γ. Consider the tensor field $_j$: while $\Omega\setminus\{\ _j\}$ is not simply connected, $\Omega\setminus\gamma$ is, so that there exists a field $_j$ on $\Omega\setminus\gamma$ such that $_j = \nabla\ _j$, and

$$[\![\ _j]\!] = -\ _j,$$

where $[\![\ _j]\!]$ is the jump of $_j$ across γ, defined by

$$[\![\ _j]\!](\ ) := \lim_{\ \to\ ,\ \cdot\ >0}\ _j(\ ) - \lim_{\ \to\ ,\ \cdot\ <0}\ _j(\ )$$

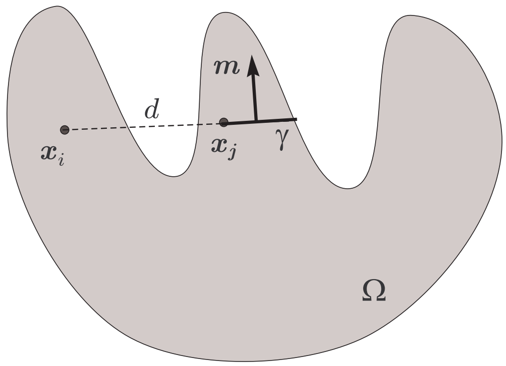


FIG. 5.1. *Connecting cut γ.*

for $\in \gamma$. Applying the divergence theorem to $\Omega\backslash\gamma$, and noting that $\partial(\Omega\backslash\gamma) = \partial\Omega\cup\gamma$, we find

$$\int_{\Omega} C[\ (\ _i)] \cdot \ _j \, da = \int_{\Omega\backslash\gamma} C[\ (\ _i)] \cdot \nabla \ _j \, da$$
$$= \int_{\partial\Omega} \ _j \cdot C[\ (\ _i)] \ \ ds - \int_{\gamma} [\![\ _j]\!] \cdot C[\ (\ _i)] \ \ ds.$$

The first integral in the above expression remains bounded as $_i - \ _j \to \mathbf{0}$, since $_i(\ ) = \ _i(\ ;\ _i) \to \ _i(\ ;\ _j)$ uniformly on $\partial\Omega$ as $_j - \ _j \to \mathbf{0}$. As to the second integral, write $_i = \frac{1}{\varrho_i} \ _i(\vartheta_i)$, and choose $s = \varrho_i - d$, $\vartheta_i = \bar{\vartheta}$ on γ with $d = |\ _j - \ _i|$: then

$$-\int_{\gamma} [\![\ _j]\!] \cdot C[\ (\ _i)] \ \ ds = \int_0^{\bar{s}} \ _j \cdot C[\ (\ _i(\bar{\vartheta}))] \ \ \frac{1}{d+s} \, ds$$
$$= \ _j \cdot C[\ (\ _i(\bar{\vartheta}))] \ \ \left(\ln \frac{1}{d} + \ln(d + \bar{s})\right),$$

which proves (5.15) since by a straightforward computation using (3.8) and (3.9) we have

$$_j \cdot C[\ (\ _i(\bar{\vartheta}))] \ \ = \frac{\mu(\lambda+\mu)}{\pi(\lambda+2\mu)} \ _i \cdot \ _j. \qquad \square$$

6. The force on a dislocation. We prove in this section that the derivative of the renormalized energy with respect to defect position coincides with the resultant of the Eshelby stress

$$= W(\ (\ ))\mathbf{1} - \ ^{\top} C[\ (\ )] \tag{6.1}$$

on the dislocation.[7]

To highlight the dependence of the minimizers on the location of the dislocations, we write

$$_0(\ ;\ _1, \ldots, \ _N)$$

[7]Here $\mathbf{1}$ is the identity tensor.

for a minimizer of I_0 relative to a system of dislocations located at $(\ _1, \dots, \ _N)$ in Ω, and

$$(6.2) \qquad \ _0 = \ _0(\ ; \ _1, \dots, \ _N) := \sum_{i=1}^{N} \ _i(\ ; \ _i) + \nabla \ _0(\ ; \ _1, \dots, \ _N)$$

for the corresponding solution of (4.2) as in Theorem 2.1. Let also $\ \in \mathbb{R}^2$ be a fixed vector and $t \in I \subset \mathbb{R}$ a real parameter.

LEMMA 6.1. *The field* $\ _i(\ ; \ _i)$, $\ _0(\ ; \ _1, \dots, \ _N)$ *and* $\ _0(\ ; \ _1, \dots, \ _N)$ *are smooth with respect to* $\ _i$ *for* $i \in \{1, \dots, N\}$.

In particular,

$$(6.3) \qquad \dot{\ }_i(\ ) := \frac{d}{dt} \ _i(\ ; \ _i + t \ )\Big|_{t=0} = -\nabla (\ _i(\ ; \ _i) \ ) = -\nabla (\ _i(\ ) \ ) .$$

Moreover, if for a fixed $k \in \{1, \dots, N\}$ *we denote by* $\dot{\ }_0$ *and* $\dot{\ }_0$ *the smooth fields such that*

$$\dot{\ }_0(\ ) := \frac{d}{dt} \ _0(\ ; \ _1, \dots, \ _k + t \ , \dots, \ _N)\Big|_{t=0},$$

$$\dot{\ }_0(\ ) := \frac{d}{dt} \ _0(\ ; \ _1, \dots, \ _k + t \ , \dots, \ _N)\Big|_{t=0},$$

then

$$(6.4) \qquad \dot{\ }_0 = \nabla \qquad \text{with} \qquad = \dot{\ }_0 - \ _k \ .$$

Proof. Smoothness of $\ _i$ follows upon recalling that $\ _i(\ ; \ _i)$ is a smooth function of $\ - \ _i$ (cf. (3.3)). This implies that $\ _i(\ ; \ _i + t \ ) = \ _i(\ - t \ ; \ _i) = \ _i(\ - t \ )$, so that

$$\dot{\ }_i(\ ) = -\nabla(\ _i(\ )) \ ,$$

where $\nabla(\ _i)$ is the tensor field defined by the identity $[\nabla(\ _i) \ ] = [\nabla(\ _i \ )]$ for any constant vector $\ \in \mathbb{R}^2$. Since Curl $\ _i = 0$ in $\mathbb{R}^2 \setminus \{ \ _i\}$,

$$\nabla(\ _i) \ = \nabla(\ _i \ ),$$

which yields (6.3).

Now, since $\ _0$ minimizes the functional I_0, and satisfies the corresponding Euler equations, it may be written in the form

$$(6.5) \qquad \ _0(\ ; \ _1, \dots, \ _N) = \int_{\partial\Omega} (\ , \ ) \left(\sum_{i=1}^{N} \ _i(\ ; \ _1, \dots, \ _N) \ (\ ) \right) da \ ,$$

modulo an infinitesimal rigid body motion, where $(\ , \ )$ is the Green's function for the Neumann problem in plane elasticity. Since

$$\sum_{i=1}^{N} \ _i(\ ; \ _1, \dots, \ _N) = \sum_{i=1}^{N} C[\ (\ _i(\ ; \ _i))],$$

the smoothness of $\ _0$ follows from the smoothness of $\ _i$.

Finally, the smoothness of $\ _0$ follows from (6.2), and (6.4) can be directly verified. ∎

LEMMA 6.2. *Let $f = f(\ ,t)$, $g = g(\ ,t)$, $r = r(\ ,t)$ be smooth functions defined on $B_R(\ _0+t\ )$, $\partial B_R(\ _0+t\ )$ and $\Omega \setminus B_R(\ _0+t\ )$ for $t \in I \subset \mathbb{R}$, respectively, with $R > 0$ and $\ $ a constant vector. Then*

$$\frac{d}{dt}\int_{B_R(\ _0+t\ )} f(\ ,t)\,da\bigg|_{t=0} = \int_{B_R(\ _0)} D_t f(\ ,0)\,da$$

$$= \int_{B_R(\ _0)} \partial_t f(\ ,0)\,da + \int_{\partial B_R(\ _0)} f(\ ,0)\ \ \cdot\ \ ds, \tag{6.6}$$

and

$$\frac{d}{dt}\int_{\partial B_R(\ _0+t\ )} g(\ ,t)\,ds\bigg|_{t=0} = \int_{\partial B_R(\ _0)} D_t g(\ ,0)\,ds. \tag{6.7}$$

Moreover,

$$\frac{d}{dt}\int_{\Omega\setminus B_R(\ _0+t\ )} r(\ ,t)\,da\bigg|_{t=0} = \int_{\Omega\setminus B_R(\ _0)} \partial_t r(\ ,0)\,da - \int_{\partial B_R(\ _0)} r(\ ,0)\ \ \cdot\ \ ds, \tag{6.8}$$

where

$$D_t f = \partial_t f + \nabla f \cdot\ \ . \tag{6.9}$$

Proof. The first identity follows upon applying the classical theorem of derivation under the integral

$$\frac{d}{dt}\int_{B_R(\ _0+t\ )} f(\ ,t)\,da = \frac{d}{dt}\int_{B_R(\ _0)} f(\ \ +t\ \ ,t)\,da$$

$$= \int_{B_R(\ _0)} (\partial_t f(\ \ +t\ \ ,t) + \ \ \cdot \nabla f(\ \ +t\ \ ,t))\,da;$$

letting $t = 0$ and using the divergence theorem we obtain (6.6). Relation (6.7) follows from a similar argument. To prove (6.8), denote by $\hat{r}(\ ,t)$ a smooth extension of $r(\ ,t)$ to Ω for all $t \in I$. Then

$$\begin{aligned}
\frac{d}{dt}\int_{\Omega\setminus B_R(\ _0+t\ )} r(\ ,t)\,da &= \frac{d}{dt}\int_{\Omega} \hat{r}(\ ,t)\,da - \frac{d}{dt}\int_{B_R(\ _0+t\ )} \hat{r}(\ ,t)\,da \\
&= \int_{\Omega} \partial_t \hat{r}(\ ,t)\,da - \int_{B_R(\ _0+t\ )} \partial_t \hat{r}(\ ,t)\,da \\
&\quad - \int_{\partial B_R(\ _0+t\ )} \hat{r}(\ ,t)\ \ \cdot\ \ ds \\
&= \int_{\Omega\setminus B_R(\ _0+t\ )} \partial_t \hat{r}(\ ,t)\,da - \int_{\partial B_R(\ _0+t\ )} \hat{r}(\ ,t)\ \ \cdot\ \ ds,
\end{aligned}$$

which proves the assertion. ∎

The next theorem, one of our main results, shows that the derivative of the renormalized energy coincides with the force on a dislocation and thus, as mentioned in the introduction, it may be viewed as the generalization of Eshelby's notion of force on a defect, when bad singularities are associated to the defect itself.

THEOREM 6.3. *Let* $\;_0$ *be defined by* (6.2), *and let* $F = F(\;_1, \ldots, \;_N)$ *be the renormalized energy* (5.3)*: then*

$$\nabla\;_k F = -\int_{\partial B_R(\;_k)} \left\{ W(\;(\;_0))\mathbf{1} - \;_0^\top C[\;(\;_0)] \right\} \; ds \tag{6.10}$$

for $R < \bar{R}$, *where* $\bar{R}$ *is defined in* (5.1).

Proof. The first step is to rewrite the renormalized energy as the sum of two contributions,

$$F(\;_1, \ldots, \;_N) = F_R(\;_1, \ldots, \;_N) + G(\;_1, \ldots, \;_N), \tag{6.11}$$

where we have omitted the constant logarithmic term and

$$F_R(\;_1, \ldots, \;_N) = \int_{\Omega_R} W(\;(\;_0))\, da, \tag{6.12}$$

and

$$\begin{aligned} G(\;_1, \ldots, \;_N) &= \sum_{i=1}^{N} \sum_{h \neq i} \int_{B_R(\;_h)} W(\;(\;_i))\, da + \sum_{h=1}^{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \int_{B_R(\;_h)} C[\;(\;_i)] \cdot \;_j\, da \\ &\quad + \sum_{h=1}^{N} \int_{B_R(\;_h)} W(\;(\;_0))\, da + \sum_{h=1}^{N} \sum_{i=1}^{N} \int_{\partial B_R(\;_h)} \;_0 \cdot C[\;(\;_i)] \; da. \end{aligned} \tag{6.13}$$

Consider now a variation of the position of the kth dislocation of the form $\;_k \to \;_k + t$ with a fixed vector. Then

$$\begin{aligned} &\frac{d}{dt} F_R(\;_1, \ldots, \;_k + t\;, \ldots, \;_N)\bigg|_{t=0} \\ &= \frac{d}{dt} \int_{\Omega_R \setminus (B_R(\;_1) \cup \cdots \cup B_R(\;_k + t\;) \cup \cdots \cup B_R(\;_N))} W(\;(\;_0(\;;\;_1, \ldots, \;_k + t\;, \ldots, \;_N)))\, da \bigg|_{t=0} \end{aligned}$$

and applying Lemma 6.2 we obtain

$$\int_{\Omega_R} C[\;(\;_0)] \cdot \dot{\;}_0\, da - \int_{\partial B_R(\;_k)} W(\;(\;_0)) \;\cdot\; ds.$$

Recalling now Lemma 6.1, writing $\dot{\;}_0 = \nabla\;$, applying the divergence theorem, and recalling that $C[\;(\;_0)]\; = 0$ on $\partial\Omega$ this becomes

$$\begin{aligned} &\int_{\Omega_R} C[\;(\;_0)] \cdot \nabla\; \; da - \int_{\partial B_R(\;_k)} W(\;(\;_0)) \;\cdot\; ds \\ &\qquad = -\sum_{i=1}^{N} \int_{\partial B_R(\;_i)} \;\cdot C[\;(\;_0)]\; ds - \int_{\partial B_R(\;_k)} W(\;(\;_0)) \;\cdot\; ds, \end{aligned}$$

which in turn equals

$$-\int_{\partial B_R(\ _k)} \cdot (W(\ (\ _0))\mathbf{1} - \ _0^{\top} C[\ (\ _0)])\ \ ds$$
$$-\int_{\partial B_R(\ _k)} (\ +\ \ _0\ ) \cdot C[\ (\ _0)]\ \ ds - \sum_{i\neq k} \int_{\partial B_R(\ _i)} \cdot C[\ (\ _0)]\ \ ds.$$

The first term is the resultant of the Eshelby stress on the kth dislocation, while the second and third term may be written as

$$-\int_{\partial B_R(\ _k)} (\ \dot{}_0 + (\nabla\ _0)\ \ + \sum_{i\neq k}\ _i\ ) \cdot C[\ (\ _0)]\ \ ds$$
$$-\sum_{i\neq k} \int_{\partial B_R(\ _i)} (\ \dot{}_0 - \ _k\ ) \cdot C[\ (\ _0)]\ \ ds \tag{6.14}$$

since $=\ \dot{}_0 - \ _k$ and

$$+\ \ _0\ = \ \dot{}_0 - \ _k\ + \left(\sum_i\ _i + \nabla\ _0\right)\ = \ \dot{}_0 + (\nabla\ _0)\ + \sum_{i\neq k}\ _i\ .$$

We now prove that (6.14) cancels with the derivative of (6.13), i.e.,

$$\frac{d}{dt} G(\ _1, \ldots, \ _k + t\ , \ldots, \ _N)\Big|_{t=0}.$$

In fact, rewriting (6.13) as the sum of the terms

$$\int_{B_R(\ _k)} \left\{ \sum_{i\neq k} W(\ (\ _i)) + \sum_{i<j} C[\ (\ _i)] \cdot \ _j + W(\ (\ _0)) \right\} da$$
$$+\int_{\partial B_R(\ _k)} \sum_i (\ _0 \cdot C[\ (\ _i)]\ \ )\ ds \tag{6.15}$$

and

$$\sum_{h\neq k} \int_{B_R(\ _h)} \left\{ \sum_{i\neq h} W(\ (\ _i)) + \sum_{i<j} C[\ (\ _i)] \cdot \ _j + W(\ (\ _0)) \right\} da$$
$$+\sum_{h\neq k} \int_{\partial B_R(\ _h)} \sum_i (\ _0 \cdot C[\ (\ _i)]\ \ )\ ds \tag{6.16}$$

and computing the derivative of (6.15) with respect to t and using Lemma 6.2, we find

$$\int_{B_R(\ _k)} \left\{ \sum_{i\neq k} C[\ (\ _i)] \cdot D_t\ \ _i + \sum_{i<j} C[\ (\ _i)] \cdot D_t\ \ _j \right.$$
$$\left. + \sum_{i<j} C[\ (\ _j)] \cdot D_t\ \ _i + C[\ (\ _0)] \cdot \nabla(D_t\ \ _0) \right\} da$$
$$+ \int_{\partial B_R(\ _k)} \left\{ \sum_i\ \ _0 \cdot C[\ (D_t\ \ _i)]\ \ + \sum_i D_t\ \ _0 \cdot C[\ (\ _i)]\ \ \right\} ds,$$

where $D_t \; _j := \dot{} _j + \nabla(\; _j)$ and $D_t \; _0 = \dot{}_0 + (\nabla \; _0)$. Since $D_t \; _k = 0$ and $D_t \; _j = \nabla(\; _j) \; = \nabla(\; _j \;)$ for $j \neq k$ (since $\dot{} _j = 0$ for $j \neq k$), using the divergence theorem and noting that $\operatorname{Div} C[\; (D_t \; _i)] = 0$ on $\mathbb{R}^2 \setminus \{ \; _i\}$, we find

$$\int_{\partial B_R(\; _k)} \Bigg\{ D_t \; _0 \cdot C[\; (\; _0)] \; + \sum_{i\neq k} (\; _i \;) \cdot C[\; (\; _0)] + \sum_{i\neq k} (\; _i \;) \cdot C[\; (\; _i)] \; + \sum_{i\neq k}\sum_{j\neq i} (\; _i \;) \cdot C[\; (\; _j)] \; \Bigg\} \, ds,$$

which becomes finally

$$\int_{\partial B_R(\; _k)} (D_t \; _0 + \sum_{i\neq k} \; _i \;) \cdot C[\; (\; _0)] \; ds. \tag{6.17}$$

Consider now the derivative of (6.16) with respect to t: using Lemma 6.2, we find

$$\sum_{h\neq k} \int_{B_R(\; _h)} \Bigg\{ C[\; (\; _k)] \cdot \dot{} _k + \sum_{i\neq k} C[\; (\; _i)] \cdot \dot{} _k + C[\; (\; _0)] \cdot \nabla \dot{}_0 \Bigg\} \, da + \sum_{h\neq k} \int_{\partial B_R(\; _h)} \Bigg\{ \; _0 \cdot C[\; (\dot{} _k)] \; + \sum_i \dot{}_0 \cdot C[\; (\; _i)] \; \Bigg\} \, ds,$$

and using the fact that $\operatorname{Div} C[\; (\dot{} _k)] = 0$ on $\mathbb{R}^2 \setminus \{ \; _k\}$, and $\dot{} _k = -\nabla(\; _k \;)$ we finally obtain

$$\sum_{h\neq k} \int_{\partial B_R(\; _h)} (\dot{}_0 - \; _k \;) \cdot C[\; (\; _0)] \; ds. \tag{6.18}$$

Relation (6.10) follows now upon noting that the sum of (6.14), (6.17), and (6.18) vanishes. ∎

Appendix: Poincaré and Korn inequalities for Ω_ε.

The basic tool to study the compactness of a sequence of minimizers of problem (2.9) is Korn's inequality: we prove here that for a perforated domain such as Ω_ε, under mild regularity assumptions, Korn's and Poincaré's inequalities hold uniformly in ε as $\varepsilon \to 0$. Also, we show that the trace constant for Ω_ε may be chosen independent of ε.

We begin by proving that Poincaré's inequality holds for each domain Ω_ε uniformly in ε, when Ω has the cone property.

Proposition A.1. *Let Ω be a bounded open connected domain in $\mathbb{R}^2$ with the cone property: then, for any $\; \in H^1(\Omega_\varepsilon; \mathbb{R}^2)$,*

$$\int_{\Omega_\varepsilon} | \; - \; _B|^2 \, d \; \leq c \int_{\Omega_\varepsilon} |\nabla \; |^2 \, d \; , \tag{A.1}$$

where

$$_B := \frac{1}{|B|} \int_B \; da, \tag{A.2}$$

B is any fixed ball contained in $\Omega \setminus \{\ _1, \dots, \ _N\}$, and the constant c is independent of ε (but may depend on $\Omega \setminus \{\ _1, \dots, \ _N\}$ and B).

Proof. Fix $R > 2\epsilon$ so small that $B_R(\ _i) \subset \Omega \setminus B$ for any $i = 1, \dots, N$, $B_R(\ _i) \cap B_R(\ _j) = \emptyset$ for $i \neq j$, and decompose Ω_ε as the union of the annuli $C_{\varepsilon, \frac{R}{2}}(\ _i) = B_{\frac{R}{2}}(\ _i) \setminus B_\varepsilon(\ _i)$ and its complement Ω', which is independent of ε. Since Ω' is still connected and has the cone property, by the classical Poincaré's inequality we may find a constant c depending on Ω' and B such that

$$\text{(A.3)} \qquad \int_{\Omega'} |\ \ -\ \ _B|^2\, d\ \ \leq c \int_{\Omega'} |\nabla\ \ |^2\, d\ \ .$$

For each fixed $i = 1, \dots, N$ let (ϱ, ϑ) be polar coordinates centered at $\ _i$: for $\mathcal{L}^1$ almost everywhere $\varepsilon \leq s \leq \varrho < R$ and $\theta \in [0, 2\pi]$ we have

$$\ (s, \vartheta) = \ \ (\varrho, \vartheta) - \int_s^\varrho \frac{\partial\ }{\partial \varrho}(r, \vartheta)\, dr,$$

from which it follows, by Hölder's inequality,

$$|\ \ (s, \vartheta)|^2 \leq 2|\ \ (\varrho, \vartheta)|^2 + 2R \int_s^R \left| \frac{\partial\ }{\partial \varrho}(r, \vartheta) \right|^2 dr.$$

Integrating with respect to ϑ and multiplying by s yields

$$\text{(A.4)} \qquad \int_0^{2\pi} s|\ \ (s, \vartheta)|^2\, d\vartheta \leq 2 \int_0^{2\pi} \varrho|\ \ (\varrho, \vartheta)|^2\, d\vartheta + 2R \int_{C_{\varepsilon, R}(\ _i)} |\nabla\ \ |^2\, da,$$

where we have used the fact that $s \leq \varrho$, r. By integrating first with respect to s in $[\varepsilon, \frac{R}{2}]$ and then to ϱ in $[\frac{R}{2}, R]$ we obtain

$$\int_{C_{\varepsilon, \frac{R}{2}}(\ _i)} |\ \ |^2\, da \leq 2 \int_{C_{\frac{R}{2}, R}(\ _i)} |\ \ |^2\, da + 2R \int_{C_{\varepsilon, R}(\ _i)} |\nabla\ \ |^2\, da.$$

If we now replace $\ $ with $\ \ -\ \ _B$ in the previous inequality we get

$$\int_{C_{\varepsilon, \frac{R}{2}}(\ _i)} |\ \ -\ \ _B|^2\, da \leq 2 \int_{C_{\frac{R}{2}, R}(\ _i)} |\ \ -\ \ _B|^2\, da + 2R \int_{C_{\varepsilon, R}(\ _i)} |\nabla\ \ |^2\, da,$$

which together with (A.3) and the fact that $C_{\frac{R}{2}, R} \subset \Omega'$ concludes the proof. ◻

We now turn to Korn's inequality. First we notice that if Ω has the cone property, then for all $\varepsilon > 0$ Ω_ε has the cone property, and the following version of this inequality holds [9, 17, 24].

Proposition A.2 (Korn's inequality I). *Let $\Omega \subset \mathbb{R}^2$ be an open bounded connected domain with the cone property, and let $\ \ \in H^1(\Omega_\varepsilon; \mathbb{R}^2)$ such that*

$$\text{(A.5)} \qquad \int_{\Omega_\varepsilon} \left(\nabla\ \ - (\nabla\ \ )^\top\right) da = 0;$$

then there exists a constant c_ε such that

$$\text{(A.6)} \qquad \int_{\Omega_\varepsilon} |\nabla\ \ |^2 da \leq c_\varepsilon \int_{\Omega_\varepsilon} |\ \ (\ \ )|^2 da.$$

We continue to denote by c_ε the infimum of all constants satisfying (A.6) and refer to it as Korn's constant for Ω_ε. The following result shows that c_ε is bounded from above independently of ε as $\varepsilon \to 0$.

Proposition A.3. *Let c_ε be Korn's constant for Ω_ε as defined in* (A.6)*: then there exists a constant $c < \infty$, independent of ε (but possibly depending on Ω), such that*

$$c_\varepsilon \leq c \tag{A.7}$$

for all $\varepsilon > 0$.

Proof. Consider first the case $N = 1$, and let $\ _1 = \mathbf{0}$, so that $\Omega_\varepsilon = \Omega \setminus B_\varepsilon(\mathbf{0})$. The proof follows from two results of [9]. The first result states that the minimum value for Korn's constant of the annulus $C_{\varepsilon,R}(\mathbf{0})$ with internal radius ε and external radius R (under the constraint (A.5)) is

$$4\left[1 - \left(\frac{3R^2\varepsilon^2}{R^4 + R^2\varepsilon^2 + \varepsilon^4}\right)\right]^{-1}, \tag{A.8}$$

which tends to Korn's constant for the circle $c = 4$ as $\varepsilon \to 0$. The second result states that if Korn's inequality (A.5)–(A.6) holds for two open bounded domains Ω_1 and Ω_2 such that $|\Omega_1 \cap \Omega_2| > 0$, then it also holds for $\Omega_1 \cup \Omega_2$, and

$$c_{12} \leq c_1 + c_2 + \frac{\min\{|\Omega_1|, |\Omega_2|\}}{|\Omega_1 \cap \Omega_2|}\left(\sqrt{c_1} + \sqrt{c_2}\right)^2 \tag{A.9}$$

with c_{12}, c_1, and c_2 the Korn's constants of $\Omega_1 \cup \Omega_2$, Ω_1, and Ω_2, respectively.

To prove (A.7), choose R such that $2R < d(\mathbf{0}, \partial\Omega)$, let $\Omega_1 = C_{\varepsilon,2R}(\mathbf{0})$ and $\Omega_2 = \Omega_\varepsilon \setminus C_{\varepsilon,R}(\mathbf{0})$, and apply (A.9): since c_2 is independent of ε and, by (A.8), $c_1 \to 4$ as $\varepsilon \to 0$, Korn's constant $c_\varepsilon = c_{12}$ is bounded from above, and the thesis follows for $N = 1$.

When $N > 1$, to obtain the thesis it is sufficient to iterate the above procedure: define

$$\begin{cases} \tilde{\Omega}_0 := \Omega \setminus \left(\cup_{i=1}^N B_R(\ _i)\right) \\ \vdots \\ \tilde{\Omega}_i := \tilde{\Omega}_{i-1} \cup \left(B_{2R}(\ _i) \setminus B_\varepsilon(\ _i)\right), \qquad i = 1, \dots, N-1, \\ \vdots \\ \tilde{\Omega}_N := \Omega_\varepsilon. \end{cases}$$

Applying (A.9) to each $\tilde{\Omega}_i$ we obtain

$$\begin{aligned} \tilde{c}_i &\leq \tilde{c}_{i-1} + \tilde{c}_\varepsilon + \frac{\min\{|\tilde{\Omega}_{i-1}|, |c_{\varepsilon,2R}(\ _i)|\}}{|\tilde{\Omega}_{i-1} \cap c_{\varepsilon,2R}(\ _i)|}\left(\sqrt{\tilde{c}_{i-1}} + \sqrt{\tilde{c}_\varepsilon}\right)^2 \\ &= \tilde{c}_{i-1} + \tilde{c}_\varepsilon + \frac{4R^2 - \varepsilon^2}{3R^2}\left(\sqrt{\tilde{c}_{i-1}} + \sqrt{\tilde{c}_\varepsilon}\right)^2, \end{aligned}$$

where $\tilde{c}_i$ and $\tilde{c}_\varepsilon$ are Korn's constants for $\tilde{\Omega}_i$ and $C_{\varepsilon,2R}(\ _i)$, respectively. Using the relation $(a+b)^2 \leq 2(a^2 + b^2)$, and taking $\varepsilon = 0$, this relation implies

$$\tilde{c}_i \leq \frac{11}{3}(\tilde{c}_{i-1} + \tilde{c}_\varepsilon),$$

which yields in turn

$$\tilde{c}_N \leq \left(\frac{11}{3}\right)^N \tilde{c}_0 + \left(\sum_{i=1}^{N}\left(\frac{11}{3}\right)^i\right)\tilde{c}_\varepsilon.$$

Since $\tilde{c}_\varepsilon$ is given by (A.8) and is bounded from above, and $\tilde{c}_0$ is independent of ε, this relation shows that $c_\varepsilon = \tilde{c}_N$ is also bounded from above as $\varepsilon \to 0$. ∎

Korn's inequality extends trivially to displacement fields which do not satisfy (A.5).

COROLLARY A.4 (Korn's inequality I′). *Let $\Omega \subset \mathbb{R}^2$ be an open bounded connected domain with the cone property, and $\in H^1(\Omega_\varepsilon;\mathbb{R}^2)$: then there exists a constant c, independent of ε, such that*

$$\int_{\Omega_\varepsilon} |\nabla \quad - \quad |^2 da \leq c \int_{\Omega_\varepsilon} | \quad (\quad)|^2 da, \tag{A.10}$$

where

$$= \frac{1}{2}\int_{\Omega_\varepsilon} (\nabla \quad - (\nabla \quad)^\top)\, da. \tag{A.11}$$

Combining (A.1), (A.6), and (A.7), we finally obtain the following basic inequality.

PROPOSITION A.5 (Korn's inequality II). *Let $\in H^1(\Omega_\varepsilon;\mathbb{R}^2)$ such that*

$$\int_{\Omega_\varepsilon} \left(\nabla \quad - (\nabla \quad)^\top\right) da = 0; \tag{A.12}$$

then there exists a constant c, independent of ε, such that

$$\int_{\Omega_\varepsilon} | \quad - \quad _B|^2 da + \int_{\Omega_\varepsilon} |\nabla \quad |^2 da \leq c \int_{\Omega_\varepsilon} | \quad (\quad)|^2 da, \tag{A.13}$$

where

$$_B := \frac{1}{|B|}\int_B \quad da, \tag{A.14}$$

and B is any fixed ball contained in $\Omega \setminus \{ \ _1, \ldots, \ _N\}$.

We now show that the trace constant for Ω_ε may be chosen to be independent of ε.

PROPOSITION A.6. *Let $\in H^1(\Omega_\varepsilon;\mathbb{R}^2)$: then there exists a positive constant c, independent of ε, such that*

$$\int_{\partial\Omega_\varepsilon} | \quad |^2 ds \leq c\left(\int_{\Omega_\varepsilon} | \quad |^2 da + \int_{\Omega_\varepsilon} |\nabla \quad |^2 da\right). \tag{A.15}$$

Proof. Taking $s = \varepsilon$ in (A.4) we have

$$\int_{\partial B_\varepsilon(\ _i)} | \quad |^2 ds \leq 2\int_0^{2\pi} \varrho\, | \quad (\varrho,\vartheta)|^2\, d\vartheta + 2R\int_{C_{\varepsilon,R}(\ _i)} |\nabla \quad |^2\, da.$$

By averaging with respect to ϱ over $[\varepsilon, R]$ we obtain

$$\int_{\partial B_\varepsilon(\ _i)} | \quad |^2 ds \leq \frac{2}{R-\varepsilon}\int_{C_{\varepsilon,R}(\ _i)} | \quad |^2\, da + 2R\int_{C_{\varepsilon,R}(\ _i)} |\nabla \quad |^2\, da.$$

Hence, for ε sufficiently small, there exists a constant c_1 such that

$$\int_{\partial B_\varepsilon(\ _i)} |\ |^2 ds \le c_1 \left(\int_{\Omega_\varepsilon} |\ |^2\, da + \int_{\Omega_\varepsilon} |\nabla\ |^2\, da \right) \tag{A.16}$$

for each $i = 1, \dots, N$. Now, let c_2 be the trace constant for Ω_R, so that

(A.17)
$$\int_{\partial\Omega} |\ |^2 ds \le c_2 \left(\int_{\Omega_R} |\ |^2\, da + \int_{\Omega_R} |\nabla\ |^2\, da \right) \le c_2 \left(\int_{\Omega_\varepsilon} |\ |^2\, da + \int_{\Omega_\varepsilon} |\nabla\ |^2\, da \right).$$

Adding the expressions above we finally obtain the thesis. □

Finally, we show that functions in $H^1(\Omega_\varepsilon;\mathbb{R}^2)$ can be extended to Ω with extension constant independent of ε.

Proposition A.7. *Let* $\ \in H^1(\Omega_\varepsilon;\mathbb{R}^2)$*: then* $\ $ *admits an extension* $\hat{\ } \in H^1(\Omega,\mathbb{R}^2)$ *such that*

$$\|\hat{\ }\|_{H^1(\Omega,\mathbb{R}^2)} \le c \|\ \|_{H^1(\Omega_\varepsilon;\mathbb{R}^2)}, \tag{A.18}$$

where the constant c *is independent of* ε*.*

Proof. First notice that, using a partition of unity, we may assume that $\ \in H^1(\mathbb{R}^2 \setminus B_\varepsilon(\mathbf{0});\mathbb{R}^2)$, and $\ = 0$ outside a compact in $\mathbb{R}^2$. Let

$$\hat{\ }(\ ) = \begin{cases} \ \left(\dfrac{\varepsilon^2}{|\ |^2}\ \right), & \ \in B_\varepsilon(\mathbf{0}), \\ \ (\ ), & \ \in \mathbb{R}^2 \setminus B_\varepsilon(\mathbf{0}), \end{cases}$$

so that $\hat{\ }(\ ) = \ (\ )$ when $|\ | = \varepsilon$. Then

$$\int_{B_\varepsilon(\mathbf{0})} |\hat{\ }(\ )|^2\, da = \int_{B_\varepsilon(\mathbf{0})} \left| \ \left(\frac{\varepsilon^2}{|\ |^2}\ \right)\right|^2 da = \int_{\mathbb{R}^2\setminus B_\varepsilon(\mathbf{0})} \frac{\varepsilon^4}{|\ |^4} |\ (\ )|^2\, da$$
$$\le \int_{\mathbb{R}^2\setminus B_\varepsilon(\mathbf{0})} |\ (\ )|^2\, da,$$

since the modulus of the Jacobian of the transformation $\ \to \varepsilon^2\ /|\ |^2$ is $|J| = \varepsilon^4/|\ |^4$, and $\varepsilon^4/|\ |^4 \le 1$ for $|\ | \ge \epsilon$. Also, notice that

$$\nabla\hat{\ }(\ ) = \nabla\ \left(\frac{\varepsilon^2}{|\ |^2}\ \right)\left[\frac{\varepsilon^2}{|\ |^2}\left(\mathbf{1} - \frac{2}{|\ |^2}\ \otimes\ \right)\right]$$

and

$$\left|\mathbf{1} - \frac{2}{|\ |^2}\ \otimes\ \right|^2 = 1.$$

Hence,

$$\int_{B_\varepsilon(\mathbf{0})} |\nabla\hat{\ }(\ )|^2\, da \le M \int_{B_\varepsilon(\mathbf{0})} \frac{\varepsilon^4}{|\ |^4} \left|\nabla\ \left(\frac{\varepsilon^2}{|\ |^2}\ \right)\right|^2 da$$
$$= M \int_{\mathbb{R}^2\setminus B_\varepsilon(\mathbf{0})} \frac{\varepsilon^4}{|\ |^4} \frac{|\ |^4}{\varepsilon^4} |\nabla\ (\ )|^2\, da$$
$$\le M \int_{\mathbb{R}^2\setminus B_\varepsilon(\mathbf{0})} |\nabla\ (\ )|^2\, da$$

with M a positive constant independent of ε. Since $\hat{} =$ on $\mathbb{R}^2 \setminus B_\varepsilon(\mathbf{0})$, it follows that there exists a constant c independent of ε such that

$$\|\hat{}\|_{H^1(\mathbb{R}^2,\mathbb{R}^2)} \leq c\| \quad \|_{H^1(\mathbb{R}^2\setminus B_\varepsilon(\mathbf{0});\mathbb{R}^2)},$$

which implies the thesis. $\square$

Acknowledgments. The authors wish to thank A. Garroni for interesting conversations on the subject of this paper.

REFERENCES

[1] E. C. Aifantis, *The physics of plastic deformation*, Internat. J. Plasticity, 3 (1987), pp. 211–247.

[2] G. Alberti, S. Baldo, and G. Orlandi, *Variational convergence for functionals of Ginzburg-Landau type*, Indiana Univ. Math. J., 54 (2005), pp. 1411–1472.

[3] R. J. Asaro, *Micromechanics of crystals and polycrystals*, Adv. Appl. Mech., 23 (1983), pp. 1–115.

[4] F. Bethuel, H. Brezis, B. D. Coleman, and F. Hélein, *Bifurcation analysis of minimizing harmonic maps describing the equilibrium of nematic phases between cylinders*, Arch. Ration. Mech. Anal., 118 (1992), pp. 149–168.

[5] F. Bethuel, H. Brezis, and F. Hélein, *Ginzburg-Landau Vortices*, Birkhäuser, Basel, 1994.

[6] F. Bethuel, H. Brezis, and G. Orlandi, *Asymptotics for the Ginzburg-Landau equation in arbitrary dimensions*, J. Funct. Anal., 186 (2001), pp. 432–520.

[7] P. Cermelli and M. E. Gurtin, *The motion of screw dislocations in materials undergoing anti-plane shear: Glide, cross-slip, fine cross-slip*, Arch. Ration. Mech. Anal. 148 (1999), pp. 3–52.

[8] P. Cermelli, M. E. Gurtin, and G. Leoni, *Energies for incoherent films: An analytical approach*, Interfaces Free Boundaries, 1 (1999), pp. 81–105.

[9] C. M. Dafermos, *Some remarks on Korn's inequality*, Z. Angew. Mat. Phys., 19 (1968), pp. 913–920.

[10] R. Dautray and J. L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, vol. 3, Springer, Berlin, 1988.

[11] J. L. Ericksen, *Remarks concerning forces on line defect*, Z. Angew. Mat. Phy., 46 (1995), pp. S247–S271.

[12] J. D. Eshelby, *The force on elastic singularity*, Phil. Trans. R. Soc. London A, 244 (1951), pp. 87–112.

[13] J. D. Eshelby, *The continuum theory of lattice defects*, in Solid State Physics 3, F. Seitz and D. Turnbull, eds., Academic Press, New York, 3 1956, pp. 79–144.

[14] L. C. Evans, *Weak convergence methods for nonlinear partial differential equations*, Regional Conference Ser. Math., 74, American Mathematical Society, Providence, RI, 1990.

[15] N. A. Fleck and J. W. Hutchinson, *Strain gradient plasticity*, Adv. Appl. Mech., 33 (1997), pp. 295–361.

[16] D. François, A. Pineau, and A. O Zaoui, *Mechanical Behaviour of Materials* I, Kluwer, Dordrecht, The Netherlands, 1998.

[17] K. O. Friedrichs, *On the boundary-value problems of the theory of elasticity and Korn's inequality*, Ann. Math., 48 (1947), pp. 441–471.

[18] M. E. Gurtin, *Configurational Forces as Basic Concepts of Continuum Physics*, Springer-Verlag, New York, 2000.

[19] M. E. Gurtin, *On the plasticity of single crystals: Free energy, microforces, plastic-strain gradients*, J. Mech. Phys. Solids, 48 (2000), pp. 989–1036.

[20] M. E. Gurtin, *A gradient theory of single-crystal viscoplasticity that accounts for geometrically necessary dislocations*, J. Mech. Phys. Solids, 50 (2002), pp. 5–32.

[21] R. L. Jerrard and H. M. Soner, *Dynamics of Ginzburg-Landau vortices*, Arch. Ration. Mech. Anal., 142 (1998), pp. 99–125.

[22] R. L. Jerrard and H. M. Soner, *The Jacobian and the Ginzburg-Landau energy*, Calc. Var. Partial Differential Equations, 14 (2002), pp. 151–191.

[23] J. P. Hirth and J. Lothe, *Theory of Dislocations*, 2nd ed., John Wiley, New York, 1982.

[24] C. O. Horgan, *Korn's inequalities and their applications in continuum mechanics*, SIAM Rev., 37 (1995), pp. 491–511.

[25] D. Kuhlmann-Wilsdorf, *Theory of plastic deformation: Properties of low energy dislocation structures*, Materials Sci. Engrg., A113 (1989), pp. 1–41.
[26] F.-H. Lin and C. Liu, *Static and dynamic theories of liquid crystals*, J. Partial Differential Equations, 14 (2001), pp. 289–330.
[27] F.-H. Lin and J. X. Xin, *On the dynamical law of the Ginzburg-Landau vortices on the plane*, Comm. Pure Appl. Math., 52 (1999), pp. 1189–1212.
[28] N. I. Muskhelishvili, *Some Basic Problems in the Mathematical Theory of Elasticity*, Noordhoff, Groningen, The Netherlands, 1963.
[29] F. R. N. Nabarro, *Theory of Crystal Dislocations*, Clarendon Press, Oxford, UK, 1967.
[30] M. Peach and J. S. Köhler, *The forces exerted on dislocations and the stress field produced by them*, Phys. Rev., 80 (1950), pp. 436–439.

SIAM J. MATH. ANAL.
Vol. 37, No. 4, pp. 1161–1206

CORRECTIONS TO THE KdV APPROXIMATION FOR WATER WAVES*

J. DOUGLAS WRIGHT†

This article is dedicated to the memory of Dr. Joseph Hammack

Abstract. In order to investigate corrections to the common KdV approximation for surface water waves in a canal, we derive modulation equations for the evolution of long wavelength initial data. We work in Lagrangian coordinates. The equations which govern corrections to the KdV approximation consist of linearized and inhomogeneous KdV equations plus an inhomogeneous wave equation. These equations are explicitly solvable and we prove estimates showing that they do indeed give a significantly better approximation than the KdV equation alone.

Key words. water wave equation, KdV equation, linearized KdV equation, modulation equations, solitary waves, solitons, collisions of solitary waves

AMS subject classifications. 76B15, 35Q51, 35Q53

DOI. 10.1137/S0036141004444202

1. Introduction.

It is often easier to write down a partial differential equation which models a physical phenomena than it is to study solutions of such an equation. Equations which model the evolution of the surface of a fluid in a canal have been known since at least the 19th century; however, it has only been in recent years that questions of existence and uniqueness for general initial data have been answered (see [35], [36], and [22]). Moreover, numerical simulations of water waves and similarly complex phenomena are frequently time consuming and challenging to implement. Consequently, it can be quite difficult to say much about the behavior of a general solution. Therefore scientists often restrict their attention to limiting cases—for instance, one may assume that solutions are of long wavelength and small amplitude (see Figure 1). Under such a supposition, a *modulation equation* may be (formally) derived. In particular, one hopes that the modulation equation

- is well-posed,
- is either explicitly solvable or easy to solve numerically, and
- captures the essential behavior of the original system.

Remarkably, many seemingly disparate physical phenomena possess modulation equations of the same form. For solutions of long wavelength, Korteweg–de Vries (KdV) equations are often used as modulation equations for a wide variety of nonlinear dispersive systems, including the water wave equation, the Euler–Poisson equations for plasma dynamics, and the Fermi–Pasta–Ulam equation for the interaction of particles in an infinite lattice.

Despite the fact that modulation equations have been in use for over a hundred years—the KdV equation was first proposed as a model for water waves by Boussinesq in 1872 and also by Korteweg and de Vries in 1895—only recently have attempts been made to rigorously connect the behavior of the modulation equations to the original

*Received by the editors May 31, 2004; accepted for publication (in revised form) November 8, 2004; published electronically January 6, 2006. This research was supported by NSF grant DMS-0103915.
http://www.siam.org/journals/sima/37-4/44420.html
†School of Mathematics, University of Minnesota, Minneapolis, MN 55455 (jdoug@math.umn.edu).

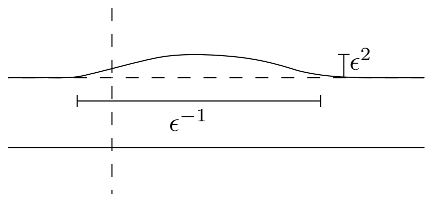


FIG. 1. *The long wave, small amplitude scaling.*

physical problem. In particular, through the work of Craig [10], Kano and Nishida [18], [19], Kalyakin [17], Schneider [27], Ben Youssef and Colin [2], and Schneider and Wayne [28], [29], the validity of KdV equations as a leading order approximation to the evolution of long wavelength water waves and to a number of other dispersive partial differential equations has been established.

In many respects, the KdV equation is an ideal modulation equation; it is simple in form and explicitly solvable via the inverse scattering transform. Nevertheless, both experimentally and numerically one observes deviations from the predictions of the KdV approximation. In this paper we derive a hierarchy of modulation equations which govern corrections to the KdV model and also prove rigorously that these higher order equations do indeed improve the accuracy of the approximation. While the correction is valid in general long wavelength/small amplitude settings, heuristically the model is set up to better approximate interactions between solitary waves—both counterpropagating collisions and unidirectional interactions.

Note that in [33], as a case study, Wayne and Wright examined higher order corrections to the KdV approximation to a Boussinesq equation. As the KdV equation is in some sense a universal approximation for long waves, we expect that the equations for corrections to this approximation will also be universal. Indeed, our results show that the higher order corrections for the water wave equation are nearly identical to those for the Boussinesq equation. Also, since a significant part of the work for the Boussinesq problem consists of showing that the modulation equations have well-behaved solutions over the time scales of interest, this is of use in tackling the water wave problem.

We now describe our results in some detail. The equations of motion for a water wave in an infinitely long canal (commonly called the *water wave equation*) are

$$\begin{gathered} x_{tt}(1+x_\alpha) + y_\alpha(1+y_{tt}) = 0, \\ y_t = K(x,y)x_t, \\ \alpha \in \mathbb{R}, \quad t \geq 0, \quad (x(\alpha,t), y(\alpha,t)) \in \mathbb{R}^2, \end{gathered} \tag{WW}$$

where $K(x,y)$ is a complicated operator (see section 3 for the definition of K) and $(\alpha + x(\alpha,t), y(\alpha,t))$ parameterizes the free surface. According to the KdV approximation results of [28], to the order of the approximation long wavelength solutions of (WW) split up into two pieces, one a right-moving wave train and one a left-moving wave train. Each of these wave trains evolves according to a KdV equation, and *there is no interaction between the left- and right-moving pieces*. That is, for $0 < \epsilon \ll 1$, if we scale amplitudes to be $O(\epsilon^2)$ (i.e., small) and wavelengths to be $O(\epsilon^{-1})$ (i.e., long), then for times of $O(\epsilon^{-3})$, solutions to (WW) satisfy

$$-x_\alpha(\alpha,t) = \epsilon^2 U(\epsilon(\alpha - t), \epsilon^3 t) + \epsilon^2 V(\epsilon(\alpha + t), \epsilon^3 t) + O(\epsilon^4), \tag{1}$$

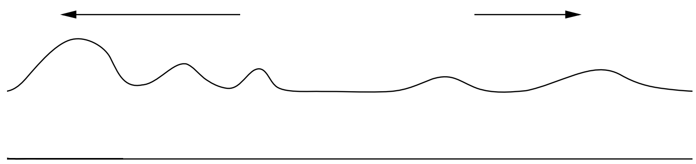


FIG. 2. *Sketch of the right- and left-moving wave trains.*

where U and V satisfy the KdV equations

$$\text{(KdV)} \qquad \begin{aligned} -2\partial_T U &= \frac{1}{3}\partial^3_{\beta_-} U + \frac{3}{2}\partial_{\beta_-}(U^2), \\ 2\partial_T V &= \frac{1}{3}\partial^3_{\beta_+} V + \frac{3}{2}\partial_{\beta_+}(V^2). \end{aligned}$$

See Figure 2. Here, $\beta_\pm = \epsilon(\beta \pm t)$ represent long wavelength moving reference frames and $T = \epsilon^3 t$ is the very long time scale coordinate. For technical reasons (which we discuss later), $-x_\alpha$ is the natural variable to estimate for the water wave equation. To lowest order, $-x_\alpha$ is proportional to the height of the wave. At higher order, this ceases to be true, though for purposes of intuition one can think of $-x_\alpha$ as representing the wave amplitude.

The KdV equation was initially derived from the water wave equation in an attempt to prove the existence of a solitary wave solution for waves in a canal. Famously, the KdV equation admits solitary wave solutions and also multisoliton solutions. (See Figure 3.) Note that the existence of a true solitary wave solution to (WW) was established first by Friedrichs and Hyers in [14] and subsequently by Beale in [1]. Also note that there exist standing wave solutions (also known as Stokes' waves) to (WW) which in many ways are similar to solitary waves, though our methods do not carry over to studying such behavior—see Toland [32] and Constantin and Strauss [8] for an overview of Stokes' waves in the irrotational and rotational problems, respectively.

We will frequently refer to multisoliton solutions as "overtaking wave" collisions. We remind the reader that the only notable first order effect after such a collision is that the waves are phase shifted after the collision.

Given the results in [28], one expects to see similar behavior in solutions systems modeled by KdV equations. Though it is unknown if these soliton-like solutions persist globally, analogous behavior is indeed observed for very long times (see [15]). The most notable deviation between true solutions and the KdV approximation is the size of the phase shift after a collision. In addition, soliton-like solutions to the type of systems we study frequently develop a very small amplitude dispersive wave train behind each soliton, which moves in the same direction; see Figure 4. The KdV approximation does not predict the existence of these dispersive wave trains. As these sorts of discrepancies are observed even in the case where there is only one wave train moving unidirectionally, we believe that they are, loosely, independent of interactions between the left- and right-moving wave trains. They reflect intrinsic differences between the approximation and the original system.

On the other hand, there is evidence that a noticeable interaction takes place between the left- and right-moving waves. One can see from the form of the approximation in (1) that during a head-on collision of waves moving in opposite directions, the KdV approximation predicts that the heights of the waves add linearly. In true head-on collisions in solutions to the water wave equation, however, the height of the

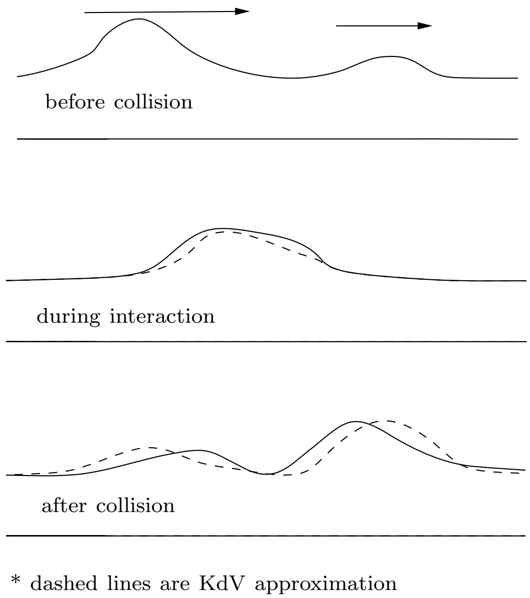


FIG. 3. *Sketch of the overtaking wave interaction.*

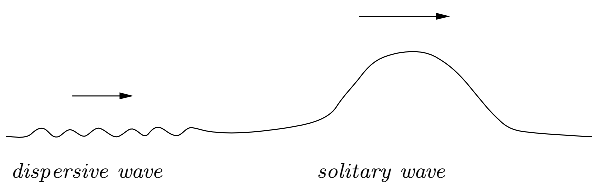


FIG. 4. *Sketch of the dispersive wave.*

waves is slightly different from the sum of the heights of the waves taken separately—it is slightly larger. (See the works of Maxworthy [23], Byatt-Smith [6], [7], Cooker, Weidman, and Bale [9], and Su and Mirie [30], [31].) We sketch this in Figure 5.

Thus we might expect two types of corrections to the KdV approximation:

- corrections due to the fact that, even in the case of a purely right- (or left-) moving wave train, solutions to the water wave equation are not exactly described by solutions to the KdV equation—we will refer to this source of error as *unidirectional error*;
- corrections due the fact that the left- and right-moving wave trains will interact at higher order—we call such errors *counterpropagation error*.

Both of these types of corrections are apparent in our results, and to incorporate these two types of corrections, we add an additional three functions to the KdV wave trains. The first two, F and G, will correct for unidirectional errors. The third, P, will correct for counterpropagation errors. We scale the amplitudes of these three functions so they are $O(\epsilon^4)$, which is the same as the order of the error in using only the KdV equations. F and G will take the same functional form as U and V, as they correct for differences between the approximate and actual wave trains. Therefore we

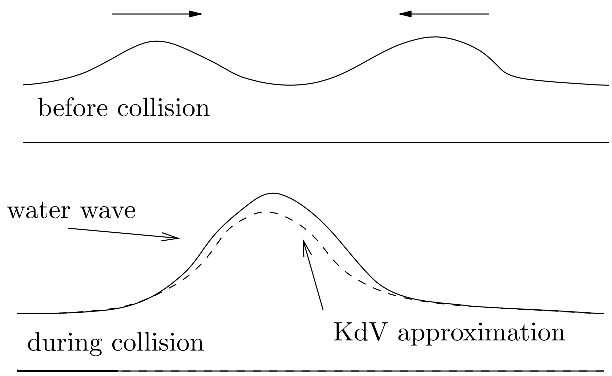


FIG. 5. *Sketch of the head-on collision.*

add

$$\epsilon^4 F(\epsilon(\alpha - t), \epsilon^3 t) + \epsilon^4 G(\epsilon(\alpha + t), \epsilon^3 t)$$

to the first order approximation (1).

We do not expect P to be moving strictly left or right, as it corrects for the interaction between waves moving in opposite directions. Thus its spatial dependence will be on $\beta = \epsilon\alpha$. Suppose that the functions U and V are solitary wave solutions. In the long wavelength variables we are considering, this means that the right- and left-moving wave packets are large only over a length of $O(\epsilon^{-1})$. In addition, the reference frame moves with unit velocity. Thus we expect any interaction of the two waves to last a time of $O(\epsilon^{-1})$. Accordingly, we let P depend on the time variable $\tau = \epsilon t$. That is, we add a correction term of the form

$$\epsilon^4 P(\epsilon\alpha, \epsilon t)$$

to the KdV model.

Through formal means we find that P satisfies an inhomogeneous wave equation

$$\text{(IW)} \qquad \partial_\tau^2 P - \partial_\beta^2 P = 3\partial_\beta^2 \left(U(\beta - \tau, \epsilon^2\tau) V(\beta + \tau, \epsilon^2\tau) \right).$$

Similarly, F and G satisfy a pair of driven, linearized KdV equations

$$\text{(LK)} \qquad \begin{aligned} -2\partial_T F =& \frac{1}{3}\partial_{\beta_-}^3 F + 3\partial_{\beta_-}(UF) + J^-, \\ 2\partial_T G =& \frac{1}{3}\partial_{\beta_+}^3 G + 3\partial_{\beta_+}(VG) + J^+. \end{aligned}$$

Notice that these equations are linearized about the KdV solutions U and V. The inhomogeneous terms J^- and J^+ are made up of a combination of sums and products of U, V, and P. For explicit forms of these driving terms see equations (22). Linearized KdV equations are explicitly solvable, though this is a complicated matter (see [25] and [16]). However, solutions are simple to compute numerically.

One can solve inhomogeneous wave equations explicitly and easily via the method of characteristics. Moreover, we can reduce such systems to a pair of transport equations by the following fact.

FACT 1. *If* $\partial_t f - \partial_x f = 1/2\partial_x h$ *and* $\partial_t g + \partial_x g = -1/2\partial_x h$, *then* $q = f + g$ *satisfies* $\partial_t^2 q - \partial_x^2 q = \partial_x^2 h$.

Thus, we have

$$\begin{aligned} P^-_\tau + P^-_\beta &= -\frac{3}{2}\partial_\beta (U(\beta-\tau,\epsilon^2\tau)V(\beta+\tau,\epsilon^2\tau)), \\ P^+_\tau - P^+_\beta &= \frac{3}{2}\partial_\beta (U(\beta-\tau,\epsilon^2\tau)V(\beta+\tau,\epsilon^2\tau)), \end{aligned} \tag{T}$$

where

$$P(\beta,\tau) = P^+(\beta,\tau) + P^-(\beta,\tau).$$

We remark that the initial data for the modulation equations is determined from initial conditions for the original system in ways described in section 6. Also, this hierarchy of higher order modulation equations is nearly identical to that derived in Wayne and Wright [33] for the Boussinesq equation—the chief difference lying in the specific forms of the inhomogeneous terms $J^\pm$.

To enforce the notion of spatial localization, we will be considering initial data which is of rapid decay, that is, initial data in

$$H^s(m) = \left\{ f(\alpha) | (1+\alpha^2)^{m/2} f(\alpha) \in H^s \right\}.$$

The inner product on $H^s(m)$ is given by

$$(f(\cdot), g(\cdot))_{H^s(m)} = \left((1+\cdot^2)^{m/2} f(\cdot), (1+\cdot^2)^{m/2} g(\cdot) \right)_{H^s},$$

where we use the standard inner product in H^s. In particular, the known soliton solutions of the KdV equations are in such spaces.

That the KdV equations have solutions for all times with this sort of initial data is well known. In particular, we have the following theorem from [28].

THEOREM 1. *Let* $\sigma \geq 4$. *Then for all* $C_I, T_0 > 0$ *there exists* $C_1 > 0$ *such that if* U, V *satisfy* (KdV) *with initial conditions* U_0, V_0 *and*

$$\max\{\|U_0\|_{H^\sigma(4)\cap H^{\sigma+4}(2)\cap H^{\sigma+9}}, \|V_0\|_{H^\sigma(4)\cap H^{\sigma+4}(2)\cap H^{\sigma+9}}\} < C_I, \tag{2}$$

then

$$\sup_{T\in[0,T_0]} \left\{ \|U(\cdot,T)\|_{H^\sigma(4)\cap H^{\sigma+4}(2)\cap H^{\sigma+8}}, \|V(\cdot,T)\|_{H^\sigma(4)\cap H^{\sigma+4}(2)\cap H^{\sigma+8}} \right\} < C_1. \tag{3}$$

On the other hand, it is less clear that solutions of (IW) and (LK) will remain bounded over the very long time scales necessary for the KdV approximation. In [33] we proved the following result which guarantees that the solutions of the modulation equations remain bounded for sufficiently long times.

PROPOSITION 1. *Fix* $T_0 > 0$ *and* $\sigma > 11/2$. *Suppose that* U_0, V_0 *satisfy* (2) *and* U, V, $P^\pm$, F, *and* G *satisfy* (KdV), (T), *and* (LK). *Then there exists a constant* C_2, *independent of* ϵ, *such that the solutions of* (T) *and* (LK) *satisfy the estimates*

$$\sup_{\tau\in[0,T_0\epsilon^{-2}]} \|P^\pm(\cdot,\tau)\|_{H^{\sigma+3}} \leq C_2,$$

$$\sup_{T\in[0,T_0]} \left\{ \|F(\cdot,T)\|_{H^\sigma\cap H^{\sigma-4}(2)}, \|G(\cdot,T)\|_{H^\sigma\cap H^{\sigma-4}(2)} \right\} \leq C_2.$$

Moreover, $P^{\pm}(\beta,\tau) = \varphi^{\pm}(\beta_{\pm}, T)$ *with*

$$\sup_{T\in[0,T_0]} \|\varphi^{\pm}(\cdot,T)\|_{H^{\sigma+3}\cap H^{\sigma-1}(2)} \leq C_2.$$

Finally we note that since $\partial_\tau P = \partial_\beta P^+ - \partial_\beta P^-$ we have $\|\partial_\tau P\|_s \leq \|\partial_\beta P\|_s$.

With this preliminary result in hand we can state our principal results. Denote the sum of the modulation functions, properly scaled, as

$$\begin{aligned} -\psi^d(\alpha,t) &= \epsilon^2 U\left(\epsilon(\alpha-t),\epsilon^3 t\right) + \epsilon^2 V\left(\epsilon(\alpha+t),\epsilon^3 t\right) \\ &\quad + \epsilon^4 F\left(\epsilon(\alpha-t),\epsilon^3 t\right) + \epsilon^4 G\left(\epsilon(\alpha+t),\epsilon^3 t\right) \\ &\quad + \epsilon^4 P(\epsilon\alpha,\epsilon t). \end{aligned} \tag{4}$$

As mentioned earlier, (x,y) are not the natural coordinates to study solutions to (WW). The coordinates we use are $(x_\alpha, y, x_t)^{tr}$. x_α is approximated by ψ^d, and the functions y and x_t are approximated by functions we denote ψ^y and ψ^u, respectively. They are given by

$$\begin{aligned} \psi^y(\alpha,t) &= \epsilon^2 U(\epsilon(\alpha-t),\epsilon^3 t) + \epsilon^2 V(\epsilon(\alpha+t),\epsilon^3 t) \\ &\quad + \epsilon^4 F(\epsilon(\alpha-t),\epsilon^3 t) + \epsilon^4 G(\epsilon(\alpha+t),\epsilon^3 t) + \epsilon^4 P(\epsilon\alpha,\epsilon t) \\ &\quad + \frac{1}{3}\epsilon^4 \partial_{\beta_-}^2 U(\epsilon(\alpha-t),\epsilon^3 t) + \frac{1}{3}\epsilon^4 \partial_{\beta_+}^2 V(\epsilon(\alpha+t),\epsilon^3 t) \\ &\quad + \epsilon^4 \left(U(\epsilon(\alpha-t),\epsilon^3 t) + V(\epsilon(\alpha+t),\epsilon^3 t)\right)^2 \end{aligned} \tag{5}$$

and

$$\begin{aligned} \psi^u(\alpha,t) &= \epsilon^2 U(\epsilon(\alpha-t),\epsilon^3 t) - \epsilon^2 V(\epsilon(\alpha+t),\epsilon^3 t) \\ &\quad + \epsilon^4 F(\epsilon(\alpha-t),\epsilon^3 t) - \epsilon^4 G(\epsilon(\alpha+t),\epsilon^3 t) \\ &\quad + \epsilon^4 \varphi^-(\epsilon(\alpha-t),\epsilon^3 t) - \epsilon^4 \varphi^+(\epsilon(\alpha+t),\epsilon^3 t) \\ &\quad + \frac{1}{6}\epsilon^4 \partial_{\beta_-}^2 U(\epsilon(\alpha-t),\epsilon^3 t) - \frac{1}{6}\epsilon^4 \partial_{\beta_+}^2 V(\epsilon(\alpha+t),\epsilon^3 t) \\ &\quad + \frac{3}{4}\epsilon^4 U^2(\epsilon(\alpha-t),\epsilon^3 t) - \frac{3}{4}\epsilon^4 V^2(\epsilon(\alpha+t),\epsilon^3 t). \end{aligned} \tag{6}$$

We discuss the origin of these equations in section 4.

The approximation will be valid in the space

$$\mathfrak{H}^s = H^s \times H^s \times H^{s-1/2}.$$

Our main result is the following.

THEOREM 2. *Fix* T_0, $C_I > 0$, $s > 4$, $\sigma \geq s+7$. *Suppose that* U, V, P, F, *and* G *satisfy* (KdV), (IW), *and* (LK) *and that* ψ^d, ψ^y, *and* ψ^u *are the combinations of these functions given in* (4), (5), *and* (6). *Then there exist* $\epsilon_0 > 0$ *and* $C_F > 0$ *such that the following is true. If the initial conditions for* (WW) *are of the form* $((x_\alpha(\alpha,0), y(\alpha,0), x_t(\alpha,0))^{tr} = \left(0, \epsilon^2\Theta_y(\epsilon\alpha), \epsilon^2\Theta_u(\epsilon\alpha)\right)^{tr}$ *with*

$$\max_{i=y,z}\left\{\|\Theta_i(\cdot)\|_{H^\sigma(4)\cap H^{\sigma+4}(2)\cap H^{\sigma+9}}\right\} \leq C_I,$$

then for $\epsilon \in (0,\epsilon_0)$ *there is a reparameterization of the free surface such that the unique solution to* (WW) *satisfies*

$$\left\| \begin{pmatrix} x_\alpha(\cdot,t) \\ y(\cdot,t) \\ x_t(\cdot,t) \end{pmatrix} - \begin{pmatrix} \psi^d(\cdot,t) \\ \psi^y(\cdot,t) \\ \psi^u(\cdot,t) \end{pmatrix} \right\|_{\mathfrak{H}^s} \leq C_F \epsilon^{11/2}$$

for $t \in [0, T_0\epsilon^{-3}]$*. The constant* C_F *does not depend on* ϵ*.*

Remark 1. The loss of the one-half power of ϵ in Theorem 2 is caused by the long wave scaling and not a lack of sharpness in the estimates.

Remark 2. It is clear that the form of the initial conditions specified in the hypotheses of this theorem does not agree with that found by setting $t = 0$ in the approximation inequality (unless, of course, $\psi^d \equiv 0$). This is precisely why we mention the need to reparameterize the free surface. We discuss this at length in section 6.

Remark 3. In [28], Schneider and Wayne prove the analogous result for the KdV (i.e., $O(\epsilon^4)$) approximation alone. The methods used here follow largely from their work.

Remark 4. The long time scale on which this theorem holds is natural given the form of the approximation—$t \sim O(\epsilon^{-3})$ corresponds to times T of $O(1)$ in the KdV equation. The methods of the proof rely on a Gronwall-type estimate which imposes this time scale. Moreover, while Proposition 1 shows that the functions F and G remain $O(1)$ for $t \sim O(\epsilon^{-3})$, estimates show that beyond this they are likely to grow very rapidly, ruining any chance for the higher order correction to be valid for longer times. Nonetheless, it is unclear whether or not the approximation of water waves by the KdV solutions alone is valid for even longer times. Numerical experiments done for the KdV approximation to the Euler–Poisson equations by Haragus, Nicholls, and Sattinger in [15] indicate that the approximation is valid for very long times, but little is known analytically.

Less technically, this theorem states that solutions to (WW), in the long wavelength limit, satisfy

$$x_\alpha(\alpha, t) = \psi^d(\alpha, t) + O(\epsilon^6)$$

for times of $O(\epsilon^{-3})$. This is, as expected, a marked improvement over the use of KdV alone.

We note that this is not the first time that linearized KdV equations have been put forward as a means to improve the accuracy of the KdV approximation. Other instances where linearized KdV equations appear include Sachs [26], Haragus, Nicholls, and Sattinger [15], Kodama and Taniuti [21], and Drazin and Johnson [12]. Moreover, there have been numerous models put forward over the years which model water waves in the same scaling regime we are considering. We refer the reader to Kodama [20], Olver [24], Bona, Pritchard, and Scott [5], Craig and Groves [11], Dullin, Gottwald, and Holm [13], and Bona and Chen [4]. Much of the work done in the above papers pertains to analyzing the behavior of the model equations and not to their connection to the original system. A notable exception is the recent work by Bona, Colin, and Lannes [3], wherein they prove the rigorous validity of a large number of Boussinesq-style models. There, they prove that such models are at least as accurate as the KdV approximation alone (though not as accurate as the model described here) and for the same time scales. Their models are designed, in part, to be easy to approximate numerically and are not, in general, explicitly solvable. Our particular combination of linearized KdV equations with an inhomogeneous wave equation appears to be unique and is asymptotically the most accurate model for long wavelength solutions to the water wave equation which has currently been justified rigorously.

The remainder of this paper is organized as follows. First, in section 2, we conduct a preliminary discussion of the water wave equation. Sections 3 and 5 contain a thorough discussion of the operator $K(x, y)$. Then, in section 4 we derive the higher order modulation equations and prove an important estimate. In section 6 we prove

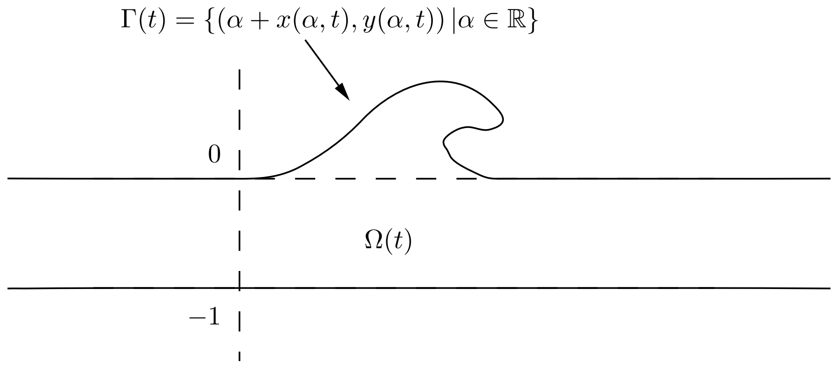


FIG. 6. *The water wave in Lagrangian coordinates.*

the validity of the approximation, i.e., Theorem 2. Finally, section 7 contains the details for a number of proofs.

2. Preliminaries. We begin by discussing the water wave problem in greater detail. Consider an infinitely long canal of unit mean depth in two dimensions (see Figure 6). We denote the region occupied by the fluid at time t as $\Omega(t)$, and the upper surface as $\Gamma(t)$. We parameterize $\Gamma(t)$ by $(\tilde{x}(\alpha,t), y(\alpha,t))$, where $\alpha \in \mathbb{R}$ is the parameter and $\tilde{x}$ and y are the real-valued coordinate functions. It is useful to break $\tilde{x}$ up as follows:

$$\tilde{x}(\alpha,t) = \alpha + x(\alpha,t).$$

We consider fluids which are inviscid and incompressible and flows which are irrotational. Also, we assume that the pressure on the top surface is constant and that the acceleration due to gravity is 1. With these assumptions, the evolution of x and y are given by the equations

$$\begin{aligned} x_{tt}(1+x_\alpha) + y_\alpha(1+y_{tt}) &= 0, \\ y_t &= K(x,y)x_t. \end{aligned} \tag{WW}$$

(See [10].)

The first of these two equations is found from Euler's equations for fluid motion. The operator K in the second line is a transformation which is linear in x_t, but depends nonlinearly on (x,y). That such an operator exists and gives a relationship between x_t and y_t is discussed in sections 3 and 5, along with an analysis of K. Much of the difficulty in answering questions about the water wave equation is related to this operator. If the surface of the water is perfectly flat, i.e., $x = y = 0$, then we have $K(0,0) = K_0$, where K_0 is a linear operator defined by $\widehat{K_0 f}(k) = \widehat{K_0}(k)\widehat{f}(k)$, with $\widehat{K_0}(k) = -i\tanh(k)$. Notice that since tanh is a bounded function, K_0 is bounded from H^s to H^s. K_0 will be appearing frequently.

This formulation of the water wave problem is said to be in *Lagrangian*, or *material*, coordinates. In this point of view we are not in a fixed "lab" frame, but instead we are tracking the position of each "particle" of water separately. That is, $(\alpha + x(\alpha,t), y(\alpha,t))$ gives the location of the particle which was initially at $(\alpha + x(\alpha,0), y(\alpha,0))$. The *laboratory*, or *Eulerian*, point of view is to fix a system of coordinates on the fluid domain and to measure the velocity of the fluid at each point of this fixed reference frame. For our purposes, it is far more convenient to use Lagrangian coordinates; however, experimentalists work with Eulerian coordinates. The coordinate change from one frame to the other is nontrivial and adapting our approximation results to the Eulerian point of view requires some involved computations. We will carry the change of variables out and give formulae for the approximation in terms of Eulerian coordinates in a future publication. The interested reader may also see the author's thesis [34] for this information. We remark that the results of Bona, Colin, and Lannes on the approximation of water waves by Boussinesq models in [3] are proven in Eulerian coordinates, and as such do not contend with this issue.

Since the water wave equation is second order in time for both x and y, one might suppose that four functions are necessary to specify the initial state of the system—$x(\alpha,0)$, $y(\alpha,0)$, $x_t(\alpha,0)$, and $y_t(\alpha,0)$. In fact, in general only three are needed, and for initial data with small amplitudes, only two are needed. The relationship $y_t = K(x,y)x_t$ specifies the value of $y_t(\alpha,0)$ given the other three functions. We can also "do away" with the initial condition for x, provided we are in the small amplitude, long wave limit. If $x(\alpha,0)$ and its first derivative are sufficiently small, then $\alpha + x(\alpha,0)$ will be invertible. This implies that $\Gamma(0)$ will be a graph over the horizontal coordinate. And so, without loss of generality, we can reparameterize the initial conditions so that

$$\Gamma(0) = \{(\alpha, y(\alpha,0))|\alpha \in \mathbb{R}\}.$$

Thus we need only to choose $y(\alpha,0)$ and $x_t(\alpha,0)$. As it turns out, we need to reparameterize the system one more time to prove the approximation theorem, but we will leave this technicality until section 6. The essential point here is that due to the freedom in choosing the initial parameterization, we can eliminate two of our initial conditions.

Even though we can assume that $x(\alpha,0) = 0$ (or, alternately, is small), this coordinate grows linearly in time. (See the linear estimates in Chapter 2 of [28].) As we are concerned with very long time scales, this is a problem. As was shown in [28], one can replace x with the new coordinate $z = K_0 x$, which is well behaved over long times. Rewriting (WW) with this new variable, we have

$$\begin{aligned}
\partial_t z &= K_0 u,\\
\partial_t y &= K(z,y)u,\\
\partial_t u &= -\frac{\partial_\alpha y(1+\partial_t^2 y)}{1+K_0^{-1}\partial_\alpha z}.
\end{aligned} \tag{WW3}$$

We will see that the operator $K(x,y)$ in truth depends not on x but on z, so the abuse of the notation $K(z,y)$ above is in some sense legitimate (see section 3 for further discussion). Furthermore, even though K_0^{-1} is not well defined, as it blows up at frequency $k=0$, the composition

$$L = K_0^{-1}\partial_\alpha$$

is well defined, as its symbol, $-k/\tanh(k)$, has no singularities; L is also invertible. Finally, we notice that the Maclaurin expansion of $\widehat{K_0}(k) = -ik + O(k^3)$. Thus to lowest order $K_0 \sim -\partial_\alpha$, and so $z \sim -x_\alpha$. This is precisely the reason why, in the introduction, we stated that $-x_\alpha$ is a natural coordinate for the water wave equation.

Though we will primarily be working with the three-dimensional system (WW3), we will need to embed this system into a four-dimensional system to prove certain aspects of the validity of the approximation. We introduce the new coordinate $a = u_t$, and (WW3) becomes

$$\text{(WW4)} \qquad \begin{aligned} \partial_t z &= K_0 u, \\ \partial_t y &= K(z,y)u, \\ \partial_t u &= a, \\ \partial_t a &= -\frac{a\partial_\alpha u + \partial_\alpha \partial_t (1+\partial_t^2 y) + \partial_t y \partial_t^3 y}{1 + K_0^{-1}\partial_\alpha z}. \end{aligned}$$

Though things appear to be getting out of hand, we remark that this is as large a system as we will need. Results in [28] prove that solutions to (WW), (WW3), and (WW4) do indeed exist for long times. We will be considering solutions to the four-dimensional system which are in

$$\mathfrak{H}_e^s = H^s \times H^s \times H^{s-1/2} \times H^{s-1}.$$

The main goal of this paper is to prove Theorem 2. To do this, we first prove a similar theorem for solutions to (WW3), from which Theorem 2 will follow. Let

$$\Psi^d(\alpha,t) = \psi^d(\alpha,t) + \epsilon^6 W_3(\epsilon\alpha, \epsilon t). \tag{7}$$

The additional function W_3 will solve an equation we specify later. While we assure the reader that we will be following Chekhov's rule and that this gun which appears in the first act will be fired in the third, interested parties may look ahead to equation (23) in section 4 for more information about W_3. Define the functions

$$\begin{aligned} \Psi^z =& L^{-1}\Psi^d, \\ \Psi^y =& \Psi^z + \epsilon^4 \Delta_1 + \epsilon^6 \Delta_2, \\ \Psi^u =& \partial_\alpha^{-1}\partial_t \Psi^d. \end{aligned}$$

Δ_1 and Δ_2 are combinations of solutions to the modulation equations and are given in (17) and (19) in section 4. We justify the presence of the inverse derivative in Ψ^u by means of example. Ψ^d contains the term U which solves a KdV equation. Thus $\partial_t \Psi^d$ contains the terms from the right-hand side of the KdV equation, all of which are perfect space derivatives. These then are cancelled by the inverse derivative. It is simple to check that this method applies to all terms in Ψ^d.

With these function, we have the following results.

THEOREM 3. *Fix T_0, $C_I > 0$, $s > 4$, $\sigma \geq s + 7$. Let Ψ^d, Ψ^z, Ψ^y, and Ψ^u be as above. Moreover assume that the initial conditions for* (KdV) *satisfy*

$$\max\{\|U_0\|_{H^\sigma(4)\cap H^{\sigma+4}(2)\cap H^{\sigma+9}}, \|V_0\|_{H^\sigma(4)\cap H^{\sigma+4}(2)\cap H^{\sigma+9}}\} < C_I.$$

Then there exist $\epsilon_0 > 0$ and $C_F = C_F(T_0, C_I, s) > 0$ such that if the initial conditions for (WW3) *are of the form*

$$\begin{pmatrix} z(\alpha,0) \\ y(\alpha,0) \\ u(\alpha,0) \end{pmatrix} = \begin{pmatrix} \Psi^z(\alpha,0) \\ \Psi^y(\alpha,0) \\ \Psi^u(\alpha,0) \end{pmatrix} + \epsilon^{11/2}\bar{R}_0(\alpha)$$

with $\|\bar{R}_0\|_{\mathfrak{H}^s} \leq C_I$, then the unique solution of (WW3) *satisfies the estimate*

$$\left\| \begin{pmatrix} z(\cdot,t) \\ y(\cdot,t) \\ u(\cdot,t) \end{pmatrix} - \begin{pmatrix} \Psi^z(\cdot,t) \\ \Psi^y(\cdot,t) \\ \Psi^u(\cdot,t) \end{pmatrix} \right\|_{\mathfrak{H}^s} \leq C_F \epsilon^{11/2}$$

for $t \in [0, T_0\epsilon^{-3}]$. The constant C_F does not depend on ϵ.

3. The operator ( ). Part I: Basics and basic expansions. This is the first of two sections where we discuss the operator $K(x,y)$ which gives the relation between y_t and x_t in the water wave equation. Here we briefly discuss the origin of this operator and report expansions of K found in previous work. We also state some very basic facts about these expansions. In section 5 we quote more complicated results and prove some new technical extensions needed for our purposes.

$K(x,y)$ (or, more precisely $-K(x,y)$) is sometimes called the Hilbert transform for the region $\Omega(t)$. Loosely, given a region Ω in the complex plane, and any function F which is analytic in Ω, the Hilbert transform for Ω, $H(\Omega)$, is a linear operator which relates the real and imaginary parts of F on the boundary of Ω. That is,

$$\mathrm{Im}(F)|_{\partial\Omega} = H(\Omega)\mathrm{Re}(F)|_{\partial\Omega}.$$

For example, if Ω were the lower half-plane, then the Hilbert transform would be the operator H, given by $\widehat{Hf} = i\ \mathrm{sgn}(k)\widehat{f}$. (This particular operator H is also frequently called the Hilbert transform.) The nature of the operator depends greatly on the region being studied. Unsurprisingly, the proof that such an operator exists is connected to the Riemann mapping theorem and to techniques for solving boundary value problems for Laplace's equation in the plane. In this problem, since the region $\Omega(t)$ is completely specified by the coordinate functions x and y, we denote the Hilbert transform by $H(\Omega(t)) = -K(x,y)$.

As we are considering a fluid which is incompressible and a flow which is irrotational, $x_t(\alpha,t) - iy_t(\alpha,t)$ is the value, on the upper boundary $\Gamma(t)$, of an analytic function on $\Omega(t)$, $\omega(\alpha+i\beta) = v^x(\alpha,\beta) - iv^y(\alpha,\beta)$. Here (v^x, v^y) is the velocity field for the fluid in the whole region. Thus, given that $K(x,y)$ exists, we have

$$y_t = K(x,y)x_t.$$

Of course, the boundary of $\Omega(t)$ is not just $\Gamma(t)$, but also includes the bottom of the canal (i.e., where $\beta = -1$). As we do not have fluid flow through the bottom, we have $v^y(\alpha,-1) = 0$.

Under these conditions, $K(x,y)$ has been analyzed extensively by Craig in [10] and Schneider and Wayne in [28]. In particular Craig shows that $K(x,y)$ has the following expansion:

$$K(x,y)u = K_0u + K_1(x,y)u + S_2(x,y)u, \tag{8}$$

where

$$\widehat{K_0u}(k) = -i\tanh(k)\widehat{u}(k),$$

$$K_1(x,y)u = [x,K_0]\partial_\alpha u - (y + K_0(yK_0))\partial_\alpha u,$$

and S_2 is quadratic in (x,y).

First of all we note that K_0 is a bounded operator from H^s to H^s since $\tanh(k)$ is a bounded function. That is,

$$\|K_0u\|_s \leq \|u\|_s.$$

The operator $L = K_0^{-1}\partial_\alpha$, which is well defined as we discussed in section 2, will also be used frequently. L is not a bounded operator on H^s. It effectively takes one derivative. That is,

$$\|Lu\|_s \leq \|u\|_{s+1}. \tag{9}$$

On the other hand, L^{-1} replaces one derivative. That is, since

$$|\tanh(k)/k| \leq (1+k^2)^{-1/2},$$

we know

$$\|L^{-1}u\|_s \leq \|u\|_{s-1}. \tag{10}$$

We will be considering functions which are of long wavelength, that is, functions of the form $f(\alpha) = F(\beta)$ where $\beta = \epsilon\alpha$. We define operators $K_{0,\epsilon}$ and L_ϵ via

$$\begin{aligned} K_{0,\epsilon}F(\beta) &= K_0 f(\alpha), \\ L_\epsilon F(\beta) &= Lf(\alpha). \end{aligned}$$

Taking the Maclaurin series expansion for $\tanh(k)$ shows that formally

$$K_{0,\epsilon} = -\epsilon\partial_\beta - \frac{1}{3}\epsilon^3\partial_\beta^3 - \frac{2}{15}\epsilon^5\partial_\beta^5 + O(\epsilon^7).$$

Similarly L_ϵ and L_ϵ^{-1} have expansions in terms of derivatives

$$\begin{aligned} L_\epsilon &= -1 + \frac{1}{3}\epsilon^2\partial_\beta^2 + \frac{1}{45}\epsilon^4\partial_\beta^4 + O(\epsilon^6), \\ L_\epsilon^{-1} &= -1 - \frac{1}{3}\epsilon^2\partial_\beta^2 - \frac{2}{15}\epsilon^4\partial_\beta^4 + O(\epsilon^6). \end{aligned}$$

We call such expansions of Fourier multiplier operators *long wave approximations*. The rigorous connection between a long wave approximation and the original operator is given in the following lemma, whose simple proof is contained in section 7.

LEMMA 1. *Suppose A and A_n are linear operators defined by $\widehat{Af}(k) = \widehat{A}(k)\widehat{f}(k)$ and $\widehat{A_n f}(k) = \widehat{A_n}(k)\widehat{f}(k)$, where $\widehat{A}(k)$ and $\widehat{A_n}(k)$ are complex-valued functions. Also suppose that $|\widehat{A}(k) - \widehat{A_n}(k)| \leq C|k|^n$ (e.g., $\widehat{A_n}$ is a Taylor polynomial for $\widehat{A}$). Then for $f \in H^{s+n}$ we have*

$$\|Af(\cdot) - A_n f(\cdot)\|_s \leq C\|\partial_\alpha^n f(\cdot)\|_s.$$

Moreover, if $f(\alpha)$ is of long wavelength form—that is, if $f(\alpha) = F(\epsilon\alpha)$, with $F \in H^{s+n}$—then for $0 < \epsilon < 1$ there exists C independent of ϵ such that

$$\|Af(\cdot) - A_n f(\cdot)\|_s \leq C\epsilon^{n-1/2}\|F(\cdot)\|_{s+n}.$$

In [28], Schneider and Wayne show that the operator $K(x,y)$ does not depend on x per se, but rather on $z = K_0 x$. We confuse the notation for the operators intentionally. They showed that $K(z,y)$ has the following expansion:

$$K(z,y)u = K_0 u + K_1(z,y)u + S_2(z,y)u,$$

where

$$K_1(z,y)u = M_1(z)\partial_\alpha u - (y + K_0 y K_0)\partial_\alpha u$$

with

$$\mathfrak{F}[M_1(z)v](k) = -\int \frac{\widehat{K_0}(k) - \widehat{K_0}(l)}{\widehat{K_0}(k-l)}\widehat{z}(k-l)\widehat{v}(l)dl.$$

S_2 is an operator which depends quadratically on z and y. Section 5 contains an analysis of these operators.

By using the hyperbolic trigonometric identity

$$\frac{\tanh(l) - \tanh(k)}{\tanh(l-k)} = 1 - \tanh(k)\tanh(l), \tag{11}$$

we can simplify the expression for K_1 to

$$\begin{aligned} K_1(z,y)u &= M_1(z+y)\partial_\alpha u \\ &= -\left(z + y + K_0(z+y)K_0\right)\partial_\alpha u. \end{aligned} \tag{12}$$

Since we know K_0 is a bounded operator, it is clear that

$$\|M_1(z)v\|_s \le \|zv\|_s. \tag{13}$$

4. The derivation. In this section we will derive the higher order correction to the KdV approximation. For technical reasons, it is most convenient to work with the water wave equation written in the form (WW3). Suppose that one is given the function $\Psi(\alpha,t) = (\Psi^z(\alpha,t), \Psi^y(\alpha,t), \Psi^u(\alpha,t))^{tr}$. The amount that this function fails to satisfy (WW3) is called the *residual* and is given by $\mathrm{Res}[\Psi] = (\mathrm{Res}_z, \mathrm{Res}_y, \mathrm{Res}_u)^{tr}$ with

$$\begin{aligned} \mathrm{Res}_z &= \partial_t \Psi^z - K_0 \Psi^u, \\ \mathrm{Res}_y &= \partial_t \Psi^y - K(\Psi^z, \Psi^y)\Psi^u, \\ \mathrm{Res}_u &= \partial_t \Psi^u + \partial_\alpha \Psi^y \frac{1 + \partial_t^2 \Psi^y}{1 + L\Psi^z}. \end{aligned}$$

For a true solution, notice that $\mathrm{Res}[\Psi]$ is identically zero.

Remark 5. We will also consider the four-dimensional system (WW4). If we let $\Psi^a = \partial_t \Psi^u$, then we have the additional Res function

$$\begin{aligned} \mathrm{Res}_a &= \partial_t \Psi^a + \frac{\Psi^a \partial_\alpha \Psi^u + \partial_\alpha \partial_t (1 + \partial_t^2 \Psi^y) + \partial_t \Psi^y \partial_t^3 \Psi^y}{1 + L\Psi^z} \\ &= \partial_t \mathrm{Res}_u + \frac{\partial_\alpha \Psi^u}{1 + L\Psi^z} \mathrm{Res}_u. \end{aligned}$$

The main goal when deriving modulation equations is to choose a system of equations such that solutions to this system yield a very small residual. This is different from (but connected to) showing that solutions to the modulation equations are close to true solutions for the original problem. This latter issue is precisely that answered by the main results, Theorem 2 and Proposition 3, and is discussed in section 6. Here, we will perform a series of calculations on the residual and derive equations (KdV), (IW), and (LK). In this process we guarantee the smallness of the residual. While several of the steps will initially seem to have little mathematical justification (i.e., they are formal), once the calculation is completed it will be obvious that all steps are valid. For example, we will take

$$\Psi^u = K_0^{-1} \partial_t \Psi^z. \tag{14}$$

With this choice

$$\mathrm{Res}_z = 0,$$

which is small indeed! However, K_0^{-1} is not in general a well-defined operator. Nonetheless, when we eventually select Ψ^z, $K_0^{-1}\partial_t\Psi^z$ will make perfect sense.

We are looking for solutions which are small in amplitude and long in wavelength. Thus we let

$$\begin{aligned}\Psi^z(\alpha,t) &= \epsilon^2 Z(\beta,\tau)\\ &= \epsilon^2 Z_1(\beta,\tau)+\epsilon^4 Z_2(\beta,\tau)+\epsilon^6 Z_3(\beta,\tau).\end{aligned}$$

Recall $\beta=\epsilon\alpha$ and $\tau=\epsilon t$. We require $\mathrm{Res}[\Psi]$ to be $O(\epsilon^{17/2})$. Loosely, we need three powers of ϵ more than the expected error of $O(\epsilon^{11/2})$ to account for the long times $(O(\epsilon^{-3}))$ over which our approximation will be valid. See Schneider and Wayne [28] and Wayne and Wright [33].

Remark 6. More specifically, if we wish to prove Theorem 2 in the space $\mathfrak{H}^s$ we will need

$$\begin{aligned}\|\mathrm{Res}_z\|_s &\le C\epsilon^{17/2},\\ \|\mathrm{Res}_y\|_s &\le C\epsilon^{17/2},\\ \|\mathrm{Res}_u\|_{s-1} &\le C\epsilon^{17/2},\\ \|\mathrm{Res}_a\|_{s-1} &\le C\epsilon^{19/2}\end{aligned}$$

for $0\le t\le T_0\epsilon^{-3}$. Given the definition of Res_a, the final estimate will follow automatically from the estimate on Res_u.

We have already chosen Ψ^u in terms of Ψ^z. We will first use the expression for Res_y to similarly determine Ψ^y in terms of Ψ^z. This is not as simple a matter because while K_0 commutes with ∂_t, the full operator $K(z,y)$ does not.

We have

$$\begin{aligned}\mathrm{Res}_y &= -\partial_t\Psi^y + K_0\Psi^u + M_1(\Psi^z+\Psi^y)\partial_\alpha\Psi^u + S_2(\Psi^y,\Psi^z)\Psi^u\\ &= -\partial_t\Psi^y + \partial_t\Psi^z + M_1(\Psi^z+\Psi^y)K_0^{-1}\partial_\alpha\partial_t\Psi^z + S_2(\Psi^y,\Psi^z)\Psi^u.\end{aligned}$$

Notice that in the above expression we can cancel the linear terms by taking $\Psi^y\sim\Psi^z$. More precisely, we set

$$\Psi^y(\alpha,t)=\epsilon^2 Z(\beta,\tau)+\epsilon^4\Delta_1(\beta,\tau)+\epsilon^6\Delta_2(\beta,\tau)$$

for as yet undetermined functions Δ_i. Thus

$$\begin{aligned}\mathrm{Res}_y &= -\epsilon^5\partial_\tau\Delta_1-\epsilon^7\partial_\tau\Delta_2+M_1(2\epsilon^2 Z)L_\epsilon\epsilon^3\partial_\tau Z\\ &\quad + M_1(\epsilon^4\Delta_1+\epsilon^6\Delta_2)L_\epsilon\epsilon^3\partial_\tau Z + S_2(\Psi^y,\Psi^z)\Psi^u\\ &= -\epsilon^5\partial_\tau\Delta_1-\epsilon^7\partial_\tau\Delta_2-2\epsilon^5 Z L_\epsilon\partial_\tau Z-\epsilon^6 K_{0,\epsilon}\left(2Z\partial_\beta\partial_\tau Z\right)\\ &\quad -\epsilon^7\Delta_1 L_\epsilon\partial_\tau Z-\epsilon^8 K_{0,\epsilon}\left(\Delta_1\partial_\beta\partial_\tau Z\right)+\epsilon^9 M_1(\Delta_2)L_\epsilon\partial_\tau Z\\ &\quad + S_2(\Psi^y,\Psi^z)\Psi^u.\end{aligned}$$

A number of the terms in Res_y are already $O(\epsilon^{17/2})$. By Lemma 1 we have the following estimate on $K_{0,\epsilon}$:

$$\|K_{0,\epsilon}F\|_s \le \epsilon^{1/2}\|F\|_{s+1}. \tag{15}$$

Thus terms containing $K_{0,\epsilon}$ can be considered to be a power of ϵ smaller than they appear (though this costs a derivative). On the other hand, K_0 is bounded so we have

$$\|K_{0,\epsilon}F\|_s \le \epsilon^{-1/2}\|F\|_s. \tag{16}$$

Thus we can use $K_{0,\epsilon}$ either as a bounded functional *or* to gain powers of ϵ, but not both. Notice that L_ϵ does not contribute any additional powers of ϵ in any case. We separate out all the terms that are already sufficiently small into error terms. That is,

$$\begin{aligned}\mathrm{Res}_y =& -\epsilon^5\partial_\tau\Delta_1 - \epsilon^7\partial_\tau\Delta_2 - 2\epsilon^5 Z_1 L_\epsilon\partial_\tau Z_1\\ &- 2\epsilon^7 Z_1 L_\epsilon \partial_\tau Z_2 - 2\epsilon^7 Z_2 L_\epsilon\partial_\tau Z_1\\ &- \epsilon^6 K_{0,\epsilon}\left(2Z_1\partial_\beta\partial_\tau Z_1\right) - \epsilon^7\Delta_1 L_\epsilon\partial_\tau Z_1\\ &+ E^y_{small} + E^y_{S_2}\end{aligned}$$

with

$$\begin{aligned}E^y_{small} =& 2\epsilon^9 Z_3 L_\epsilon\partial_\tau Z + 2\epsilon^9(Z_1+\epsilon^2 Z_2)L_\epsilon\partial_\tau Z_3 + 2\epsilon^9 Z_2 L_\epsilon\partial_\tau Z_2\\ &- \epsilon^8 K_{0,\epsilon}\big((2Z_2+2\epsilon^2 Z_3)\partial_\beta\partial_\tau Z\big) - \epsilon^8 K_{0,\epsilon}\big(2Z_1\partial_\beta\partial_\tau(Z_2+\epsilon^2 Z_3)\big)\\ &- \epsilon^9\Delta_1 L_\epsilon\partial_\tau(Z_2+\epsilon^2 Z_3) - \epsilon^8 K_{0,\epsilon}\left(\Delta_1\partial_\beta\partial_\tau Z\right) + \epsilon^9 M_1(\Delta_2)L_\epsilon\partial_\tau Z\\ E^y_{S_2} =& S_2(\Psi^y,\Psi^z)\Psi^u.\end{aligned}$$

It is clear that E^y_{small} is $O(\epsilon^{17/2})$. That is,

$$\|E^y_{small}\|_s \le C\epsilon^{17/2}.$$

The constant C depends on various norms of the functions Z_i, $\partial_\tau Z_i$, and Δ_i. Specifically, chasing through the various terms in E^y_{small} and applying the estimates in (9), (13), (15), and (16), one can show that C depends on $\|Z_1\|_{s+1}$, $\|Z_2\|_{s+1}$, $\|Z_3\|_s$, $\|\partial_\tau Z_1\|_{s+2}$, $\|\partial_\tau Z_2\|_{s+2}$, $\|\partial_\tau Z_3\|_{s+1}$, $\|\Delta_1\|_{s+1}$, and $\|\Delta_2\|_s$.

The term $E^y_{S_2}$ is also $O(\epsilon^{17/2})$, though this is not as obvious. We prove this in Proposition 3 in section 5. The proof of this relies strongly on the fact that we have taken Ψ^y and Ψ^z such that $\Psi^y - \Psi^z$ is $O(\epsilon^4)$. This causes a cancellation in S_2, which in turn makes this term small.

We now expand L_ϵ and $K_{0,\epsilon}$ in the remaining low order terms in Res_y to find

$$\begin{aligned}\mathrm{Res}_y =& -\epsilon^5\partial_\tau\Delta_1 - \epsilon^7\partial_\tau\Delta_2 + 2\epsilon^5 Z_1\partial_\tau Z_1 + \frac{4}{3}\epsilon^7 Z_1\partial_\tau\partial_\beta^2 Z_1\\ &+ 2\epsilon^7 Z_1\partial_\tau Z_2 + 2\epsilon^7 Z_2\partial_\tau Z_1\\ &+ 2\epsilon^7\partial_\beta Z_1\partial_\beta\partial_\tau Z_1 + \epsilon^7\Delta_1\partial_\tau Z_1\\ &+ E^y_{small} + E^y_{S_2} + E^y_{lwa}\end{aligned}$$

with

$$\begin{aligned}E^y_{lwa} =& -2\epsilon^5 Z_1\left(L_\epsilon + 1 - \frac{1}{3}\epsilon^2\partial_\beta^2\right)\partial_\tau Z_1 - 2\epsilon^7 Z_1(L_\epsilon+1)\partial_\tau Z_2\\ &- 2\epsilon^7 Z_2(L_\epsilon+1)\partial_\tau Z_1 - 2\epsilon^6(K_{0,\epsilon}+\epsilon\partial_\beta)(Z_1\partial_\beta\partial_\tau Z_1)\\ &- \epsilon^7\Delta_1(L_\epsilon+1)\partial_\tau Z_1.\end{aligned}$$

Each term in E^y_{lwa} is $O(\epsilon^{17/2})$ by Lemma 1. That is,

$$\|E^y_{lwa}\| \le C\epsilon^{17/2},$$

where C depends on $\|Z_1\|_{s+3}$,$\|Z_2\|_s$, $\|\partial_\tau Z_1\|_{s+4}$,$\|\partial_\tau Z_2\|_{s+2}$, and $\|\Delta_1\|_s$. (The subscript "lwa" stands for "long wave approximation.")

The only $O(\epsilon^5)$ terms remaining in Res_y are

$$-\epsilon^5 \partial_\tau \Delta_1 + 2\epsilon^5 Z_1 \partial_\tau Z_1$$

which we remove by selecting

$$\Delta_1 = Z_1^2. \tag{17}$$

Therefore

$$\begin{aligned}\mathrm{Res}_y = &- \epsilon^7 \partial_\tau \Delta_2 + \frac{4}{3}\epsilon^7 Z_1 \partial_\tau \partial_\beta^2 Z_1 \\ &+ 2\epsilon^7 Z_1 \partial_\tau Z_2 + 2\epsilon^7 Z_2 \partial_\tau Z_1 \\ &+ 2\epsilon^7 \partial_\beta Z_1 \partial_\beta \partial_\tau Z_1 + \epsilon^7 Z_1^2 \partial_\tau Z_1 \\ &+ E^y_{small} + E^y_{S_2} + E^y_{lwa}.\end{aligned}$$

The remaining $O(\epsilon^7)$ terms in Res_y are all perfect time derivatives with the exception of

$$\frac{4}{3} Z_1 \partial_\tau \partial_\beta^2 Z_1.$$

Notice, however, that

$$\partial_\tau \left(\frac{4}{3} Z_1 \partial_\beta^2 Z_1 - \frac{2}{3} (\partial_\tau Z_1)^2 \right) = \frac{4}{3} Z_1 \partial_\tau \partial_\beta^2 Z_1 + \frac{4}{3} \partial_\tau Z_2 \left(\partial_\beta^2 Z_1 - \partial_\tau^2 Z_1 \right).$$

Given the form of the approximation in (1), it is not unreasonable to suspect that

$$\partial_\beta^2 Z_1 - \partial_\tau^2 Z_1 \sim O(\epsilon^2). \tag{18}$$

We are now in a position to select Δ_2. Taking

$$\Delta_2 = (\partial_\beta Z_1)^2 + 2 Z_1 Z_2 + \frac{1}{3} Z_1^3 + \frac{4}{3} Z_1 \partial_\beta^2 Z_1 - \frac{2}{3} (\partial_\tau Z_1)^2 \tag{19}$$

gives

$$\mathrm{Res}_y = E^y_{small} + E^y_{S_2} + E^y_{lwa} + E^y_{switch},$$

where

$$E^y_{switch} = \frac{4}{3} \epsilon^7 \partial_\tau Z_1 (\partial_\beta^2 Z_1 - \partial_\tau^2 Z_1).$$

Given that our assumption (18) is valid, we have shown, with our choices for Ψ^y and Ψ^u in terms of Ψ^z, that $\mathrm{Res}_y = O(\epsilon^{17/2})$. More specifically, we have shown that if

$$\|\partial_\beta^2 Z_1 - \partial_\tau^2 Z_1\|_s \le C\epsilon^{3/2},$$

then

$$\mathrm{Res}_y \le C\epsilon^{17/2},$$

where the constant depends only on $\|Z_1\|_{s+3}$, $\|Z_2\|_{s+1}$, $\|Z_3\|_s$, $\|\partial_\tau Z_1\|_{s+4}$, $\|\partial_\tau Z_2\|_{s+2}$, and $\|\partial_\tau Z_3\|_{s+1}$.

Now that we have computed Ψ^y and Ψ^u in terms of Ψ^z, we now turn our attention to determining Ψ^z by examining Res_u:

$$\mathrm{Res}_u = K_0^{-1}\partial_t^2\Psi^z + \partial_\alpha\Psi^y\frac{1+\partial_t^2\Psi^y}{1+L\Psi^z}.$$

We expand $(1+L\Psi^z)^{-1}$ by the geometric series to find

$$\mathrm{Res}_u = K_0^{-1}\partial_t^2\Psi^z + \partial_\alpha\Psi^y(1+\partial_t^2\Psi^y)(1-L\Psi^z+(L\Psi^z)^2) + E_1^u,$$

where

$$E_{geo}^u = \partial_\alpha\Psi^y\frac{1+\partial_t^2\Psi^y}{1+L\Psi^z} - \partial_\alpha\Psi^y(1+\partial_t^2\Psi^y)(1-L\Psi^z+(L\Psi^z)^2).$$

Since Ψ^z is "small" this error term can be shown to be $O(\epsilon^{17/2})$. We have the following lemma.

LEMMA 2. *Let $f \in H^{s+1}$, $s > 0$. Take ϵ_0 such that $\epsilon_0^2\|f\|_{L^\infty} \le 1/2$. Then for $0 < \epsilon < \epsilon_0$ we have that the function*

$$g(\epsilon x) = \frac{1}{1+\epsilon^2 f(\epsilon x)} - 1 + \epsilon^2 f(\epsilon x)$$

satisfies $\|g(\epsilon\cdot)\|_s \le C\epsilon^{7/2}$ for C independent of ϵ.

Remark 7. Under the same hypotheses as in Lemma 2, arguments similar to the proof of that lemma show that

- $(1+\epsilon^2 f(\epsilon x))^{-1} \in C^s$ and is bounded there independent of ϵ,
- $(1+\epsilon^2 f(\epsilon x))^{-1} - 1 \in H^s$ and has norm there bounded by $C\epsilon^{3/2}$ for C independent of ϵ, and
- $(1+\epsilon^2 f(\epsilon x))^{-1} - 1 + \epsilon^2 f(\epsilon x) - \epsilon^4 f^2(\epsilon x) \in H^s$ and has norm there bounded by $C\epsilon^{11/2}$ for C independent of ϵ.

Now, after substituting from the definitions of Ψ^y and Ψ^z, we collect all the terms which are smaller than $O(\epsilon^{17/2})$ and find

$$\begin{aligned}\mathrm{Res}_u = {}& \epsilon^4 K_{0,\epsilon}^{-1}\partial_\tau^2(Z_1+\epsilon^2 Z_2+\epsilon^4 Z_3) + \epsilon^3\partial_\beta Z_1 - \epsilon^5\partial_\beta Z_1 L_\epsilon Z_1\\ &+ \epsilon^5\partial_\beta\left(Z_2+Z_1^2\right) + \epsilon^7\partial_\beta Z_1\left(\partial_\beta^2 Z_1 - L_\epsilon Z_2 + (L_\epsilon Z_1)^2\right)\\ &- \epsilon^7\partial_\beta(Z_2+Z_1^2)L_\epsilon Z_1 + \epsilon^7\partial_\beta Z_3\\ &+ \epsilon^7\partial_\beta\left((\partial_\beta Z_1)^2 + 2Z_1Z_2 + \frac{1}{3}Z_1^3\right)\\ &+ \epsilon^7\partial_\beta\left(\frac{4}{3}Z_1\partial_\beta^2 Z_1 - \frac{2}{3}(\partial_\tau Z_1)^2\right)\\ &+ E_{geo}^u + E_{small}^u + E_{switch}^u.\end{aligned}$$

We omit the exact expression for E_{small}^u because it is both lengthy and uninteresting. We have

$$\|E_{small}^u\|_s \le C\epsilon^{17/2}, \tag{20}$$

where the constant C depends on H^{s+1} norms of the functions Z_i and the H^s norms of $\partial_\tau^2 Z_i$. We have also replaced one instance of $\partial_\tau^2 Z_1$ with $\partial_\beta^2 Z_1$ (much as we did earlier); thus

$$E_{switch}^u = \epsilon^7\partial_\beta Z_1\left(\partial_\beta^2 Z_1 - \partial_\tau^2 Z_1\right).$$

We now define new functions W_i by $Z_i = L_\epsilon^{-1} W_i$. This seemingly mysterious (and sudden!) change of variables will seem less so if we remind the reader that at the end of the day we wish to model not z but rather the function x_α. Accordingly, if we approximate x_α by a function

$$\Psi^d = \epsilon^2 W_1 + \epsilon^4 W_2 + \epsilon^6 W_3,$$

then it is logical to take

$$\Psi^z = L^{-1}\Psi^d,$$

and in the long wavelength limit we arrive at these functions W_i.

Therefore we have

$$\begin{aligned}
\mathrm{Res}_u = {} & \epsilon^3 \partial_\beta^{-1}\partial_\tau^2 (W_1 + \epsilon^2 W_2 + \epsilon^4 W_3) + \epsilon^2 K_{0,\epsilon} W_1 - \epsilon^4 W_1 K_{0,\epsilon} W_1 \\
& + \epsilon^4 K_{0,\epsilon} W_2 + \epsilon^5 \partial_\beta (L_\epsilon^{-1} W_1)^2 \\
& + \epsilon^2 K_{0,\epsilon} W_1 \left(\epsilon^3 K_{0,\epsilon} \partial_\beta W_1 - \epsilon^4 W_2 + \epsilon^4 W_1^2\right) \\
& - (\epsilon^6 K_{0,\epsilon} W_2 + \epsilon^7 \partial_\beta (L_\epsilon^{-1} W_1)^2) W_1 + \epsilon^6 K_{0,\epsilon} W_3 \\
& + \epsilon^5 \partial_\beta \left((K_{0,\epsilon} W_1)^2 + 2\epsilon^2 L_\epsilon^{-1} W_1 L_\epsilon^{-1} W_2 + \frac{1}{3}\epsilon^2 (L_\epsilon^{-1} W_1)^3 \right) \\
& + \epsilon^7 \partial_\beta \left(\frac{4}{3} L_\epsilon^{-1} W_1 L_\epsilon^{-1} \partial_\beta^2 W_1 - \frac{2}{3} (L_\epsilon^{-1} \partial_\tau W_1)^2 \right) \\
& + E^u_{geo} + E^u_{small} + E^u_{switch}.
\end{aligned}$$

At this time, the presence of inverse β derivatives may seem problematic. Notice that each such inverse derivative precedes a time derivative. Once we select the functions W_i we will see that there can be an exchange between time and space derivatives, which will justify the instances of ∂_β^{-1}.

Now we replace $K_{0,\epsilon}$ and L_ϵ^{-1} with their long wave approximates and find

$$\begin{aligned}
\mathrm{Res}_u = {} & \epsilon^3 \partial_\beta^{-1}\partial_\tau^2 (W_1 + \epsilon^2 W_2 + \epsilon^4 W_3) \\
& - \epsilon^2 \left(\epsilon\partial_\beta + \frac{1}{3}\epsilon^3 \partial_\beta^3 + \frac{2}{15}\epsilon^5 \partial_\beta^5 \right) W_1 \\
& + \epsilon^4 W_1 \left(\epsilon\partial_\beta + \frac{1}{3}\epsilon^3 \partial_\beta^3 \right) W_1 \\
& - \epsilon^4 \left(\epsilon\partial_\beta + \frac{1}{3}\epsilon^3 \partial_\beta^3 \right) W_2 + \epsilon^5 \partial_\beta (W_1)^2 + \frac{2}{3}\epsilon^7 \partial_\beta (W_1 \partial_\beta^2 W_1) \\
& - \epsilon^7 \partial_\beta W_1 \left(-\partial_\beta^2 W_1 - W_2 + W_1^2 \right) \\
& - \epsilon^7 W_1 (-\partial_\beta W_2 + \partial_\beta (W_1)^2) - \epsilon^7 \partial_\beta W_3 \\
& + \epsilon^7 \partial_\beta \left((\partial_\beta W_1)^2 + 2 W_1 W_2 - \frac{1}{3} W_1^3 \right) \\
& + \epsilon^7 \partial_\beta \left(\frac{4}{3} W_1 \partial_\beta^2 W_1 - \frac{2}{3} (\partial_\tau W_1)^2 \right) \\
& + E^u_{geo} + E^u_{small} + E^u_{switch} + E^u_{lwa}.
\end{aligned}$$

The error made by the long wave approximations is denoted E^u_{lwa}. By Lemma 1 we have

$$\|E^u_{lwa}\|_s \le C\epsilon^{17/2},$$

where C depends on $\|W_1\|_{s+7}$, $\|W_2\|_{s+5}$, $\|W_3\|_{s+3}$, and $\|\partial_\tau W_1\|_{s+2}$.

Now we organize the above as

$$
\begin{aligned}
\mathrm{Res}_y = {} & \epsilon^3 \partial_\beta^{-1}\partial_\tau^2 W_1 - \epsilon^3 \partial_\beta W_1 \\
& - \epsilon^5 \frac{1}{3}\partial_\beta^3 W_1 + \epsilon^5 \frac{3}{2}\partial_\beta (W_1)^2 \\
& + \epsilon^5 \partial_\beta^{-1}\partial_\tau^2 W_2 - \epsilon^5 \partial_\beta W_2 \\
& - \epsilon^7 \frac{1}{3}\partial_\beta^3 W_2 + \epsilon^7 3\partial_\beta (W_1 W_2) \\
& - \frac{2}{15}\epsilon^7 \partial_\beta^5 W_1 + \epsilon^7 \frac{1}{3} W_1 \partial_\beta^3 W_1 \\
& + \epsilon^7 2\partial_\beta (W_1 \partial_\beta^2 W_1) + \epsilon^7 \frac{3}{2}\partial_\beta (\partial_\beta W_1)^2 \\
& - \epsilon^7 \frac{4}{3}\partial_\beta (W_1)^3 - \epsilon^7 \frac{2}{3}\partial_\beta (\partial_\tau W_1)^2 \\
& + \epsilon^7 \partial_\beta^{-1}\partial_\tau^2 W_3 - \epsilon^7 \partial_\beta W_3 \\
& + E^u_{geo} + E^u_{small} + E^u_{switch} + E^u_{lwa}.
\end{aligned}
\tag{21}
$$

The term on the first line of the right-hand side looks formally like an inverse derivative of a wave equation:

$$\partial_\beta^{-1}\partial_\tau^2 W_1 - \partial_\beta W_1 = \partial_\beta^{-1}\left(\partial_\tau^2 W_1 - \partial_\beta^2 W_1\right).$$

We cancel this term (to lowest order) by taking W_1 of the form

$$W_1(\beta,\tau) = -U(\beta-\tau, \epsilon^2\tau) - V(\beta+\tau, \epsilon^2\tau).$$

Recall $\beta_\pm = \beta \pm \tau$ and $T = \epsilon^2 \tau$. The "minus" signs may seem arbitrary, but are included at this stage so that they agree with previous work in the area. Noting that the third line looks very much like the first, we also set

$$W_2(\beta,\tau) = -F(\beta-\tau, \epsilon^2\tau) - G(\beta+\tau, \epsilon^2\tau) - P(\beta,\tau).$$

These choices for W_1 and W_2 are precisely those described heuristically in the introduction.

The first three lines in (21) become

$$
\begin{aligned}
& \epsilon^5 \left(2\partial_T U + \frac{1}{3}\partial_{\beta_-}^3 U + \frac{3}{2}\partial_{\beta_-} U^2\right) \\
& + \epsilon^5 \left(-2\partial_T V + \frac{1}{3}\partial_{\beta_+}^3 V + \frac{3}{2}\partial_{\beta_+} V^2\right) \\
& + \epsilon^5 \left(\partial_\beta^{-1}\left(\partial_\beta^2 P - \partial_\tau^2 P\right) + 3\partial_\beta (UV)\right) \\
& + \epsilon^7 \left(2\partial_T F - 2\partial_T G - \partial_\beta^{-1}\left(\partial_T^2 U + \partial_T^2 V\right)\right) \\
& - \epsilon^9 \partial_\beta^{-1}\left(\partial_T^2 F + \partial_T^2 G\right).
\end{aligned}
$$

We cancel everything multiplied by ϵ^5 by taking

$$
\begin{aligned}
-2\partial_T U &= \frac{1}{3}\partial_{\beta_-}^3 U + \frac{3}{2}\partial_{\beta_-} U^2, \\
2\partial_T V &= \frac{1}{3}\partial_{\beta_+}^3 V + \frac{3}{2}\partial_{\beta_+} V^2, \\
\partial_\tau^2 P - \partial_\beta^2 P &= 3\partial_\beta^2 (UV),
\end{aligned}
$$

which are precisely equations (KdV) and (IW). By Proposition 1 we know that the solutions to these equations are well behaved over the long time scales.

Given that the functions U and V have been chosen to solve (KdV), one computes that

$$\partial_T^2 U = \partial_{\beta_-}\left(\frac{1}{36}\partial_{\beta_-}^5 U + \frac{9}{4}U^2\partial_{\beta_-}U + \frac{1}{2}U\partial_{\beta_-}^3 U + \frac{3}{4}\partial_{\beta_-}U\partial_{\beta_-}^2 U\right),$$
$$\partial_T^2 V = \partial_{\beta_+}\left(\frac{1}{36}\partial_{\beta_+}^5 V + \frac{9}{4}V^2\partial_{\beta_+}V + \frac{1}{2}V\partial_{\beta_+}^3 V + \frac{3}{4}\partial_{\beta_+}V\partial_{\beta_+}^2 V\right).$$

Thus the term $\partial_\beta^{-1}\left(\partial_T^2 U + \partial_T^2 V\right)$ is perfectly well defined. For brevity, we will continue to write these terms with the inverse derivatives instead of in the longer form above.

Moreover, now we can put more precise estimates on E^y_{switch} and E^u_{switch}. In particular, since each time derivative for solutions to KdV equations counts for three space derivatives, we have

$$\|\partial_\beta^2 W_1 - \partial_\tau^2 W_1\|_s \le C\epsilon^{3/2},$$

where C depends on $\|W_1\|_{s+6}$.

Recall from Fact 1 and Proposition 1 that solutions to (IW) can be rewritten as

$$\begin{aligned} P(\beta,\tau) &= P^+(\beta,\tau) + P^-(\beta,\tau) \\ &= \varphi^+(\beta_+,T) + \varphi^-(\beta_-,T). \end{aligned}$$

The functions $\varphi^\pm$ are rapidly decaying. We make this decomposition so that every remaining term in (21)

- will be a unidirectional term which is rapidly decaying;
- will be a product of two such terms which are moving in opposite directions;
- or will include a derivative of W_3.

That is,

$$\begin{aligned} \mathrm{Res}_y = {} & \epsilon^7\left(2\partial_T F + \frac{1}{3}\partial_{\beta_-}^3 F + 3\partial_{\beta_-}(UF) + J^-\right) \\ & + \epsilon^7\left(-2\partial_T G + \frac{1}{3}\partial_{\beta_+}^3 G + 3\partial_{\beta_+}(VG) + J^+\right) \\ & + \epsilon^7\left(\partial_\beta^{-1}\left(\partial_\tau^2 W_3 - \partial_\beta^2 W_3\right) + J^s\right) \\ & - \epsilon^9\partial_\beta^{-1}\left(\partial_T^2 F + \partial_T^2 G\right) \\ & + E^u_{geo} + E^u_{small} + E^u_{switch} + E^u_{lwa} + E^u_{time}, \end{aligned}$$

where

$$
\begin{aligned}
J^- = {} & 3\partial_{\beta_-}(U\varphi^-) + 4U^2\partial_{\beta_-}U + \frac{7}{3}U\partial^3_{\beta_-}U \\
& + \frac{11}{3}\partial_{\beta_-}U\partial^2_{\beta_-}U + \frac{2}{15}\partial^5_{\beta_-}U + \frac{1}{3}\partial^3_{\beta_-}\varphi^- \\
& - \partial^{-1}_{\beta_-}\partial^2_T U, \\
J^+ = {} & 3\partial_{\beta_+}(V\varphi^+) + 4V^2\partial_{\beta_+}V + \frac{7}{3}V\partial^3_{\beta_+}V \\
& + \frac{11}{3}\partial_{\beta_+}V\partial^2_{\beta_+}V + \frac{2}{15}\partial^5_{\beta_+}V + \frac{1}{3}\partial^3_{\beta_+}\varphi^+ \\
& - \partial^{-1}_{\beta_+}\partial^2_T V, \\
J^s = {} & \partial_\beta\left(U\left(3G + 3\varphi^+ + 4V^2 + \frac{7}{3}\partial^2_{\beta_+}V\right)\right) \\
& + \partial_\beta\left(V\left(3F + 3\varphi^- + 4U^2 + \frac{7}{3}\partial^2_{\beta_-}U\right)\right) \\
& + 4\partial_\beta\left(\partial_{\beta_-}U\partial_{\beta_+}V\right),
\end{aligned}
\tag{22}
$$

and

$$
\begin{aligned}
E^u_{time} = {} & \epsilon^9\frac{4}{3}\partial_\beta\left(\left(\partial_{\beta_-}U - \partial_{\beta_+}V\right)\left(\partial_T U + \partial_T V\right)\right) \\
& - \epsilon^{11}\frac{2}{3}\partial_\beta\left(\left(\partial_T U + \partial_T V\right)^2\right).
\end{aligned}
$$

Notice that $J^\pm = J^\pm(\beta_\pm, T)$. E^u_{time} (so called because each term in it contains some sort of time derivative) is clearly $O(\epsilon^{17/2})$. That is,

$$\|E^u_{time}\|_s \le C\epsilon^{17/2}.$$

The constant above depends on $\|U\|_{s+4}$ and $\|V\|_{s+4}$.

The term $\epsilon^9\partial^{-1}_\beta\left(\partial^2_T F + \partial^2_T G\right)$ is not included in E^u_{time} for the following reason. In a moment, when we select the equations F and G solve, a consequence will be that there will be terms in $\partial^2_T F$ and $\partial^2_T G$ which are $O(\epsilon^{-2})$.

By taking

$$
\begin{aligned}
-2\partial_T F &= \frac{1}{3}\partial_{\beta_-}(UF) + \frac{3}{2}\partial^3_{\beta_-}F + J^-, \\
2\partial_T G &= \frac{1}{3}\partial_{\beta_+}(VG) + \frac{3}{2}\partial^3_{\beta_+}G + J^+,
\end{aligned}
$$

we cancel nearly all the terms which are not in the various E^u terms. These are the linearized KdV equations (LK) discussed in the introduction. Proposition 1 guarantees that the solutions are well behaved. We are left with

$$
\begin{aligned}
\mathrm{Res}_y = {} & + \epsilon^7\left(\partial^{-1}_\beta\left(\partial^2_\tau W_3 - \partial^2_\beta W_3\right) + J^s\right) \\
& - \epsilon^9\partial^{-1}_\beta\left(\partial^2_T F + \partial^2_T G\right) \\
& + E^u_{geo} + E^u_{small} + E^u_{switch} + E^u_{lwa} + E^u_{time}.
\end{aligned}
$$

Now we consider the terms in $\partial_\beta^{-1}\left(\partial_T^2 F + \partial_T^2 G\right)$. Notice that

$$\begin{aligned}-\partial_{\beta_-}^{-1}\partial_T^2 F &= \partial_{\beta_-}^{-1}\partial_T\left(\frac{1}{6}\partial_\beta^3 F + \frac{3}{2}\partial_\beta(UF) + \frac{1}{2}J^-\right)\\ &= \frac{1}{6}\partial_{\beta_-}^2\partial_T F + \frac{3}{2}\partial_T(UF) + \frac{1}{2}\partial_{\beta_-}^{-1}\partial_T J^-.\end{aligned}$$

J^- contains the term $3\partial_{\beta_-}(U\varphi^-) + \frac{1}{3}\partial_{\beta_-}^3\varphi^-$. From the definition of φ^- we know that

$$\partial_T\varphi^- = -\epsilon^{-2}\frac{3}{2}\partial_\beta(UV).$$

Thus we have

$$\begin{aligned}\frac{1}{2}\partial_{\beta_-}^{-1}\partial_T\left(3\partial_{\beta_-}(U\varphi^-) + \frac{1}{3}\partial_{\beta_-}^3\varphi^-\right) &= \frac{3}{2}\varphi^-\partial_T U + \frac{3}{2}U\partial_T\varphi^- + \frac{1}{6}\partial_{\beta_-}^2\partial_T\varphi^-\\ &= \frac{3}{2}\varphi^-\partial_T U - \frac{9}{4}\epsilon^{-2}U\partial_\beta(UV) + \frac{1}{4}\epsilon^{-2}\partial_\beta^3(UV).\end{aligned}$$

We treat $\partial_T^2 G$ in the same fashion. Therefore we can write

$$-\epsilon^9\partial_\beta^{-1}\left(\partial_T^2 F + \partial_T^2 G\right) = E_{F,G}^u - \epsilon^7\left(\frac{9}{4}(U+V)\partial_\beta(UV) + \frac{1}{2}\partial_\beta^3(UV)\right).$$

By construction $E_{F,G}^u$ satisfies the estimate

$$\|E_{F,G}^u\|_s \le C\epsilon^{17/2},$$

where C depends on $\|U\|_{s+7}$, $\|V\|_{s+7}$, $\|F\|_{s+5}$, and $\|G\|_{s+5}$.

We have

$$\begin{aligned}\mathrm{Res}_y = \;&\epsilon^7\left(\partial_\beta^{-1}\left(\partial_\tau^2 W_3 - \partial_\beta^2 W_3\right) + J^s\right)\\ &- \epsilon^7\left(\frac{9}{4}(U+V)\partial_\beta(UV) + \frac{1}{2}\partial_\beta^3(UV)\right)\\ &+ E_{geo}^u + E_{small}^u + E_{switch}^u + E_{lwa}^u + E_{time}^u + E_{F,G}^u.\end{aligned}$$

By selecting

$$\partial_\tau^2 W_3 - \partial_\beta^2 W_3 = \partial_\beta\left(\frac{9}{4}(U+V)\partial_\beta(UV) + \frac{1}{2}\partial_\beta^3(UV) - J^s\right) \tag{23}$$

the gun goes off and we cancel all remaining $O(\epsilon^7)$ terms. Thus

$$\mathrm{Res}_y = E_{geo}^u + E_{small}^u + E_{switch}^u + E_{lwa}^u + E_{time}^u + E_{F,G}^u.$$

Each of the E^u is $O(\epsilon^{17/2})$.

Unlike the previous equations (KdV), (IW), and (LK), Proposition 1 does not tell us that the solutions to (23) are controllable. Nonetheless, (23) is an inhomogeneous wave equation where the inhomogeneity consists entirely of terms which are products of left- and right-moving rapidly decaying functions. From Wayne and Wright [33] we have the following lemma.

LEMMA 3. *Suppose*

$$\partial_\tau u \pm \partial_\beta u = l(\beta+\tau, \epsilon^2\tau) r(\beta-\tau, \epsilon^2\tau), \quad u(X,0)=0,$$

with $\|l(\cdot,T)\|_{H^s(2)} \le C$ *and* $\|r(\cdot,T)\|_{H^s(2)} \le C$ *for* $T\in[0,T_0]$; *then*

$$\|u(\beta,\tau)\|_s \le C$$

for all $\tau \in [0, T_0\epsilon^{-2}]$. *The constant* C *is uniform in* ϵ.

Thus W_3 will remain $O(1)$.

Remark 8. If we are in the situation in which Proposition 1 applies, we see that least regular parts in the driving term are $\partial_\beta^2(UG)$ and $\partial_\beta^2(VF)$, which are in $H^{\sigma-6}(2)$. Thus, by this lemma we have that $W_3 \in H^{\sigma-5}$ for all times of interest.

At this time we have derived the modulation equations and shown that the residual is small. The only remaining order of business in this section is to determine how smooth the solutions to our modulation equations need to be in order for $\mathrm{Res}[\Psi]$ to be appropriately regular. This may seem to be a fairly tiresome task, but fortunately the least regular terms in all of the sundry E functions come from only one term—E^u_{small}! This is because E^u_{small} contains many time derivatives.

We need to control Res_a in H^{s-1}. For this we need $\partial_t \mathrm{Res}_u \in H^{s-1}$, which in turn implies that we must have $\partial_t E^u_{small} \in H^{s-1}$. Recalling (20), we see that this will require $\partial_\tau^3 Z_2 \in H^{s-1}$, or rather (since L_ϵ^{-1} saves a derivative) $\partial_\tau^3 W_2 \in H^{s-2}$. For this, we need $\partial_T^3 F$ and $\partial_T^3 G$ in H^{s-2}. Given that F solves (LK) where J^+ contains the terms $\partial_{\beta\_}^5 U$, $\partial_{\beta\_}^3 \varphi^-$, one sees that $\partial_T^3 F$ will include the terms $\partial_{\beta\_}^9 F$, $\partial_{\beta\_}^{11} U$, and $\partial_{\beta\_}^9 \varphi^-$. Thus $\|\partial_T^3 F\|_{s-2}$ is controlled by the H^{s+9} norms of U and V, and the H^{s+7} norms of φ^-, F, and G. The analogous result is true for $\partial_T^3 G$. We also need $\partial_\tau^3 W_3 \in H^{s-2}$. Since W_3 solves (23), we require $W_3 \in H^{s+1}$.

In summary we have the following proposition.

PROPOSITION 2. *Take* Ψ^d *as in* (7), *with* U, V, F, G, P, *and* W_3 *solving their respective equations. Let*

$$\begin{aligned}
\Psi^z &= L^{-1}\Psi^d,\\
\Psi^y &= \Psi^z + \epsilon^4\Delta_1 + \epsilon^6\Delta_2,\\
\Psi^u &= \partial_\alpha^{-1}\partial_t\Psi^d,\\
\Psi^a &= \partial_\alpha^{-1}\partial_t^2\Psi^d,
\end{aligned}$$

with Δ_1 *and* Δ_2 *as in* (17) *and* (19), *and form* $Res[\Psi]$ *as in* (4). *Then*

$$\begin{aligned}
\|Res_z\|_s &\le C\epsilon^{17/2},\\
\|Res_y\|_s &\le C\epsilon^{17/2},\\
\|Res_u\|_{s-1} &\le C\epsilon^{17/2},\\
\|Res_a\|_{s-1} &\le C\epsilon^{19/2},
\end{aligned}$$

where C *is a constant which depends on* $\|U\|_{s+9}$, $\|V\|_{s+9}$, $\|P\|_{s+7}$, $\|F\|_{s+7}$, $\|G\|_{s+7}$, *and* $\|W_3\|_{s+1}$. *The estimate* (6) *holds as long as these quantities remain bounded. The constant* C *does not depend on* ϵ.

In light of Proposition 1 and Remark 8, we see, if we take the initial conditions for U and V to satisfy (2) with $\sigma \ge s+7$, that $\|U\|_{s+9}$, $\|V\|_{s+9}$, $\|P\|_{s+7}$, $\|F\|_{s+7}$, $\|G\|_{s+7}$, and $\|W_3\|_{s+1}$ are all $O(1)$ for $t\in[0,T_0\epsilon^{-3}]$. Thus we move on.

5. The operator ( ). Part II: Estimates and extensions. In this section we will describe a few more estimates related to $K(x,y)$. All such estimates are either smoothing estimates or ones which show that certain terms are small in the long wavelength setting.

First, since $1+\widehat{K_0}^2(k)$ goes to zero exponentially fast as $|k|\to\infty$, the operator $1+K_0^2$ is smoothing. That is, for all $s\geq 0$

$$\|(1+K_0^2)u\|_s \leq C\|u\|_{L^2}.$$

Also, commutators involving K_0 are smoothing. We quote the following lemma from [28].

LEMMA 4. *Let $r\geq 0$, $q>1/2$, and $0\leq p\leq q$. Then there exists a $C>0$ such that*

$$\|[f,K_0]g\|_r \leq C\|f\|_{r+p}\|g\|_{q-p}.$$

Proof. See Lemma 3.12 on page 1498 of [28]. □

Schneider and Wayne show that $K_1(z,y)$ is a smoothing operator.

LEMMA 5. *For $r\geq 0$, $q\geq 1/2$ and $0\leq p\leq q$, there is C such that*

$$\|K_1(z,y)u\|_r \leq C\left(\|z\|_{r+p}+\|y\|_{r+p}\right)\|u\|_{q-p}. \tag{24}$$

Proof. See Corollary 3.13 on page 1499 of [28]. □

If we let $S_1(z,y)=K(z,y)-K_0$, we also have the following estimates from [28].

LEMMA 6. *Fix $s\geq 4$. If the free surface is sufficiently smooth, then for $j=1,2$ we have*

$$\text{(a)}\qquad \|S_j(z,y)u\|_s \leq C\left(\|z\|_s^j+\|y\|_s^j\right)\|u\|_3,$$

that is, S_j is a smoothing operator;

$$\text{(b)}\qquad \|\partial_\alpha(S_j(z,y)u)\|_s \leq C\left(\|z\|_s^j+\|y\|_s^j\right)\|u\|_3,$$

that is, $\partial_\alpha S_j$ is a smoothing operator;

$$\text{(c)}\qquad \|[\partial_t,S_j]u\| \leq C\left(\|z\|_s^j+\|y\|_s^j\right)\|u\|_3,$$

that is, $[\partial_t,S_j]$ is a smoothing operator and this operator can be bounded independently of $\partial_t u$; and

$$\text{(d)}\qquad \|[\partial_t^2,S_j]u\| \leq C\left(\|z\|_s^j+\|y\|_s^j\right)\left(\|u\|_4+\|\partial_t u\|_4\right),$$

that is, $[\partial_t^2,S_j]$ is a smoothing operator and this operator can be bounded independently of $\partial_t^2 u$.

Proof. The proof is in [28]; see Lemmas 3.14 and 3.15 and Corollary 3.16 on pages 1500, 1506, and 1507, respectively. □

We will also need the following propositions concerning the behavior of the remainder terms S_1 and S_2. The first of these says that more or less the remainder S_2 is negligible for the sort of scalings we are considering. That is to say, the term $E^y_{S_2}$ in section 4 is very small.

PROPOSITION 3. *Fix $s>5/2$. Suppose $z=\epsilon^2Z(\epsilon\alpha)$, $y=\epsilon^2Y(\epsilon\alpha)$, and $f=\epsilon^2F(\epsilon\alpha)$, with $Z,Y,F\in H^{s+1}(2)$. Moreover, assume $z-y=\epsilon^4\Delta(\epsilon\alpha)$ with*

$\Delta \in H^{s+1}(2)$. *Then there exists* ϵ_0 *such that for* $\epsilon \in [0, \epsilon_0]$ *there is a constant* C *independent of* ϵ *such that*

$$\|S_2(z, y)f\|_s \leq C\epsilon^{17/2}.$$

The second proposition is a technical version of the mean value theorem as applied to the operator S_1.

PROPOSITION 4. *Suppose* $z(\alpha, t) = \epsilon^2 Z(\epsilon(\alpha \pm t), \epsilon^3 t)$, $y(\alpha, t) = \epsilon^2 Y(\epsilon(\alpha \pm t), \epsilon^3 t)$, $u(\alpha, t) = \epsilon^2 U(\epsilon(\alpha \pm t), \epsilon^3 t)$, *and* $f(\alpha, t) = \epsilon^2 F(\epsilon(\alpha \pm t), \epsilon^3 t)$ *with* $Z, Y, U, F \in H^s(2)$ *for* $t \in [0, T_0\epsilon^{-3}]$. *Also suppose* $R^z(\alpha, t)$, $R^y(\alpha, t)$, *and* $R^u(\alpha, t) \in H^s$ *for the same time interval. Then*

$$\|S_1(z(\cdot) + \epsilon^{11/2}R^z(\cdot), y(\cdot) + \epsilon^{11/2}R^y(\cdot))f(\cdot) - S_1(z(\cdot), y(\cdot))f(\cdot)\|_s \leq C\epsilon^{17/2}$$

for $t \in [0, T_0\epsilon^{-3}]$.

Proof of Proposition 3. First, notice that $x(\alpha) = \int_0^\alpha Lz(a)da = \epsilon X(\epsilon\alpha)$. We know that X is in L^∞ by the following lemma.

LEMMA 7. *Suppose* $f(\alpha) = \epsilon^2 F(\epsilon\alpha)$ *with* $F \in H^s(2)$. *Then for all* α

$$\left|\int_0^\alpha f(a)da\right| \leq C\epsilon\|F\|_{H^s(2)}.$$

Proof. See section 7. □

Let $\Phi(\tilde{x}, y) = (\Phi_1, \Phi_2)$ be the analytic map which takes $\Omega(t)$ to

$$P^- = \{(\xi, \gamma) | \gamma \in [-1, 0]\}.$$

That such a map exists and is analytic is guaranteed by the Riemann mapping theorem. Let

$$h(\alpha) = \Phi_1(\tilde{x}(\alpha), y(\alpha))$$

and $Qf = f \circ h$. From [28], we know that

$$K(x, y)f(\alpha) = Q \circ K_0 \circ Q^{-1} f(\alpha). \tag{25}$$

We can derive a very useful implicit formula for h^{-1} as follows. The function $\Phi^{-1}(\xi, \gamma)$ is analytic on P^-; thus it satisfies the Cauchy–Riemann equations. If we set

$$\Phi^{-1}(\xi, \gamma) = (\xi + u_1(\xi, \gamma)) + i(\gamma - v_1(\xi, \gamma))$$

and notice that Φ^{-1} sends the bottom and top of P^- to the bottom and top of $\Omega(t)$, respectively, we see that we have the following system:

$$\begin{gathered}\partial_\xi u_1 + \partial_\gamma v_1 = 0,\\ \partial_\gamma u_1 - \partial_\xi v_1 = 0,\\ v_1(\xi, -1) = 0, \quad v_1(\xi, 0) = \eta(\xi),\end{gathered}$$

where $\eta(\xi) = y(h^{-1}(\xi))$. One can solve this system with the use of Fourier transforms relatively simply. If one does so, one finds that

$$u_1(\xi, 0) = -\int_0^\xi \eta(\xi_1)d\xi_1 - M\eta(\xi),$$

where M is the pseudodifferential operator given by

$$\widehat{M\eta}(k) = \frac{k\cosh(k) - \sinh(k)}{ik\sinh(k)}\widehat{\eta}(k).$$

Notice that to lowest order, M is $C\partial_\alpha$.

Now, notice that $h^{-1}(\xi) = \tilde{x}^{-1}(u(\xi,0))$ and so we have an implicit equation for h^{-1}:

$$\begin{aligned} h^{-1}(\xi) &= \tilde{x}^{-1}\left(\xi - \int_0^\xi \eta(\xi_1)d\xi_1 - M\eta(\xi)\right) \\ &= \tilde{x}^{-1}\left(\xi - \int_0^\xi y(h^{-1}(\xi_1))d\xi_1 - M(y\circ h^{-1})(\xi)\right). \end{aligned} \tag{26}$$

Remark 9. In [28, p. 1494], Schneider and Wayne make a minor error in calculating this same function. As a result, they claim that the above representation gives an explicit formula for h^{-1}. Our correction here changes nothing about subsequent steps in their proofs.

Since $\tilde{x} = \alpha + \epsilon X(\epsilon\alpha)$, where X is well behaved, we can expect a similar form for $\tilde{x}^{-1}$.

LEMMA 8. *Suppose $f(\alpha) = \alpha + g(\alpha)$ with $\|g\|_{C^2} \le 1/2$. Then*

$$f^{-1}(\xi) = \xi - g(\xi) + g(\xi)g'(\xi) + E,$$

where $E = O(\|g\|^3_{W^{2,\infty}})$. More specifically

$$E \le C\left(\|g'\|^2_{L^\infty}\|g\|_{L^\infty} + \|g\|^2_{L^\infty}\|g''\|_{L^\infty}\right).$$

In particular, notice that if $g(\alpha) = \epsilon G(\epsilon\alpha)$ this means that $E = O(\epsilon^5)$.

Proof. See section 7. ∎

We apply this lemma to $\tilde{x}$ and find that

$$\tilde{x}^{-1}(\xi) = \xi - \epsilon X(\epsilon\xi) + \epsilon^3 X(\epsilon\xi)\partial_\beta X(\epsilon\xi) + O(\epsilon^5).$$

Combining this with (26) we can determine $h^{-1}(\xi)$ (and therefore h) in terms of x and y to any order we wish. To lowest order we see that

$$h^{-1}(\xi) = \xi + O(\epsilon). \tag{27}$$

Therefore now we have

$$\begin{aligned} h^{-1}(\xi) = \xi &- \int_0^\xi y(h^{-1}(\xi_1))d\xi_1 - M(y\circ h^{-1})(\xi) \\ &- \epsilon X\left(\epsilon\left(\xi - \int_0^\xi y(h^{-1}(\xi_1))d\xi_1 - M(y\circ h^{-1})(\xi)\right)\right) \\ &+ O(\epsilon^3) \\ = \xi &- \int_0^\xi \epsilon^2 Y(\epsilon h^{-1}(\xi_1))d\xi_1 - M(\epsilon^2 Y\circ \epsilon h^{-1})(\xi) \\ &- \epsilon X\left(\epsilon\left(\xi - \int_0^\xi \epsilon^2 Y(\epsilon h^{-1}(\xi_1))d\xi_1 - M(\epsilon^2 Y\circ\epsilon h^{-1})(\xi)\right)\right) \\ &+ O(\epsilon^3). \end{aligned}$$

If we insert (27) into the above and expand we have

$$h^{-1}(\xi) = \xi - \epsilon X(\epsilon\xi) - \int_0^\xi \epsilon^2 Y(\epsilon\xi_1))d\xi_1 - \epsilon^2 M(Y(\epsilon\cdot))(\xi) + O(\epsilon^3).$$

One can continue in this manner and determine the next order terms in the expansion of h^{-1}. If we let

$$\epsilon G_1(\epsilon\alpha) = -\epsilon X(\epsilon\alpha) - \int_0^\alpha \epsilon^2 Y(\epsilon\alpha_1))d\alpha_1 - \epsilon^2 M(Y(\epsilon\cdot))(\alpha),$$

the expansion is

$$h^{-1}(\xi) = \xi + \epsilon G_1(\epsilon\xi) + \epsilon^3 B_1(\epsilon\xi) + O(\epsilon^5),$$

where

$$\begin{aligned} B_1(\xi) = &\int_0^\xi \epsilon G_1(\epsilon a)\partial_\beta \epsilon^3 Y(\epsilon a)da \\ &+ M(\epsilon^3 G_1(\epsilon\cdot)\partial_\beta Y(\epsilon\cdot))(\xi) + \epsilon^3 G_1(\epsilon\xi)\partial_\beta X(\epsilon\xi). \end{aligned}$$

Notice that since M is $C\partial_\alpha$ to lowest order, $-\epsilon^2 M(Y(\epsilon\cdot))(\alpha)$ is $O(\epsilon^3)$. Moreover, by hypothesis, we have $\epsilon^2 Z(\epsilon\alpha) - \epsilon^2 Y(\epsilon\alpha) = \epsilon^4\Delta(\epsilon\alpha)$. Thus

$$\begin{aligned} -\epsilon X(\epsilon\alpha) - \int_0^\alpha \epsilon^2 Y(\epsilon a)da &= -\int_0^\alpha \left(\epsilon^2\partial_\beta X(\epsilon a) + \epsilon^2 Y(\epsilon a)\right) da \\ &= -\int_0^\alpha \left(\epsilon^2 L(Z(\epsilon\cdot))(a) + \epsilon^2 Y(\epsilon a)\right) da \\ &= \int_0^\alpha \left(\epsilon^2 Z(\epsilon a) - \epsilon^2 Y(\epsilon a)\right) da + O(\epsilon^3) \\ &= \int_0^\alpha \left(\epsilon^4 \Delta(\epsilon a)\right) da + O(\epsilon^3) \\ &= O(\epsilon^3). \end{aligned}$$

That is, ϵG_1 is really $O(\epsilon^3)$! This cancellation is the crucial step in this proof. Since ϵG_1 appears in each term in B_1, we have shown

$$h^{-1}(\xi) = \xi + \epsilon^3 G(\epsilon\xi) + O(\epsilon^5)$$

with $\epsilon^3 G = \epsilon G_1$. We appeal to Lemma 8 again, and we have

$$h(\alpha) = \alpha - \epsilon^3 G(\epsilon\alpha) + O(\epsilon^5).$$

Now that we have particularly good estimates on h and h^{-1}, we can begin our discussion of K in earnest. For notational simplicity, we will let

$$\begin{aligned} h(\alpha) &= \alpha + g_1(\alpha), \\ h^{-1}(\xi) &= \xi + g_2(\xi). \end{aligned}$$

If we let

$$\tilde{f} = Q^{-1} f,$$

we can make the following formal approximation using Taylor's theorem:

$$\begin{aligned}K(x,y)f(\alpha) &= Q \circ K_0 \tilde{f}(\alpha)\\ &= K_0 \tilde{f}(h(\alpha))\\ &= K_0 \tilde{f}(\alpha + g_1(\alpha))\\ &= K_0 \tilde{f}(\alpha) + g_1(\alpha) K_0 \partial_\alpha \tilde{f}(\alpha) + \text{h.o.t.}\end{aligned}$$

Also by Taylor's theorem,

$$\tilde{f}(\alpha) = f(\alpha) + g_2(\alpha)\partial_\alpha f(\alpha) + \text{h.o.t.}$$

Putting these together we have

$$K(x,y)f(\alpha) = K_0 f(\alpha) + g_1(\alpha) K_0 f'(\alpha) + K_0(g_2 f')(\alpha) + \text{h.o.t.}$$

Thus let

$$\begin{aligned}E_1 f &= K_0 f(\alpha) + g_1(\alpha) K_0 f'(\alpha) + K_0(g_2 f')(\alpha),\\ E_2 f &= K(x,y) f - E_1 f.\end{aligned}$$

We prove Proposition 3 if we can prove

- $\|E_1 f - K_0 f - K_1(x,y) f\|_s \leq C\epsilon^{17/2}$ and
- $\|E_2 f\|_s \leq C\epsilon^{17/2}$.

Let us deal with $E_2 f$ first. We can rewrite $E_2 f$ as

$$E_2 f = E_2^1 f + E_2^2 f + E_2^3 f$$

with

$$\begin{aligned}E_2^1 f &= K(x,y) f - K_0 \tilde{f} - g_1 K_0 \partial_\alpha \tilde{f},\\ E_2^2 f &= K_0 \tilde{f} - K_0 f - K_0(g_2 \partial_\alpha f),\\ E_2^3 f &= g_1 K_0 \partial_\alpha \tilde{f} - g_1 K_0 \partial_\alpha f.\end{aligned}$$

As our approximation for K was determined by an application of Taylor's theorem, we need to prove a lemma which shows that this formal step can be made rigorous, at least for functions in the weighted Sobolev spaces.

Lemma 9. *Suppose $F \in H^s(n)$, $s > 1/2$, $n > 1/2$. Then for all $C_0 > 0$ there exists ϵ_0 such that for $\epsilon \in [0, \epsilon_0]$ there is a constant C independent of ϵ such that*

$$\left(\int_{|\alpha| > C_0 \epsilon^{-3}} |F(\epsilon\alpha)|^2 d\alpha\right)^{1/2} \leq C\epsilon^{2n-3/2}.$$

Moreover, for $1 \leq j \leq s$ we have

$$\left(\int_{|\alpha| > C_0 \epsilon^{-3}} |\partial_\alpha^j F(\epsilon\alpha)|^2 d\alpha\right)^{1/2} \leq C\epsilon^{2n-3/2+j}.$$

Proof. See section 7. ∎

Remark 10. If instead we are considering

$$\left(\int_{|\alpha|>C_0\epsilon^{-3}} |F(\epsilon(\alpha \pm t), \epsilon^3 t)|^2 d\alpha\right)^{1/2}$$

with $F(\cdot, T) \in H^s(n)$ for $T \in [0, T_0]$, we can maintain the same bound as above by taking $C_0 \geq 2T_0$.

We can use the above lemma to prove a version of Taylor's theorem.

LEMMA 10. *Suppose $F \in H^s(2)$, $s > 5/2$, and $g \in L^\infty$. Then*

$$\|F(\epsilon \cdot + \epsilon^2 g(\epsilon\cdot)) - F(\epsilon\cdot)\|_{L^2} \leq C\epsilon^{3/2}.$$

Proof. By Lemma 9 we have

$$\begin{aligned}
&\|F(\epsilon \cdot + \epsilon^2 g(\epsilon\cdot)) - F(\epsilon\cdot)\|^2 \\
&= \int_{|\alpha|\leq\epsilon^{-3}} |F(\epsilon\alpha + \epsilon^2 g(\epsilon\alpha)) - F(\epsilon\alpha)|^2 d\alpha \\
&+ \int_{|\alpha|\geq\epsilon^{-3}} |F(\epsilon\alpha + \epsilon^2 g(\epsilon\alpha)) - F(\epsilon\alpha)|^2 d\alpha \\
&\leq \int_{|\alpha|\leq\epsilon^{-3}} |F(\epsilon\alpha + \epsilon^2 g(\epsilon\alpha)) - F(\epsilon\alpha)|^2 d\alpha + C\epsilon^5.
\end{aligned}$$

Now, we add and subtract $\epsilon^2 g(\epsilon\alpha) F'(\epsilon\alpha)$ in the remaining integral,

$$\begin{aligned}
&\int_{|\alpha|\leq\epsilon^{-3}} |F(\epsilon\alpha + \epsilon^2 g(\epsilon\alpha)) - F(\epsilon\alpha)|^2 d\alpha \\
&\leq \int_{|\alpha|\leq\epsilon^{-3}} |F(\epsilon\alpha + \epsilon^2 g(\epsilon\alpha)) - F(\epsilon\alpha) - \epsilon^2 g(\epsilon\alpha)F'(\epsilon\alpha)|^2 d\alpha \\
&\quad + \int_{|\alpha|\leq\epsilon^{-3}} |\epsilon^2 g(\epsilon\alpha)F'(\epsilon\alpha)|^2 d\alpha \\
&\leq \int_{|\alpha|\leq\epsilon^{-3}} |F(\epsilon\alpha + \epsilon^2 g(\epsilon\alpha)) - F(\epsilon\alpha) - \epsilon^2 g(\epsilon\alpha)F'(\epsilon\alpha)|^2 d\alpha \\
&\quad + \epsilon^3 \|g(\cdot)\|_{L^\infty}^2 \|F'(\cdot)\|_{L^2}^2.
\end{aligned}$$

We naively bound the above integral and apply the mean value theorem. That is,

$$\begin{aligned}
&\int_{|\alpha|\leq\epsilon^{-3}} |F(\epsilon\alpha + \epsilon^2 g(\epsilon\alpha)) - F(\epsilon\alpha) - \epsilon^2 g(\epsilon\alpha)F'(\epsilon\alpha)|^2 d\alpha \\
&\leq C\epsilon^{-3} \sup_{|\alpha|\leq\epsilon^{-3}} |F(\epsilon\alpha + \epsilon^2 g(\epsilon\alpha)) - F(\epsilon\alpha) - \epsilon^2 g(\epsilon\alpha)F'(\epsilon\alpha)|^2 \\
&\leq C\epsilon^{-3} \sup_{|\alpha|\leq\epsilon^{-3}} |\epsilon^4 g^2(\epsilon\alpha) F''(\epsilon\alpha^*)|^2 \\
&\leq C\epsilon^5 \|g\|_{L^\infty}^4 \|F''\|_{L^\infty}^2.
\end{aligned}$$

With this, we have proven the lemma. $\square$

Remark 11. With this general technique we are also able to show that

$$\|\partial_\alpha^j (F(\epsilon \cdot + \epsilon^2 g(\epsilon\cdot)) - F(\epsilon\cdot))\|_{L^2} \leq C\epsilon^{3/2+j},$$

$$\|\partial_\alpha^j \left(F(\epsilon\cdot+\epsilon^2 g(\epsilon\cdot)) - F(\epsilon\cdot) - \epsilon^2 g(\epsilon\cdot)F'(\cdot)\right)\|_{L^2} \le C\epsilon^{7/2+j},$$

$$\|\partial_\alpha^j \left(F(\epsilon\cdot+\epsilon^2 g(\epsilon\cdot))\right) - \partial_\alpha^j \left(F(\epsilon\cdot) + \epsilon^2 g(\epsilon\cdot)F'(\cdot) + 1/2\epsilon^4 g^2(\epsilon\cdot)F''(\epsilon\cdot)\right)\|_{L^2} \le C\epsilon^{11/2+j},$$

and so on.

Now we will be able to control E_2^1. We can control the other functions in precisely the same fashion. Since f is of long wavelength and rapid decay, so is $\tilde{f}$. Thus we can use Lemma 10. In what follows, $\epsilon^3\tilde{F}(\epsilon\alpha) = K_0\tilde{f}(\alpha)$. (The extra ϵ comes from the long wave approximation of K_0.)

$$\begin{aligned}
&\|E_2^1 f(\cdot)\|_s = \|K_0\tilde{f}(\cdot+g_1(\cdot)) - K_0\tilde{f}(\cdot) - g_1(\cdot)K_0\partial_\alpha\tilde{f}(\cdot)\|_s \\
&\le C\epsilon^3\|\tilde{F}(\epsilon\cdot+\epsilon^4 G(\epsilon\cdot)) - \tilde{F}(\epsilon\cdot) - \epsilon^4 G(\epsilon\cdot)\partial_\beta\tilde{F}(\epsilon\cdot)\|_s \\
&\le C\epsilon^{17/2}.
\end{aligned}$$

Now we turn our attention to $E_1 f - K_0 f - K_1(x,y)f$. A routine calculation shows that this is equal to

$$\begin{aligned}
&-\epsilon^3[G(\epsilon\alpha),K_0]\epsilon^3\partial_\beta F(\epsilon\alpha) - K_1(\epsilon X,\epsilon^2 Y)\epsilon^2 F(\epsilon\alpha) \\
&+(g_1+\epsilon^3 G)K_0\epsilon^3\partial_\beta F(\epsilon\alpha) + K_0\left((g_2-\epsilon^3 G)\epsilon^3\partial_\beta F(\epsilon\cdot)\right)(\alpha).
\end{aligned}$$

Now, $g_1+\epsilon^3 G$ and $g_2-\epsilon^3 G$ are $O(\epsilon^5)$, so the second line above can be bounded by $C\epsilon^{17/2}$, if we make use of K_0's long wavelength approximation. Moreover, we claim that the first line is identically zero. Let

$$\begin{aligned}
b(\alpha) &= -\int_0^\alpha \epsilon^2 Y(\epsilon\alpha_1)d\alpha_1 - \epsilon^2 M(Y(\epsilon\cdot))(\alpha) \\
&= -\int_0^\alpha y(\alpha_1)d\alpha_1 - My(\alpha).
\end{aligned}$$

Thus $\epsilon^3 G(\epsilon\alpha) = b(\alpha) - x(\alpha)$. So

$$\epsilon^3[G(\epsilon\alpha),K_0]\epsilon^3\partial_\beta F(\epsilon\alpha) = -[x(\alpha),K_0]\partial_\alpha f(\alpha) + [b(\alpha),K_0]\partial_\alpha f(\alpha).$$

Taking the Fourier transform of the second term we have

$$\begin{aligned}
&\mathfrak{F}\left([b(\alpha),K_0]\partial_\alpha f(\alpha)\right)(k) \\
&= \int\left(\widehat{K_0}(l)-\widehat{K_0}(k)\right)\widehat{b}(l-k)il\widehat{f}(l)dl \\
&= \int\frac{\widehat{K_0}(l)-\widehat{K_0}(k)}{i(l-k)}i(l-k)\widehat{b}(l-k)il\widehat{f}(l)dl \\
&= \int\frac{\widehat{K_0}(l)-\widehat{K_0}(k)}{i(l-k)}\widehat{\partial_\alpha b}(l-k)il\widehat{f}(l)dl.
\end{aligned}$$

Now notice that $\partial_\alpha b = Ly$, so the above becomes

$$\begin{aligned}
&\int\frac{\widehat{K_0}(l)-\widehat{K_0}(k)}{i(l-k)}\widehat{Ly}(l-k)il\widehat{f}(l)dl = \int\frac{\widehat{K_0}(l)-\widehat{K_0}(k)}{\widehat{K_0}(l-k)}\widehat{y}(l-k)il\widehat{f}(l)dl \\
&= \int(1+\widehat{K_0}(k)\widehat{K_0}(l))\widehat{y}(l-k)il\widehat{f}(l)dl \\
&= \mathfrak{F}\left((y+K_0 y K_0)\partial_\alpha f\right)(k),
\end{aligned}$$

where we have used the trigonometric identity (11) from section 4.

That is,

$$\begin{aligned}&\epsilon^3[G(\epsilon\alpha), K_0]\epsilon^3\partial_\beta F(\epsilon\alpha) = -[x(\alpha), K_0]\partial_\alpha f(\alpha) + ((y + K_0 y K_0)\partial_\alpha f)(k)\\ =& - K_1(x, y)f\end{aligned}$$

and so we are done with the proof of Proposition 3. ∎

Proof of Proposition 4. Let h be as in Proposition 3 and h_2 be the analogous function for the configuration $(z_2, y_2) = (z + \epsilon^{11/2}R^z, y + \epsilon^{11/2}R^y)$. We also define the function $\tilde{x}_2 = \alpha + x_2$ by $\partial_\alpha x_2 = Lz_2$. Unlike in the previous lemma, here the time dependence of the functions is important. Thus we determine x and x_2 by integrating in both space and time. That is,

$$\begin{aligned}\tilde{x}(\alpha, t) &= \alpha + \epsilon\chi(\epsilon\alpha, \epsilon t),\\ \tilde{x}_2(\alpha, t) &= \alpha + \epsilon\chi(\epsilon\alpha, \epsilon t) + \epsilon^{5/2}E(t) + \epsilon^{11/2}\rho(\alpha, t)\end{aligned}$$

with

$$\begin{aligned}\epsilon\chi(\epsilon\alpha, \epsilon t) &= \left(\int_0^t u(0, s)\, ds + \int_0^\alpha Lz(w, t)\, dw\right),\\ \epsilon^{5/2}E(t) &= \epsilon^{11/2}\int_0^t R^u(0, s)\, ds,\\ \epsilon^{11/2}\rho &= \int_0^\alpha LR^z(w, t)\, dw.\end{aligned}$$

The functions satisfy the following estimates for all $t \in [0, T_0\epsilon^{-3}]$:

$$\begin{aligned}|\epsilon\chi(\epsilon\alpha, \epsilon t)| &\le C\epsilon,\\ |\epsilon^{5/2}E(t)| &\le C\epsilon^{5/2},\\ |\epsilon^{11/2}\rho(\alpha, t)| &\le C\sqrt{|\alpha|}\|R^z\|_{H^s}.\end{aligned}$$

The first estimate follows from similar estimates in the previous lemma, the second is the naive bound, and the final follows from the following simple fact.

FACT 2. *If $p(\alpha) = \int_0^\alpha r(a)da$, where $r \in L^2$, then*

$$|p(\alpha)| \le \sqrt{|\alpha|}\|r\|_{L^2}.$$

In what follows we will make strong use of the fact that $E(t)$ does not depend on α.

Using the same techniques as were used in proving Lemma 8, one can show that

$$\tilde{x}_2^{-1}(\xi, t) = \tilde{x}^{-1}(\xi, t) - \epsilon^{5/2}E(t) + \epsilon^{11/2}\rho_2(\xi),$$

where $|\rho_2(\alpha)| \le \sqrt{|\alpha|}\|R^z\|_{L^2}$. This sort of estimate carries over to the functions h. That is,

$$\begin{aligned}h_2^{-1}(\xi, t) &= h^{-1}(\xi, t) - \epsilon^{5/2}E(t) + \epsilon^{11/2}\rho_3(\xi, t),\\ h_2(\alpha, t) &= h(\alpha, t) + \epsilon^{5/2}E(t) + \epsilon^{11/2}\rho_4(\alpha, t),\end{aligned}$$

where $|\rho_3(\alpha, t)|, |\rho_4(\alpha, t)| \le \sqrt{|\alpha|}(\|R^z\|_{L^2} + \|R^y\|_{L^2})$ over the long time scale.

Now define $Q_2 f = f \circ h_2$. Therefore

$$\begin{aligned} &S_1(z+\epsilon^{11/2}R^z, y+\epsilon^{11/2})f - S_1(z,y)f \\ &= Q_2 \circ K_0 \circ Q_2^{-1} f - Q \circ K_0 \circ Q^{-1} f \\ &= Q \circ \big(Q^{-1} \circ Q_2 \circ K_0 \circ Q_2^{-1} \circ Q - K_0\big) \circ Q^{-1} f. \end{aligned}$$

Since Q and its inverse are bounded operators from H^s to H^s, we need only prove the estimate for the operator

$$\tilde{Q} \circ K_0 \circ \tilde{Q}^{-1} - K_0,$$

where $\tilde{Q} = Q^{-1} \circ Q_2$. Notice that $\tilde{Q} \circ K_0 \circ \tilde{Q}^{-1}$ is the Hilbert operator K for a domain with the "h" function given by $\tilde{h}(\alpha) = h_2(h^{-1}(\alpha))$. Moreover, from the above calculations for h and h_2 we have

$$h_2(h^{-1}(\alpha)) = \alpha + \epsilon^{5/2}E(t) + \epsilon^{11/2}\rho_5(\alpha,t),$$

with ρ_5 satisfying the same type of estimates as ρ_4.

At this point we can make an appeal to Lemma 3.14 on page 1500 of [28]. In this lemma they prove that $\|S_1(z,y)f\|_s \le C\left(\|z\|_s + \|y\|_s\right)\|f\|_3$. In the course of their proof, they show that if $h(\alpha) = \alpha + g(\alpha)$, then

$$\|Q \circ K_0 \circ Q^{-1} f(\cdot) - K_0 f(\cdot)\|_s \le C\|\partial_\alpha g\|_{s-1}\|\partial_\alpha f\|_2.$$

(See the inequalities in Cases I–IV on pages 1501–1506.) Therefore, if we set $\tilde{g} = \epsilon^{5/2}E(t) + \epsilon^{11/2}\rho_5(\alpha,t)$, we see that taking a spatial derivative leaves us with

$$\partial_\alpha \tilde{g} = O(\epsilon^{11/2}).$$

Thus, if we keep in mind that $f(\alpha,t) = \epsilon^2 F(\epsilon(\alpha \pm t), \epsilon^3 t)$,

$$\begin{aligned} \|\tilde{Q} \circ K_0 \circ \tilde{Q}^{-1} f(\cdot) - K_0 f(\cdot)\|_s &\le C\|\partial_\alpha \tilde{g}\|_{s-1}\|\partial_\alpha f\|_2 \\ &\le C\epsilon^{17/2}. \end{aligned}$$

This completes the proof of Proposition 4. □

6. The error estimates. In this section we prove that the approximation is rigorous. That is we will prove Theorem 2. We will be working with the three- and four-dimensional formulations of the water wave problem (equations (WW3) and (WW4)). From [28], we know that for initial data of the type we are considering, solutions to these equations exist over the long times we are considering. If (z,y,u) is a solution to (WW3), let

$$\begin{aligned} z(\alpha,t) &= \Psi^z(\alpha,t) + \epsilon^{11/2}R^z(\alpha,t), \\ y(\alpha,t) &= \Psi^y(\alpha,t) + \epsilon^{11/2}R^y(\alpha,t), \\ u(\alpha,t) &= \Psi^u(\alpha,t) + \epsilon^{11/2}R^u(\alpha,t), \end{aligned} \tag{28}$$

with the functions Ψ defined as above. We call R^z, R^y, and R^u "error" functions and we denote $\bar{R} = (R^z, R^y, R^u)$. Our goal will be to show that $\bar{R}$ remain $O(1)$ in $\mathfrak{H}^s = H^s \times H^s \times H^{s-1/2}$ over the long time scale. If we can do this, then we will have proven the main theorem. The first step will be to determine the equations which these functions satisfy. Loosely, we want to be able to write for each of the error functions an evolution equation of the form

$$\partial_t R = \text{quasi-linear} + \text{small and smooth}.$$

We will at times go to great lengths to achieve this!

Clearly,

$$\partial_t R^z = K_0 R^u.$$

Finding the equations for R^y and R^u is a bit more complex. First we focus on R^y. Substituting from (28) into $\partial_t y = K(z,y)u$, we have

$$\begin{aligned}
\partial_t R^y &= \epsilon^{-11/2}\left(K(\Psi^z + \epsilon^{11/2}R^z, \Psi^y + \epsilon^{11/2}R^y)(\Psi^u + \epsilon^{11/2}R^u) - \partial_t \Psi^y\right)\\
&= K(\Psi^z + \epsilon^{11/2}R^z, \Psi^y + \epsilon^{11/2}R^y)R^u\\
&\quad + \epsilon^{-11/2}\left(K(\Psi^z + \epsilon^{11/2}R^z, \Psi^y + \epsilon^{11/2}R^y)\Psi^u - \partial_t \Psi^y\right)\\
&= K_0 R^u + M_1(\Psi^z)\partial_\alpha R^u - (\Psi^y + K_0(\Psi^y K_0))\partial_\alpha R^u + N^y,
\end{aligned}$$

where

$$\begin{aligned}
N^y = \epsilon^{-11/2}\mathrm{Res}_y &+ \epsilon^{-11/2}\left(\left(S_1(\Psi^z + \epsilon^{11/2}R^z, \Psi^y + \epsilon^{11/2}R^y) - S_1(\Psi^z, \Psi^y)\right)\Psi^u\right)\\
&+ \left(K(\Psi^z + \epsilon^{11/2}R^z, \Psi^y + \epsilon^{11/2}R^y) - K_0 - K_1(\Psi^z, \Psi^y)\right)R^u.
\end{aligned}$$

We claim that N^y is "small." That is, we have the following lemma.

LEMMA 11. *For all $C_R > 0$, there exists ϵ_0 such that for all $\epsilon \in (0, \epsilon_0)$ and t such that $\sup_{0\le t' \le t} \|\bar{R}(\cdot, t')\|_{\mathfrak{H}^s} \le C_R$ we have*

$$\|N^y\|_s \le C\left(\epsilon^3 + \epsilon^3\|\bar{R}\|_{\mathfrak{H}^s} + \epsilon^{11/2}\|\bar{R}\|^2_{\mathfrak{H}^s}\right).$$

Proof. First we remark that the approximating functions Ψ and their derivatives are all bounded over the long time scales. Thus, we will not be keeping track of the dependence of the norm of N^y on the norms of these functions. By Proposition 2, we know that $\|\epsilon^{-11/2}\mathrm{Res}_y\|_s \le C\epsilon^3$.

We can bound

$$\epsilon^{-11/2}\left(S_1(\Psi^z + \epsilon^{11/2}R^z, \Psi^y + \epsilon^{11/2}R^y) - S_1(\Psi^z, \Psi^y)\right)\Psi^u$$

by Lemma 4.

Finally,

$$\begin{aligned}
&\|(K(\Psi^z + \epsilon^{11/2}R^z, \Psi^y + \epsilon^{11/2}R^y) - K_0 - K_1(\Psi^z, \Psi^y))R^u\|_s\\
&\le C\left(\epsilon^3\|\bar{R}\|_{\mathfrak{H}^s} + \epsilon^{11/2}\|\bar{R}\|^2_{\mathfrak{H}^s}\right)
\end{aligned}$$

by the estimates on K and its expansions which we saw in sections 3 and 5 (in particular Lemmas 5 and 6). ◻

Now we discuss R^u. We know that

$$\partial_t u(1 + Lz) + \partial_\alpha y(1 + \partial_t^2 y) = 0. \tag{29}$$

We would like an evolution-type equation for R^u. Notice that since $\partial_t y = K(z,y)u$, there is a "hidden" $\partial_t u$ in the term $\partial_t^2 y$. Recall that the commutator $[\partial_t, S_1(z,y)]u$ can be bounded independently of $\partial_t u$ (see Lemma 6 in section 5). Therefore, we can rewrite the above as

$$(1 + Lz + \partial_\alpha y K(z,y))\partial_t u + \partial_\alpha y(1 + [\partial_t, S_1(z,y)]u) = 0.$$

Replacing u with its definition in (28) the above becomes

$$\begin{aligned} 0 = &(1 + Lz + \partial_\alpha y K(z,y))\, \partial_t \epsilon^{11/2} R^u \\ &+ (1 + Lz + \partial_\alpha y K(z,y))\, \partial_t \epsilon^2 \Psi^u \\ &+ \partial_\alpha y\, (1 + [\partial_t, S_1(z,y)]\Psi^u) \\ &+ \partial_\alpha y \left([\partial_t, S_1(z,y)]\epsilon^{11/2} R^u \right) \end{aligned}$$

or rather

$$\begin{aligned} 0 = &(1 + Lz + \partial_\alpha y K(z,y))\, \partial_t \epsilon^{11/2} R^u \\ &+ (1 + Lz)\, \partial_t \epsilon^2 \Psi^u \\ &+ \partial_\alpha y\, (1 + \partial_t (K(z,y)\Psi^u)) \\ &+ \partial_\alpha y \left([\partial_t, S_1(z,y)]\epsilon^{11/2} R^u \right). \end{aligned}$$

We rearrange this a bit and break up y and z:

$$\begin{aligned} 0 = &(1 + Lz + \partial_\alpha y K(z,y))\, \partial_t R^u + \partial_\alpha R^y \\ &+ LR^z \partial_t \Psi^u + \partial_\alpha R^y \partial_t (K(z,y)\Psi^u) + \partial_\alpha y [\partial_t, S_1(z,y)] R^u \\ &+ \epsilon^{-11/2} \left((1 + L\Psi^z)\, \partial_t \epsilon^2 \Psi^u + \partial_\alpha \Psi^y\, (1 + \partial_t (K(z,y)\Psi^u)) \right). \end{aligned}$$

The operator

$$A(z,y) = (1 + Lz + \partial_\alpha y K(z,y))$$

is invertible since $K(z,y)$ is a bounded operator on H^s, provided z and y are small (which they are). Moreover, we can approximate A^{-1} via the Neumann series. Thus the above equation can be rewritten as

$$\partial_t R^u = -(1 - \epsilon^2 W_1)\partial_\alpha R^y + N^u,$$

where $N^u = N_1^u + N_2^u$ and

$$\begin{aligned} N_1^u = &- \epsilon^{-11/2} A^{-1} \left((1 + L\Psi^z)\, \partial_t \epsilon^2 \Psi^u + \partial_\alpha \Psi^y\, (1 + \partial_t (K(z,y)\Psi^u)) \right), \\ N_2^u = &- A^{-1} \left(LR^z \partial_t \Psi^u + \partial_\alpha R^y \partial_t (K(z,y)\Psi^u) + \partial_\alpha y [\partial_t, S_1(z,y)] R^u \right) \\ &+ \left(-A^{-1} + (1 - \epsilon^2 W_1) \right) \partial_\alpha R^y. \end{aligned}$$

LEMMA 12. *For all $C_R > 0$, there exists ϵ_0 such that for all $\epsilon \in (0, \epsilon_0)$ and t such that $\sup_{0 \le t' \le t} \|\bar{R}(\cdot, t')\|_{\mathfrak{H}^s} \le C_R$ we have*

$$\|N^u\|_{s-1} \le C \left(\epsilon^3 + \epsilon^3 \|\bar{R}\|_{\mathfrak{H}^s} + \epsilon^{11/2} \|\bar{R}\|_{\mathfrak{H}^s}^2 \right).$$

Proof. First we point out that this estimate is in H^{s-1}. The loss of regularity here is easily seen. Both LR^z and $\partial_\alpha R^y$ explicitly appear in N_2^u, and are not smoothed by any operators. Thus, losing this derivative is unavoidable. In fact, it is easy to see that the above estimates hold for N_2^u by noting that A^{-1}, K, and $[\partial_t, S_1]$ are bounded operators.

Bounding N_1^u is also easily done once we recognize that this term is almost exactly $\epsilon^{-11/2}\mathrm{Res}_u$. We have

$$\begin{aligned}
&\epsilon^{-11/2}A^{-1}\left(\left(1+L\Psi^z\right)\partial_t\epsilon^2\Psi^u+\partial_\alpha\Psi^y\left(1+\partial_t(K(z,y)\Psi^u)\right)\right)\\
&=\epsilon^{-11/2}A^{-1}\left(\left(1+L\Psi^z\right)\partial_t\epsilon^2\Psi^u+\partial_\alpha\Psi^y\left(1+\partial_t(K(\Psi^z,\Psi^y)\Psi^u)\right)\right)+N_3^u\\
&=\epsilon^{-11/2}\left(\partial_t\epsilon^2\Psi^u+\partial_\alpha\Psi^y\frac{1+\partial_t(K(\Psi^z,\Psi^y)\Psi^u)}{1+L\Psi^z}\right)+N_3^u+N_4^u\\
&=\epsilon^{-11/2}\left(\partial_t\epsilon^2\Psi^u+\partial_\alpha\Psi^y\frac{1+\partial_t^2(\Psi^y)}{1+L\Psi^z}\right)+N_3^u+N_4^u+N_5^u\\
&=\epsilon^{-11/2}\mathrm{Res}_u+N_3^u+N_4^u+N_5^u,
\end{aligned}$$

where

$$\begin{aligned}
N_3^u&=\epsilon^{-11/2}A^{-1}\left(\partial_\alpha\Psi^y\partial_t\left(K(z,y)\Psi^u-K(\Psi^z,\Psi^y)\Psi^u\right)\right),\\
N_4^u&=\epsilon^{-11/2}(A^{-1}-(1+L\Psi^z)^{-1})\\
&\quad\times\left(\left(1+L\Psi^z\right)\partial_t\epsilon^2\Psi^u+\partial_\alpha\Psi^y\left(1+\partial_t(K(\Psi^z,\Psi^y)\Psi^u)\right)\right),\\
N_5^u&=\epsilon^{-11/2}\partial_\alpha\Psi^y\frac{\partial_t(\mathrm{Res}_y)}{1+L\Psi^z}.
\end{aligned}$$

We bound N_3^u using mean value theorem arguments entirely analogous to those used when bounding N^y. To bound N_4^y, one observes that

$$\left(1+L\Psi^z\right)\partial_t\epsilon^2\Psi^u+\partial_\alpha\Psi^y\left(1+\partial_t(K(\Psi^z,\Psi^y)\Psi^u)\right)$$

is very nearly Res_u and is thus $O(\epsilon^{17/2})$. N_5^u is clearly small, as it contains $\partial_t\mathrm{Res}_y$. This completes the proof. ∎

We need to make analogous calculations for the four-dimensional system. Let

$$a(\alpha,t)=\epsilon^3\Psi^a(\epsilon\alpha,\epsilon t)+\epsilon^{11/2}R^a(\alpha,t)$$

and $\bar{R}_e=(R^z,R^y,R^u,R^a)$. This extended set of error functions lives in $\mathfrak{H}_e^s=H^s\times H^s\times H^{s-1/2}\times H^{s-1}$.

It is easy to see that

$$\partial_tR^u=R^a$$

but more difficult to determine the evolution of R^a. We begin by taking a time derivative of (29):

$$\partial_t^2u(1+Lz)+\partial_tu\partial_\alpha u+\partial_\alpha\partial_ty(1+\partial_t^2y)+\partial_\alpha y\partial_t^3y=0. \tag{30}$$

Letting

$$\begin{aligned}
I&=\partial_t^2u(1+Lz)+\partial_\alpha y\partial_t^3y,\\
II&=\partial_\alpha\partial_ty(1+\partial_t^2y)+\partial_tu\partial_\alpha u,
\end{aligned}$$

(30) is $I + II = 0$.

Manipulations very similar to those carried out in determing $\partial_t R^u$ show that

$$I = A(z,y)\partial_t \epsilon^{11/2} R^a + \partial_\alpha y[\partial_t^2, S_1(z,y)]\epsilon^{11/2} R^u \\ + (1+Lz)\partial_t \epsilon^3 \Psi^a + \partial_\alpha y \partial_t^2 (K(z,y)\Psi^u).$$

For II, we have

$$\begin{aligned} II &= (1+\partial_t^2 y)\partial_\alpha \partial_t \epsilon^{11/2} R^y + \partial_t u \partial_\alpha \epsilon^{11/2} R^u \\ &\quad + (1+\partial_t^2 y)\partial_\alpha \partial_t \Psi^y + \partial_t u \partial_\alpha \Psi^u \\ &= \epsilon^{11/2}\left((1+\partial_t^2 y)\partial_\alpha (K_0 R^u + K_1(\Psi^z,\Psi^y)R^u + N^y) + \partial_t u \partial_\alpha R^u\right) \\ &\quad + (1+\partial_t^2 y)\partial_\alpha \partial_t \Psi^y + \partial_t u \partial_\alpha \Psi^u \\ &= \epsilon^{11/2}\left((1+\partial_t^2 y - \partial_t u K_0 \cdot)K_0 \partial_\alpha R^u + \partial_\alpha K_1(\Psi^z,\Psi^y)R^u\right) \\ &\quad + (1+\partial_t^2 y)\partial_\alpha \partial_t \Psi^y + \partial_t u \partial_\alpha \Psi^u + \epsilon^{11/2} B_{II}, \end{aligned}$$

where

$$B_{II} = \partial_t u(1+K_0^2)\partial_\alpha R^u + \partial_t^2 y \partial_\alpha (K_1(\Psi^z,\Psi^y)R^u + N^y) + \partial_\alpha N^y.$$

Noting that $\partial_t^2 y = K_0 a + [\partial_t, S_1]u + S_1 a$, we see that B_{II} is smooth in the error functions, and is $O(\epsilon^3)$.

Adding I and II gives

$$\begin{aligned} 0 &= A(z,y)\partial_t R^a + (1+\partial_t^2 y - \partial_t u K_0 \cdot)K_0 \partial_\alpha R^u \\ &\quad + \partial_\alpha K_1(\Psi^z,\Psi^y)R^u + B, \end{aligned} \tag{31}$$

where

$$\begin{aligned} B &= B_{II} + \partial_\alpha y[\partial_t^2, S_1(z,y)]R^u + B_{\text{Res}}, \\ \epsilon^{11/2} B_{\text{Res}} &= (1+Lz)\partial_t \epsilon^3 \Psi^a + \partial_\alpha y \partial_t^2 (K(z,y)\Psi^u) \\ &\quad + (1+\partial_t^2 y)\partial_\alpha \partial_t \Psi^y + \partial_t u \partial_\alpha \Psi^u. \end{aligned}$$

The terms B_{II} and $\partial_\alpha y[\partial_t^2, S_1(z,y)]$ are small and smooth, and we can bound B_{Res} via the residual estimates, much as we did for N_1^u above. That is, we have

$$\|B\|_{s-1} \le C\left(\epsilon^3 + \epsilon^3 \|\bar{R}\|_{\mathfrak{H}_\epsilon^s} + \epsilon^{11/2}\|\bar{R}\|_{\mathfrak{H}_\epsilon^s}^2\right)$$

under the same hypotheses as in the above lemmas.

At this time, it is tempting to simply invert $A(z,y)$. Though we could do this, the inverse of this operator is not smoothing. In particular the presence of the term $\partial_\alpha y K_0$ in A will cause problems. We can eliminate K_0 to highest order by letting $H_0(z,y) = (1 + Lz - \partial_\alpha y K_0)$ act on (31). We have for the first term

$$\begin{aligned} &H_0(z,y)A(z,y)\partial_t R^a \\ &= (1+Lz)^2 \partial_t R^u + (1+Lz)\partial_\alpha y K_0 \partial_t R^a - \partial_\alpha y K_0\left((1+Lz)\partial_t R^a\right) \\ &\quad - \partial_\alpha y K_0\left(\partial_\alpha y K_0 \partial_t R^a\right) + H_0(z,y)(\partial_\alpha y S_1(z,y)\partial_t R^a) \\ &= \left((1+Lz)^2 + (\partial_\alpha y)^2 + H_1(z,y)\cdot\right)\partial_t R^a, \end{aligned}$$

where

$$\begin{aligned} H_1(z,y)\cdot &= \partial_\alpha y \left([Lz, K_0]\cdot - K_0[\partial_\alpha y, K_0]\cdot - (1+K_0^2)\partial_\alpha y \cdot\right) \\ &\quad + H_0(z,y)\left(\partial_\alpha y S_1(z,y)\cdot\right). \end{aligned}$$

Notice that H_1 is made up of smoothing operators and is thus a smoothing operator.

Now, for the second term in (31) we have

$$\begin{aligned} &H_0(z,y)(1+\partial_t^2 y - \partial_t u K_0\cdot)K_0\partial_\alpha R^u \\ &= (1+Lz)(1+\partial_t^2 y)K_0\partial_\alpha R^u - (1+Lz)\partial_t u K_0^2 \partial_\alpha R^u \\ &\quad - \partial_\alpha y K_0\left((1+\partial_t^2 y)K_0\partial_\alpha R^u\right) + \partial_\alpha y K_0 \left(\partial_t u K_0^2 \partial_\alpha R^u\right) \\ &= (1+Lz)(1+\partial_t^2 y)K_0\partial_\alpha R^u \\ &\quad - \partial_\alpha y [K_0, \partial_t^2 y]K_0 \partial_\alpha R^u + \partial_\alpha y K_0\left(\partial_t u K_0^2 \partial_\alpha R^u\right) \\ &\quad + \left(\partial_t u(1+Lz) + \partial_\alpha y(1+\partial_t^2 y)\right)K_0^2\partial_\alpha R^u. \end{aligned}$$

Notice that by comparing the last line of the above with (29), we see that it is identically zero! One more rearrangement of this yields

$$\left((1+Lz)(1+\partial_t^2 y) - \partial_\alpha y \partial_t u\right)K_0\partial_\alpha R^u + B_2,$$

where

$$B_2 = \partial_\alpha y\left(-[K_0, \partial_t^2 y] + [K_0, \partial_t u]K_0 + \partial_t u(1+K_0^2)\right)K_0\partial_\alpha u$$

is a smooth and small function by Lemma 4 in section 5.

If we let

$$\begin{aligned} f &= ((1+Lz)^2 + (\partial_\alpha y)^2)^{-1}, \\ g &= (1+Lz)(1+\partial_t^2 y) - \partial_\alpha y \partial_t u, \end{aligned}$$

then we have transformed (31) into

$$\begin{aligned} 0 &= (1+fH_1(z,y))\partial_t R^a + fgK_0\partial_\alpha R^u \\ &\quad + fH_0(z,y)\left(\partial_\alpha\left(K_1(\Psi^z,\Psi^y)R^u\right)\right) \\ &\quad + f\left(B_2 + H_0(z,y)B\right). \end{aligned}$$

By the Neumann series,

$$(1+fH_1(z,y)\cdot)^{-1} = 1 + \sum_{n=0}^{\infty}(-1)^n f^n H_1^n(z,y)\cdot.$$

Since H_1 is smoothing, this is the identity plus a smoothing piece. Let

$$H_2(z,y)\cdot = \sum_{n=0}^{\infty}(-1)^n f^n H_1^n(z,y)\cdot.$$

Thus,

$$\begin{aligned} 0 &= \partial_t R^a + fgK_0\partial_\alpha R^u \\ &\quad + fH_0(z,y)\left(\partial_\alpha\left(K_1(\Psi^z,\Psi^y)R^u\right)\right) - N_1^a - N_2^a, \end{aligned}$$

where

$$-N_1^a = H_2(z,y)\left(fgK_0\partial_\alpha R^u + fH_0(z,y)\left(\partial_\alpha\left(K_1(\Psi^z,\Psi^y)R^u\right)\right)\right),$$
$$-N_2^a = (1+H_2(z,y))f\left(B_2 + H_0(z,y)B\right).$$

Finally we rewrite the above as

$$\partial_t R^a = -\left(1+K_0a - Lz + N_s^a\right)\partial_\alpha K_0 R^u - \partial_\alpha\left(K_1(\Psi^z,\Psi^y)R^u\right) + N^a$$

with

$$N_s^a = fg - (1+K_0a - Lz),$$
$$N^a = N_1^a + N_2^a + N_3^a,$$
$$N_4^a = -\left(fH_0(z,y)-1\right)\left(\partial_\alpha\left(K_1(\Psi^z,\Psi^y)R^u\right)\right).$$

Notice that $(1+K_0a-Lz)$ is the first order approximation to fg. Thus, using techniques exactly like those we used in proving the bounds on N^y and N^u, we have the following.

LEMMA 13. *For all $C_R>0$, there exists ϵ_0 such that for all $\epsilon\in(0,\epsilon_0)$ and t such that* $\sup_{0\le t'\le t}\|\bar{R}_e(\cdot,t')\|_{\mathfrak{H}_e^s}\le C_R$ *we have*

$$\max\left\{\|N_s^a\|_{s-1},\|N^a\|_{s-1}\right\}\le C\left(\epsilon^3+\epsilon^3\|\bar{R}_e\|_{\mathfrak{H}_e^s}+\epsilon^{11/2}\|\bar{R}_e\|^2_{\mathfrak{H}_e^s}\right).$$

Recapping, we have shown that the three-dimensional system may be rewritten as

$$\begin{aligned}
&\partial_t R^z = K_0R^u,\\
&\partial_t R^y = K_0R^u + M_1(\Psi^z)\partial_\alpha R^u - (\Psi^y + K_0(\epsilon^2\Psi^yK_0))\partial_\alpha R^u + N^y,\\
&\partial_t R^u = -(1-\epsilon^2W_1)\partial_\alpha R^y + N^u,
\end{aligned}\tag{32}$$

and the four-dimensional system as

$$\begin{aligned}
\partial_t R^z &= K_0R^u,\\
\partial_t R^y &= K_0R^u + M_1(\Psi^z)\partial_\alpha R^u\\
&\quad - (\Psi^y + K_0(\epsilon^2\Psi^yK_0))\partial_\alpha R^u + N^y,\\
\partial_t R^u &= R^a,\\
\partial_t R^a &= -\left(1+K_0a-Lz+N_s^a\right)\partial_\alpha K_0R^u\\
&\quad - \partial_\alpha\left(K_1(\Psi^z,\Psi^y)R^u\right) + N^a.
\end{aligned}\tag{33}$$

We remark now that these are only cosmetically different from the equations which determine the evolution of the error for the KdV approximation alone in [28]. See page 1524 for the equations in three dimensions and page 1526 in four dimensions. Their variables

$$(Z_1, X_2, U_1, V_1)$$

correspond to our

$$(z, y, u, a),$$

and their functions

$$(N^2, N^3, N^4, N^5, N^8)$$

are our

$$(N^z, N^y, N^u, N_s^a, N^a).$$

The only difference of note is that their estimates contain a term they call $q(t)$ while ours do not. This term, which is related to the interaction of the left- and right-moving wavetrains, has been removed in this paper by the inclusion of the function W_3 in the approximating functions Ψ. This simplification does not adversely affect the means which they employ to prove that the error functions remain $O(1)$ over the long time scale. Therefore we appeal to their results on pages 1524–1533.

PROPOSITION 5. *For all $T_0 > 0$, $s > 4$, and $C_I > 0$, there exists ϵ_0 such that for all $0 \le \epsilon \le \epsilon_0$, the unique solution $\bar{R}_e$ of* (33)*, with initial conditions such that*

$$\|\bar{R}_e(\cdot, 0)\|_{\mathfrak{H}_e^s} \le C_I,$$

satisfies

$$\sup_{t\in[0,T_0\epsilon^{-3}]} \|\bar{R}_e(\cdot, t)\|_{\mathfrak{H}_e^s} \le C,$$

where C is independent of ϵ.

Implicit in the above proposition is the assumption that the initial conditions for the water wave problem have the form

$$\begin{pmatrix} z(\alpha,0) \\ y(\alpha,0) \\ u(\alpha,0) \end{pmatrix} = \begin{pmatrix} \Psi^z(\alpha,0) \\ \Psi^y(\alpha,0) \\ \Psi^u(\alpha,0) \end{pmatrix} + \epsilon^{11/2}\bar{R}_0(\alpha).$$

Thus we see that this proposition immediately proves Theorem 3.

Now that we have this result, there are a few small steps and one big step needed to prove Theorem 2. The first simple step is to note that the z is not a very physical coordinate and that we would prefer estimates for x_α. Since L is a bounded operator and gives the relationship between both z and x_α and Ψ^z and Ψ^d, we automatically have

$$\sup_{t\in[0,T_0\epsilon^{-3}]} \|x_\alpha(\cdot,t) - \Psi^d(\cdot,t)\|_{H^s} \le C\epsilon^{11/2}.$$

Second, the expressions for Ψ^d, Ψ^y, and Ψ^u contain terms of $O(\epsilon^6)$. These terms were needed to make the residual sufficiently small, but they are unnecessary now. Moreover, the appearance of the operator L^{-1} and inverse derivatives in the definitions of Ψ^y and Ψ^u is not very intuitive. Thus, it is a simple consequence of Lemma 1 and the triangle inequality that

$$\|\Psi^d - \psi^d\|_s \le C\epsilon^{11/2},$$
$$\|\Psi^y - \psi^y\|_s \le C\epsilon^{11/2},$$
$$\|\Psi^u - \psi^u\|_s \le C\epsilon^{11/2},$$

where ψ^d, ψ^y, and ψ^u were given in the introduction in (4), (5), and (6). Therefore we have the following corollary.

COROLLARY 1. *If the initial conditions for* (WW) *are of the form*

$$\begin{pmatrix} x_\alpha(\alpha,0) \\ y(\alpha,0) \\ u(\alpha,0) \end{pmatrix} = \begin{pmatrix} \psi^d(\alpha,0) \\ \psi^y(\alpha,0) \\ \psi^u(\alpha,0) \end{pmatrix} + \epsilon^{11/2}\bar{R}_1(\alpha) \tag{34}$$

with $\|\bar{R}_1\|_{\mathfrak{H}^s} \leq C_I$, *then the solution of* (WW) *satisfies the estimate*

$$\left\| \begin{pmatrix} x_\alpha(\cdot,t) \\ y(\cdot,t) \\ u(\cdot,t) \end{pmatrix} - \begin{pmatrix} \psi^d(\cdot,t) \\ \psi^y(\cdot,t) \\ \psi^u(\cdot,t) \end{pmatrix} \right\|_{\mathfrak{H}^s} \leq C_F \epsilon^{11/2}$$

for $t \in [0, T_0\epsilon^{-3}]$. *The constant* C_F *does not depend on* ϵ.

Finally, we must deal with initial conditions. Recall from the discussion in section 2 that it is typical to specify the initial data for the water wave problem in the long wavelength, small amplitude limit by

$$(\bar{x}_{\bar{\alpha}}(\bar{\alpha},0), \bar{y}(\bar{\alpha},0), \bar{u}(\bar{\alpha},0)) = (0, \epsilon^2\Theta_y(\epsilon\bar{\alpha}), \epsilon^2\Theta_u(\epsilon\bar{\alpha})). \tag{35}$$

However, the above results are applicable if the initial data is of the form seen in (34). We eliminate this discrepancy by altering the initial parameterization of the free surface. What should this change be? Clearly,

$$\bar{\alpha} = \alpha + x(\alpha,0). \tag{36}$$

Now set $U(\beta,0) = U_0(\beta)$, $V(\beta,0) = V_0(\beta)$, $F(\beta,0) = F_0(\beta)$, $G(\beta,0) = G_0(\beta)$, and $P(\beta,0) = 0$, and let

$$\begin{aligned} \bar{\alpha} &= \alpha + \int_0^\alpha \psi^d(a,0)da \\ &= \alpha + \epsilon X_1(\epsilon\alpha) + \epsilon^3 X_2(\epsilon\alpha), \end{aligned}$$

where

$$\begin{aligned} \epsilon X_1(\alpha) &= -\epsilon^2 \int_0^\alpha (U_0(\epsilon a) + V_0(\epsilon a))\, da, \\ \epsilon^3 X_2(\alpha) &= -\epsilon^4 \int_0^\alpha (F_0(\epsilon a) + G_0(\epsilon a))\, da. \end{aligned} \tag{37}$$

With this definition, we clearly have satisfied the first condition in (34). We also want

$$\begin{aligned} \Theta_y(\epsilon\bar{\alpha}) &= \epsilon^2\psi^y(\alpha) + O(\epsilon^{11/2}), \\ \Theta_u(\epsilon\bar{\alpha}) &= \epsilon^2\psi^u(\alpha) + O(\epsilon^{11/2}) \end{aligned}$$

or rather

$$\begin{aligned} \epsilon^2\Theta_y(\epsilon\alpha + \epsilon^2 X_1(\epsilon\alpha) + \epsilon^4 X_2(\epsilon\alpha)) &= \epsilon^2\psi^y(\alpha) + O(\epsilon^{11/2}), \\ \epsilon^2\Theta_u(\epsilon\alpha + \epsilon^2 X_1(\epsilon\alpha) + \epsilon^4 X_2(\epsilon\alpha)) &= \epsilon^2\psi^u(\alpha) + O(\epsilon^{11/2}). \end{aligned}$$

Applying Taylor's theorem we have

$$\begin{aligned}\Theta_y + \epsilon^2 X_1 \Theta_y' &= (U_0+V_0) + \epsilon^2\left(\frac{1}{3}\partial_{\beta_-}^2 U_0 + \frac{1}{3}\partial_{\beta_+}^2 V_0\right)\\ &\quad + \epsilon^2(F_0+G_0) + \epsilon^2(U_0+V_0)^2,\\ \Theta_u + \epsilon^2 X_1 \Theta_u' &= (U_0-V_0) + \epsilon^2\left(\frac{1}{6}\partial_{\beta_-}^2 U_0 - \frac{1}{6}\partial_{\beta_+}^2 V_0\right)\\ &\quad + \epsilon^2(F_0-G_0) + \epsilon^2\left(\frac{3}{4}U_0^2 - \frac{3}{4}V_0^2\right).\end{aligned}$$

We can solve the above by taking

$$\begin{aligned} U_0 &= 1/2(\Theta_y + \Theta_u),\\ V_0 &= 1/2(\Theta_y - \Theta_u)\end{aligned} \tag{38}$$

and

$$\begin{aligned} F_0 &= 1/2(h_y + h_u),\\ G_0 &= 1/2(h_y - h_u),\end{aligned} \tag{39}$$

where

$$\begin{aligned} h_y &= X_1\Theta_y' - \frac{1}{3}\partial_{\beta_-}^2 U_0 - \frac{1}{3}\partial_{\beta_+}^2 V_0 - (U_0+V_0)^2,\\ h_u &= X_1\Theta_u' - \frac{1}{6}\partial_{\beta_-}^2 U_0 + \frac{1}{6}\partial_{\beta_+}^2 V_0 - \frac{3}{4}U_0^2 + \frac{3}{4}V_0^2.\end{aligned}$$

The functions U_0, V_0, F_0, and G_0 are all in $H^s(4)$, and so the use of Taylor's theorem is justified by Lemma 10. Thus we have proven the following lemma.

LEMMA 14. *Given initial conditions for the water wave equation in the form* (35), *define U_0, V_0, F_0, G_0, X_1, and X_2 as in* (38), (39), *and* (37). *Then the reparameterization of the initial profile given by*

$$\bar{\alpha} = \alpha + \epsilon X_1(\epsilon\alpha) + \epsilon X_2(\epsilon\alpha)$$

results in initial conditions given by (34).

Remark 12. Let $\varphi^{\pm}(\beta_{\pm},0) = \varphi_0^{\pm}(\beta_{\pm})$. Then this lemma will still be true if we replace F_0 with φ_0^- and G_0 with φ^+ and set F_0 and G_0 to be identically zero. That is, we have some choice in the way we select the initial conditions for the higher order equations.

Combining this lemma with Corollary 1 we prove Theorem 2. Thus we are done.

7. Assorted proofs.

Proof of Lemma 8. Let $f^{-1}(\xi) = \xi - g_2(\xi)$. Since $f^{-1}(f(\alpha)) = \alpha$ we have

$$\alpha = f(\alpha) - g_2(f(\alpha)),$$

or rather

$$g_2(f(\alpha)) = g(\alpha). \tag{40}$$

Notice that this relation implies $\|g_2\|_{L^\infty} = \|g\|_{L^\infty}$. Taking a derivative, we have

$$g_2'(f(\alpha)) = \frac{g'(\alpha)}{1+g'(\alpha)},$$

which implies that $\|g_2'\|_{L^\infty} \leq C\|g'\|_{L^\infty}$. If we expand the left-hand side of (40) by the mean value theorem, we see

$$g_2(\alpha + g(\alpha)) = g(\alpha),$$
$$g_2(\alpha) + g_2'(\alpha^*)g(\alpha) = g(\alpha).$$

This implies $g_2(\alpha) = g(\alpha) + O(\|g'\|_{L^\infty}\|g\|_{L^\infty})$. Now, (40) can be rewritten and expanded using Taylor's theorem,

$$\begin{aligned} g_2(\xi) &= g(f^{-1}(\xi)) \\ &= g(\xi - g(\xi) + O(\|g'\|_{L^\infty}\|g\|_{L^\infty})) \\ &= g(\xi) + g'(\xi)(-g(\xi) + O(\|g'\|_{L^\infty}\|g\|_{L^\infty})) \\ &\quad + 1/2g''(\xi^*)(-g(\xi) + O(\|g'\|_{L^\infty}\|g\|_{L^\infty}))^2 \\ &= g(\xi) - g(\xi)g'(\xi) - E \end{aligned}$$

which completes the proof. □

Proof of Lemma 9. Since $(1+\beta^2)^{n/2}F(\beta) \in H^s$, by the Sobolev embedding theorem there is a C such that

$$F(\beta) \leq C(1+\beta^2)^{-n/2}.$$

So

$$\begin{aligned} &\int_{|\alpha|>C_0\epsilon^{-3}} |F(\epsilon\alpha)|^2 d\alpha \leq C\int_{|\alpha|>C_0\epsilon^{-3}} |1+(\epsilon\alpha)^2|^{-n} d\alpha \\ &\leq C\int_{|\alpha|>C_0\epsilon^{-3}} |\epsilon\alpha|^{-2n} d\alpha \\ &\leq C\epsilon^{-2n}\int_{|\alpha|>C_0\epsilon^{-3}} |\alpha|^{-2n} d\alpha \\ &\leq C\epsilon^{-2n}(\epsilon^{-3})^{-2n+1} \\ &\leq C\epsilon^{4n-3}. \end{aligned}$$

The higher derivatives are bounded in exactly the same fashion. The extra powers of ϵ come from the long wavelength scaling. □

Proof of Lemma 1. The proof is a straightforward calculation:

$$\begin{aligned} &\|Af(\cdot) - A_nf(\cdot)\|_s^2 \\ &= \int (1+k^2)^s|(\widehat{A}(k) - \widehat{A_n}(k))\widehat{f}(k)|^2 dk \\ &\leq C\int (1+k^2)^s|k^n\widehat{f}(k)|^2 dk \\ &= C\int (1+k^2)^s|\widehat{\partial_x^n f}(k)|^2 dk \\ &= C\|\partial_x^n f(\cdot)\|_s^2. \end{aligned}$$

The proof for long wavelength data follows immediately from this. □

Proof of Lemma 2. The fact that $g(\epsilon x)$ is bounded as such in L^2 follows automatically from the geometric series approximation. That is, since $|\epsilon^2 f(\epsilon x)| \leq 1/2$, we know that

$$|g(\epsilon x)| \leq C|\epsilon^2 f(\epsilon x)|^2.$$

Thus we have

$$\begin{aligned}\|g(\epsilon\cdot)\|_{L^2}^2 &\le C\int \epsilon^4|f(\epsilon x)|^4 dx\\ &\le C\epsilon^4\|f(\cdot)\|_{L^\infty}^2\int f^2(\epsilon x)dx\\ &\le C\epsilon^{7/2}\|f(\cdot)\|_s.\end{aligned}$$

Now consider the L^2 norm of $g'(\epsilon x)$. A direct calculation shows that

$$\frac{d}{dx}g(\epsilon x) = -\epsilon^3 f'(\epsilon x)\left(\frac{1}{(1+\epsilon^2 f(\epsilon x))^2} - 1\right).$$

Taylor's theorem shows that

$$\left|\frac{d}{dx}g(\epsilon x)\right| \le C|\epsilon^3 f'(\epsilon x)||\epsilon^2 f(\epsilon x)|.$$

Therefore, just as before we have that this is bounded by $C\epsilon^{9/2}$ (which is of course bounded by $C\epsilon^{7/2}$).

We could keep on going in this fashion, showing each derivative of g is bounded. This is, however, difficult as finding higher and higher derivatives is a notationally taxing job—see the expression of Faa di Bruno for proof of that!

Instead we take the following approach. Let

$$h(y) = \frac{1}{1+y} - 1 + y.$$

For $y \in [-1/2, 1/2]$, this function is real analytic and there exists another function $\tilde{h}(y)$ (real and analytic on the same interval) such that $h(y) = y\tilde{h}(y)$. Now, define $\tilde{h}_\epsilon(Y) = \tilde{h}(\epsilon^2 y)$. We have that

$$\|\tilde{h}_\epsilon(\cdot)\|_{C^s[-1/2\epsilon^{-2},1/2\epsilon^{-2}]} \le \|\tilde{h}(\cdot)\|_{C^s[-1/2,1/2]}.$$

The point here is that the C^s norm of $\tilde{h}_\epsilon$ can be bounded *independently* of ϵ.

Now notice that $g(X) = \epsilon^2 f(X)\tilde{h}_\epsilon(f(X))$. Since $f \in H^{s+1}$, we know that $f \in C^s$. This implies that $\tilde{h}_\epsilon(f(X)) \in C^s$, with the C^s norm bounded independent of ϵ. Thus we have $f(X)\tilde{h}_\epsilon(f(X)) \in H^s$, with a norm bounded independent of ϵ. With this in hand, we have that $\|g(\cdot)\|_s \le C\epsilon^2$, with $C \ne C(\epsilon)$. Now the derivatives of g can be bounded as follows:

$$\begin{aligned}\left\|\frac{d^n}{dx^n}g(\epsilon\cdot)\right\|_{L^2} &= \epsilon^n\|g^{(n)}(\epsilon\cdot)\|_{L^2}\\ &\le C\epsilon^{n-1/2}\|g(\cdot)\|_s\\ &\le C\epsilon^{n+3/2}.\end{aligned}$$

Provided $n \ge 2$, this term is small enough. Therefore we have shown that the first s derivatives are sufficiently small in L^2, and we have proved the lemma. $\square$

Acknowledgments. As this work was done principally as the author's dissertation, special thanks should go to Gene Wayne, who was the author's thesis advisor, and to the Department of Mathematics and Statistics at Boston University, where he attended graduate school.

REFERENCES

[1] J. T. Beale, *The existence of solitary water waves*, Comm. Pure Appl. Math., 30 (1977), pp. 373–389.

[2] W. Ben Youssef and T. Colin, *Rigorous derivation of Korteweg-de Vries-type systems from a general class of nonlinear hyperbolic systems*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 873–911.

[3] J. Bona, T. Colin, and D. Lannes, *Long wave approximations for water waves*, in Mathematical and Numerical Aspects of Wave Propagation—WAVES 2003, Springer, Berlin, 2003, pp. 489–494.

[4] J. L. Bona and M. Chen, *A Boussinesq system for two-way propagation of nonlinear dispersive waves*, Phys. D, 116 (1998), pp. 191–224.

[5] J. L. Bona, W. G. Pritchard, and L. R. Scott, *An evaluation of a model equation for water waves*, Philos. Trans. Roy. Soc. London Ser. A, 302 (1981), pp. 457–510.

[6] J. G. B. Byatt-Smith, *The reflection of a solitary wave by a vertical wall*, J. Fluid Mech., 197 (1988), pp. 503–521.

[7] J. G. B. Byatt-Smith, *The head-on interaction of two solitary waves of unequal amplitude*, J. Fluid Mech., 205 (1989), pp. 573–579.

[8] A. Constantin and W. Strauss, *Exact steady periodic water waves with vorticity*, Comm. Pure Appl. Math., 57 (2004), pp. 481–527.

[9] M. J. Cooker, P. D. Weidman, and D. S. Bale, *Reflection of a high-amplitude solitary wave at a vertical wall*, J. Fluid Mech., 342 (1997), pp. 141–158.

[10] W. Craig, *An existence theory for water waves and the Boussinesq and Korteweg-de Vries scaling limits*, Comm. Partial Differential Equations, 10 (1985), pp. 787–1003.

[11] W. Craig and M. D. Groves, *Hamiltonian long-wave approximations to the water-wave problem*, Wave Motion, 19 (1994), pp. 367–389.

[12] P. G. Drazin and R. S. Johnson, *Solitons: An Introduction*, Cambridge University Press, Cambridge, UK, 1989.

[13] H. R. Dullin, G. A. Gottwald, and D. D. Holm, *Camassa-Holm, Korteweg-de Vries-*5 *and other asymptotically equivalent equations for shallow water waves*, Fluid Dynam. Res., 33 (2003), pp. 73–95.

[14] K. O. Friedrichs and D. H. Hyers, *The existence of solitary waves*, Comm. Pure Appl. Math., 7 (1954), pp. 517–550.

[15] M. Haragus, D. P. Nicholls, and D. H. Sattinger, *Solitary wave interactions of the Euler-Poisson equations*, J. Math. Fluid Mech., 5 (2003), pp. 92–118.

[16] M. Hărăguş-Courcelle and D. H. Sattinger, *Inversion of the linearized Korteweg-de Vries equation at the multi-soliton solutions*, Z. Angew. Math. Phys., 49 (1998), pp. 436–469.

[17] L. A. Kalyakin, *Long-wave asymptotics. Integrable equations as the asymptotic limit of nonlinear systems*, Uspekhi Mat. Nauk, 44 (1989), pp. 5–34, 247.

[18] T. Kano and T. Nishida, *Sur les ondes de surface de l'eau avec une justification mathématique des équations des ondes en eau peu profonde*, J. Math. Kyoto Univ., 19 (1979), pp. 335–370.

[19] T. Kano and T. Nishida, *A mathematical justification for Korteweg-de Vries equation and Boussinesq equation of water surface waves*, Osaka J. Math., 23 (1986), pp. 389–413.

[20] Y. Kodama, *On integrable systems with higher order corrections*, Phys. Lett. A, 107 (1985), pp. 245–249.

[21] Y. Kodama and T. Taniuti, *Higher order corrections to the soliton-velocity and the linear dispersion relation*, Phys. Scripta, 20 (1979), pp. 486–489.

[22] D. Lannes, *Well-posedness of the water-waves equations*, J. Amer. Math. Soc., 18 (2005), pp. 605–654.

[23] T. Maxworthy, *Experiments on collisions between solitary waves*, J. Fluid Mech., 76 (1976), pp. 177–185.

[24] P. J. Olver, *Hamiltonian and non-Hamiltonian models for water waves*, in Trends and Applications of Pure Mathematics to Mechanics (Palaiseau, 1983), Lecture Notes in Phys. 195, Springer, Berlin, 1984, pp. 273–290.

[25] R. L. Sachs, *Completeness of derivatives of squared Schrödinger eigenfunctions and explicit*

solutions of the linearized KdV equation, SIAM J. Math. Anal., 14 (1983), pp. 674–683.
[26] R. L. Sachs, *A justification of the KdV approximation to first order in the case of N-soliton water waves in a canal*, SIAM J. Math. Anal., 15 (1984), pp. 468–489.
[27] G. Schneider, *The long wave limit for a Boussinesq equation*, SIAM J. Appl. Math., 58 (1998), pp. 1237–1245.
[28] G. Schneider and C. E. Wayne, *The long-wave limit for the water wave problem.* I. *The case of zero surface tension*, Comm. Pure Appl. Math., 53 (2000), pp. 1475–1535.
[29] G. Schneider and C. E. Wayne, *The rigorous approximation of long-wavelength capillary-gravity waves*, Arch. Ration. Mech. Anal., 162 (2002), pp. 247–285.
[30] C. H. Su and R. M. Mirie, *On head-on collisions between solitary waves*, J. Fluid Mech., 98 (1980), pp. 509–525.
[31] C. H. Su and R. M. Mirie, *Collisions between two solitary waves. Part* 2. *A numerical study*, J. Fluid Mech., 115 (1982), pp. 475–492.
[32] J. F. Toland, *Stokes waves*, Topol. Methods Nonlinear Anal., 7 (1996), pp. 1–48.
[33] C. E. Wayne and J. D. Wright, *Higher order modulation equations for a Boussinesq equation*, SIAM J. Appl. Dyn. Syst., 1 (2002), pp. 271–302.
[34] J. D. Wright, *Higher Order Corrections to the KdV approximation for Water Waves*, Ph.D. thesis, Boston University, 2004.
[35] S. Wu, *Well-posedness in Sobolev spaces of the full water wave problem in* 2-*D*, Invent. Math., 130 (1997), pp. 39–72.
[36] S. Wu, *Well-posedness in Sobolev spaces of the full water wave problem in* 3-*D*, J. Amer. Math. Soc., 12 (1999), pp. 445–495.

SIAM J. MATH. ANAL.
Vol. 37, No. 4, pp. 1207–1226

SELF-SIMILAR EXPANDING SOLUTIONS IN A SECTOR FOR A CRYSTALLINE FLOW*

MI-HO GIGA†, YOSHIKAZU GIGA‡, AND HIDEKATA HONTANI§

Abstract. For a given sector a self-similar expanding solution to a crystalline flow is constructed. The solution is shown to be unique. Because of self-similarity the problem is reduced to solve a system of algebraic equations of degree two. The solution is constructed by a method of continuity and obtained by solving associated ordinary differential equations. The self-similar expanding solution is useful to construct a crystalline flow from an arbitrary polygon not necessarily admissible.

Key words. crystalline flow, self-similar expanding solutions, a priori estimate

AMS subject classifications. 15A99, 34A12, 74N05, 94A08

DOI. 10.1137/040614372

1. Introduction.

A curvature flow is important to describe motion of phase boundaries in material sciences [13], [22] and also to modify contours in image analysis [18]. A crystalline flow is considered as a discrete version of an anisotropic curvature flow in the plane. It was introduced by Taylor [21] and independently by Angenent and Gurtin [1] almost a decade ago. Let us give a typical example of anisotropic curvature flow equations in $\mathbf{R}^2$:

$$V = -\mathrm{div}\xi(\vec{n}) \quad \text{on } \Gamma_t; \tag{1.1}$$

here V denotes the normal velocity of an evolving curve $\{\Gamma_t\}$ in the direction of unit normal $\vec{n}$ and $\xi = \nabla\gamma$ with the interfacial energy density $\gamma(p)$ which is positively one-homogeneous and convex in $\mathbf{R}^2$; div denotes the surface divergence. If $\gamma(p) = |p|$, then (1.1) is nothing but the curve shortening equation. If γ is piecewise linear, then (1.1) is so singular that it cannot be interpreted as a conventional partial differential equation. This is a typical situation where a crystalline flow arises. To understand (1.1) one restricts $\{\Gamma_t\}$ in a special class of polygonal curves, which is often called "admissible" [21], [1]. The boundary of the Wulff shape

$$W_\gamma = \{x \in \mathbf{R}^2;\ x \cdot m \le \gamma(m) \quad \text{for all} \quad m \in \mathbf{R}^2\}$$

*Received by the editors September 3, 2004; accepted for publication (in revised form) February 22, 2005; published electronically January 6, 2006. This work was partly supported by grant-in-aid for formation of COE "Mathematics of Nonlinear Structures via Singularities" (Hokkaido University).
http://www.siam.org/journals/sima/37-4/61437.html

†Department of Mathematics, Hokkaido University, Sapporo 060-0810, Japan. Current address: Graduate School of Mathematics, University of Tokyo, Komaba 3-8-1, Meguro, Tokyo 153-8914, Japan (mihogiga@ms.u-tokyo.ac.jp). The work of this author was supported in part by the 21st century COE program at Graduate School of Mathematical Sciences, the University of Tokyo.

‡Department of Mathematics, Hokkaido University, Sapporo 060-0810, Japan. Current address: Graduate School of Mathematical Sciences, University of Tokyo, Komaba 3-8-1, Meguro, Tokyo 153-8914, Japan (labgiga@ms.u-tokyo.ac.jp). The work of this author was partly supported by grants-in-aid for scientific research 1420411 and 1563408 from the Japan Society for the Promotion of Science (JSPS).

§Department of Informatics, Yamagata University, Yonezawa, Yamagata, 992-8510, Japan. Current address: Department of Computer Science and Engineering, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan (h.hontani@ieee.org)

is a typical example of an admissible polygon (crystal). Its weighted curvature $-\mathrm{div}\xi(\vec{n})$ formally equals -1 if $\vec{n}$ is taken outward. Thus if W_γ is a regular polygon centered at the origin, then it is reasonable to say that $\Gamma_t = t^{1/2}\partial W_\gamma$ is a "solution" of (1.1). We say that a polygon Γ is an *admissible crystal* if the orientation of each edge (facet) is one of that in ∂W_γ and the orientations of adjacent facets should be adjacent in ∂W_γ. We say that $\{\Gamma_t\}$ is an *admissible evolving crystal* if Γ_t is an admissible crystal and the motion of all vertices of Γ_t is C^1 in time t. Let $S_j(t)$ denote the jth facet of Γ_t. Then (1.1) is formally of the form

$$V_j(t) = \Lambda_j(t) \quad \text{on } S_j(t) \tag{1.2}$$

with *crystalline curvature*

$$\Lambda_j(t) = \chi_j \Delta(\vec{n}_j)/L_j(t),$$

and $V_j(t)$ denotes the normal velocity of $S_j(t)$ in the direction of $\vec{n}_j$; here $L_j(t)$ denotes the length of $S_j(t)$ and χ_j is the transition number (see section 2). The quantity $\Delta(\vec{n}_j)$ is the length of the facet of ∂W_γ whose orientation equals $\vec{n}_j$. Together with a transport equation (the first displayed formula in section 3) we have a finite system (3.1) of ordinary differential equations (ODEs), which is at least solvable locally in time. The resulting flow $\{\Gamma_t\}$ is often called a crystalline flow.

Our goal in this paper is to construct a self-similar expanding admissible evolving crystal $\{\Gamma_t\}_{t>0}$ satisfying (1.2) (shortly, self-similar expanding solution) such that it tends to the boundary of a sector as time tends to zero. We also prove its uniqueness. (Note that a self-similar expanding solution is not necessarily a dilation of the Wulff shape as numerical calculations in [15], [16] show.) The unique existence of self-similar expanding solutions has been claimed in a pioneering work by Taylor [23, Proposition 2.2(1)]. However, unfortunately the proof is rather sketchy; for example the uniqueness is not proved. It is not at all clear how to complete the proof. In this paper we shall give a complete proof.

There are several reasons such a solution is important. Here is a partial list:

(a) It is useful to construct a solution when initial data is a nonadmissible polygon;

(b) It is useful to construct a self-similar expanding solution to (1.1) for general γ when Γ_t lives in a sector and touches the boundary of the sector with prescribed contact angle.

The first reason (a) is already explained in [23, section 2.2].

Before discussing these reasons we mention relation between crystalline flow and curvature flow with smooth strictly convex interfacial energy density. One comes up with two natural questions.

(i) Is a crystalline flow approximated by a curvature flow with smooth convex γ?

(ii) Is a flow with smooth γ approximated by a crystalline flow with a piecewise linear γ? If so, crystalline flow provides a good numerical algorithm to calculate (1.1).

Fortunately, these problems are affirmatively settled by now; [5], [7], [8], [9] for (i) and [5], [12], [11], [7], [8], [9] for (ii). In [5], [7], [9] general notions of a "solution" to (1.1) based on variational principle or comparison principle are given. Moreover, it is shown that the "solution" in this sense exists for a general initial simple curve not necessarily admissible. Starting from a general polygon is important for a crystalline algorithm (ii). Practically, for a given curve it is easier to approximate it by a polygon rather than by an admissible crystal. However, it is not clear what is a crystalline flow starting from a general polygon although existence of a "solution" is known in an abstract level [5], [6], [9]. It is also numerically calculated in [4] when Γ_t is a graph by a variational inequality.

We now go back to the point (a). Suppose that the initial polygon is nonadmissible and that crystalline curvature of a pair of adjacent facets S_A and S_B equals zero. Assume that between orientations of S_A and S_B there are orientations of ∂W_γ. Then one expects that some new facets (with their missing orientations) are created instantaneously. If one has a self-similar expanding solution (with respect to a point where S_A intersects S_B), then this solution provides newly created facets. (In fact, it is a unique solution in the level set sense [8], [9]. Even if the crystalline curvature of at least one of S_A and S_B is not zero, a self-similar expanding solution represents a leading term of the length of newly created facets. We do not touch these two problems in this paper.) A self-similar expanding solution gives a definite way to create new facets (at least when S_A and S_B do not move), so it is useful to implement numerical algorithm starting with a nonadmissible polygon. We gave such an applications in [15], [16] for multiscale analysis of contour shape to extract its structure.

When the Frank diagram of γ (1-level set of γ) is strictly convex and smooth so that W_γ is smooth and strictly convex, we know that there is a (unique) self-similar expanding solution to (1.1) touching the boundary of a given sector with prescribed angle [17], [3]. Our self-similar expanding solution to (1.2), together with approximation theory [7], [9], provides a self-similar expanding solution (under prescribed contact angle condition) for more general γ whose Frank diagram is not necessarily strictly convex. As in [17], [3] our self-similar solutions represents the large time behavior of general solutions. We shall discuss this topic in a forthcoming paper. Note that our self-similar expanding solution is different from a self-similar shrinking solution studied, for example, in [19], [20].

Let us briefly mention the method of the proof. The uniqueness is proved by a geometric method (section 2)—comparison principle [21], [10]. To construct a self-similar expanding solution it suffices to find a solution of ODEs with $L_j(t) = t^{1/2}/a_j$, $a_j > 0$. The problem is reduced to find a positive a_j's solving a system of algebraic equations of degree two for a_j's. Essentially the same reduction is done by Taylor [23, p. 425]. If the number of missing orientations is small, say one or two, then one can solve it directly. However, solving the algebraic system for a greater number of equations is very tedious. The equation for a_j's is of the form

$$\begin{pmatrix} 1/a_1 \\ \vdots \\ 1/a_n \end{pmatrix} = H \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \tag{1.3}$$

and $H = (H_{ij})$ is a tridiagonal matrix; see (3.2). In [23, p. 425] a method to solve such a system of algebraic equations is suggested but the proof seems to be too short to achieve the goal. We solve this equation analytically by introducing extra parameter s, so that our algebraic system is decoupled for $s = 0$ to get $a_j = 1/\sqrt{H_{jj}}$ and for $s = 1$ it agrees with (3.2). We differentiate it with respect to s to get ODEs. We solve the ODEs from $s = 0$ to $s = 1$ by establishing a priori estimates. A crucial step is to calculate the determinant of H, in particular to prove its positiveness. An explicit and beautiful formula of $\det H$ is given in Lemma 4.2 up to an explicit constant. It is represented by angles of Wulff shapes. Note that our method provides not only the existence but also a way to calculate numerical values of a_j's by solving the ODEs numerically.

This paper is organized as follows. In section 2 we state our main results and give a proof of uniqueness. In section 3 we derive the ODEs to solve the algebraic equation (3.1). We prove the existence of a solution admitting several estimates for matrices established in section 4, which is the main technical part of this paper.

After this work was completed, we were informed of the work of Campbell [2] who partially solved our problem. In fact, she proved the unique solvability of (1.3) when the Wulff shape is a regular polygon. Her method is a kind of shooting method for a_j. When the Wulff shape is not a regular polygon, it seems that her method does not apply. Moreover, unfortunately her work is only in the form of a master's thesis and is not published. We believe it is very important to publish a complete proof in a full generality of this fundamentally important problem.

2. Main theorems. We start by formulating the problem. Let ∂C be the boundary of a given oriented cone C in $\mathbf{R}^2$ of the form $\partial C = \ell_A \cup \ell_B$, where ℓ_A and ℓ_B are maximal half lines starting from the origin O and are indexed clockwise as in Figure 1. We also call C a sector. In this paper, we assume that $\vec{n}_j = (\cos\theta_j, \sin\theta_j)$ is the outer unit normal of ℓ_j for $j = A, B$ with $|\theta_A - \theta_B| < \pi$. Let n be a nonnegative integer. Let $\Theta = \{\theta_j;\ j = 1, \dots, n\}$ with $\theta_A > \theta_1 > \theta_2 > \cdots > \theta_n > \theta_B$ (resp., $\theta_A < \theta_1 < \theta_2 < \cdots < \theta_n < \theta_B$) if $\theta_A > \theta_B$ (resp., $\theta_A < \theta_B$). We call Θ a *set of admissible angles*. We interpret that Θ is an empty set if $n = 0$.

We call a simple oriented polygonal curve S an *admissible crystal associated with* C if S is of the form $S = \cup_{j=1}^n S_j \ \cup S_A \ \cup S_B$, where S_j is a nontrivial and closed segment with the outer unit normal $\vec{n}_j = (\cos\theta_j, \sin\theta_j)$ for $j \in \{1, \dots, n\} \cup \{A, B\}$ and S_j for $j = A, B$ is a half line contained in ℓ_j. If $n = 0$, we always use the convention that $\cup_{j=1}^n S_j$ is the empty set so that $S = \partial C$. We implicitly assume that segments S_j's are numbered clockwisely. Figure 1 shows examples of C. Figure 2 shows examples of admissible crystals S associated with C.

We say that a family of polygon $\{S(t)\}_{t\in J}$, where J is a time interval, belongs to a *set of orientation-preserving evolving curves* $\mathcal{S}$ if $S(t)$ is an admissible crystal for all $t \in J$ and each corner moves continuously differentiably in time, where J is a time interval. These conditions imply that the orientation of each line (facet) is preserved for $t \in J$.

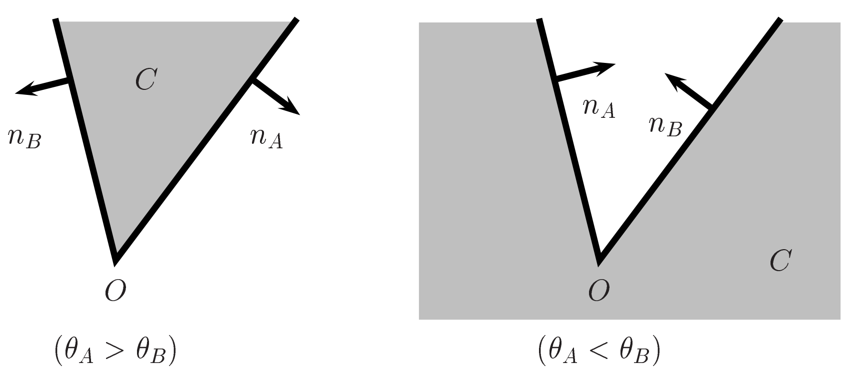


FIG. 1. *Oriented cones C.*

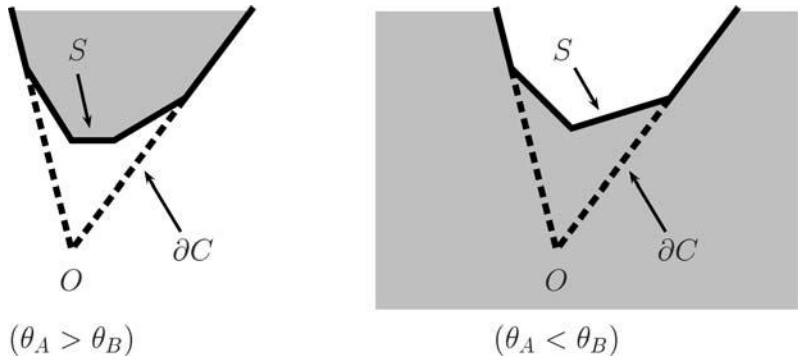


FIG. 2. *Admissible crystals S.*

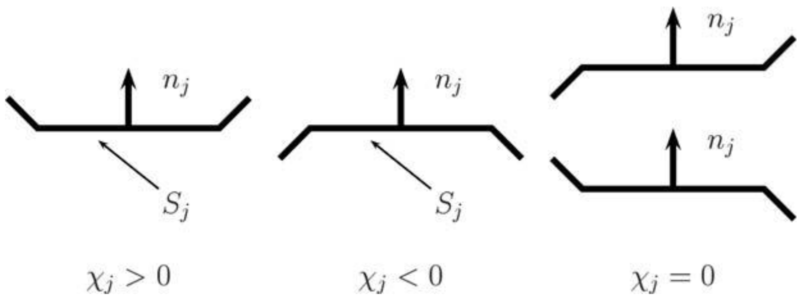


FIG. 3. *Transition number χ_j.*

We recall the notion of crystalline curvature or weighted curvature of $\{S(t)\}_{t\in J} \in S$, where $S(t)$ is of the form

$$S(t) = \bigcup_{j=1}^{n} S_j(t) \cup S_A(t) \cup S_B(t)$$

and $L_j(t)$ is the length of facet $S_j(t)$. Let $\chi_j(t)$ be a transition number defined by

$$\chi_j(t) := \begin{cases} 1 & \text{if } S(t) \text{ is concave around } S_j(t), \\ -1 & \text{if } S(t) \text{ is convex around } S_j(t), \\ 0 & \text{otherwise} \end{cases}$$

(see Figure 3). Let Δ_j be a positive given number. We define these quantities for $j = 1, 2, \ldots, n$. A crystalline curvature $\Lambda_j(t)$ of $S_j(t)$ is defined by

$$\Lambda_j(t) = \frac{\chi_j(t)\Delta_j}{L_j(t)}.$$

We use this formula to define crystalline curvatures Λ_A and Λ_B for S_A and S_B, and observe that $\Lambda_A = \Lambda_B = 0$ no matter how transition number and Δ are taken since the length of S_A and S_B is infinite.

We consider an evolving curve $\{S(t)\}_{t\in J} \in \mathcal{S}$ solving (1.2). In other words, it evolves such that the normal velocity $V_j(t)$ in the direction of $\overrightarrow{n}_j$ of jth facet $S_j(t)$ equals $\Lambda_j(t)$ for $j = 1, \ldots, n, A, B$.

We will now introduce the notion of a self-similar expanding solution.

DEFINITION 2.1. *An orientation-preserving evolving curve $\{S(t)\}_{t>0} \in \mathcal{S}$ is called a* self-similar expanding solution *to* (1.2) *in a sector C if there exists an admissible crystal S_* associated with C such that*

$$S(t) = t^{1/2} S_* = \{t^{1/2}x; x \in S_*\} \quad \textit{for} \quad t > 0; \tag{2.1}$$

$$V_j(t) = \Lambda_j(t) \quad \textit{for} \quad t > 0, \quad j = 1, \dots, n \quad \textit{if} \quad n \geq 1. \tag{2.2}$$

By definition if $n = 0$, then $S_* = \ell_a \cup \ell_b = \partial C$. Thus the only self-similar expanding solution in a sector C must be ∂C.

We note that for a self-similar solution $\{S(t)\}_{t>0}$, the transition number is unique independent of j and t, i.e., $\chi_j(t) = -1$ (resp., 1) if $\theta_A > \theta_B$ (resp., $\theta_A < \theta_B$) for all $j = 1, \dots, n$ and $t > 0$.

Our main results concern the existence and uniqueness of self-similar expanding solutions governed by crystalline curvature in a sector. They are as follows.

THEOREM 2.2 (existence). *Let C be a given oriented cone in $\mathbf{R}^2$. Let $\Theta = \{\theta_j;\ j = 1, \dots, n\}$ (with nonnegative integer n) be a set of admissible angles. Let Δ_j be a positive number for $j = 1, \dots, n$. Then there exists a self-similar expanding solution $\{S(t)\}_{t>0}$ in a sector C.*

THEOREM 2.3 (uniqueness). *Under the same hypotheses of Theorem* 2.2 *there is at most one self-similar expanding solution $\{S(t)\}_{t>0}$ in a sector C.*

We shall prove Theorem 2.2 in section 3 based on key a priori estimates shown in section 4. In the rest of this section we shall prove Theorem 2.3 by geometric argument.

Proof of Theorem 2.3. Let $\{S(t)\}_{t>0}$ and $\{R(t)\}_{t>0}$ be self-similar expanding solutions in a sector C. We may assume that $\theta_A > \theta_B$; i.e., the cone C is convex. We may also assume that $n \geq 1$. Then transition numbers of all facets of $S(t)$ and $R(t)$ are -1. Let $S(t)$ (resp., $R(t)$) be of the form $S(t) = \cup_{j=1}^n S_j(t) \cup S_A(t) \cup S_B(t)$ (resp., $R(t) = \cup_{k=1}^n R_k(t) \cup R_A(t) \cup R_B(t)$). For convenience we introduce an unbounded region $D_S(t) \subset \mathbf{R}^2$ enclosed by $S(t)$ for $t > 0$; let $D_S(t)$ denote the closure of the interior region bounded by curve $S(t)$ (see Figure 4). Let $\tilde{R}(t)$ be of the form $\tilde{R}(t) = \cup_{k=1}^n R_k(t)$.

Suppose that $S \not\equiv R$. We may assume that $\tilde{R}(1) \cap \text{ int } D_S(1) \neq \emptyset$. By this assumption, $t_0 \in (0,1)$ holds for $t_0 := \sup\{t | \tilde{R}(t) \cap \text{ int } D_S(1) = \emptyset\}$. Since $\tilde{R}(0)$ is a singleton and $S(1 - t_0) \neq \partial C$, there exists $\delta > 0$ such that $\tilde{R}(\delta) \cap \text{ int } D_S(t_1) = \emptyset$

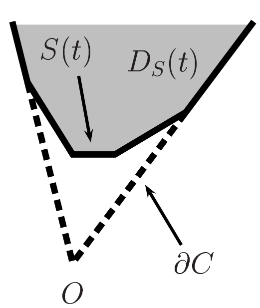


FIG. 4. *Region $D_S(t)$ enclosed by $S(t)$.*

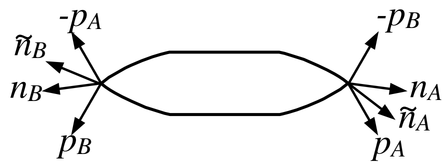


FIG. 5. *Wulff shape W.*

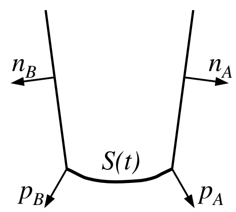


FIG. 6. *Self-similar expanding solution $S(t)$.*

with $t_1 := 1 - t_0$. We fix such δ. Setting $t_2 := \sup\{\tau > 0;\ \tilde{R}(\sigma+\delta) \cap \operatorname{int} D_S(\sigma+t_1) = \emptyset$ for $\sigma \in (0,\tau)\}$, we have $0 < t_2 < t_0$. Since $\tilde{R}(\sigma+\delta)$ touches $D_S(\sigma+t_1)$ first time at $\sigma = t_2$, there exists a facet $R_j(t_2+\delta)$ of $\tilde{R}(t_2+\delta)$ and a facet $S_j(t_2+t_1)$ of $S(t_2+t_1)$ such that the normal of $R_j(t_2+\delta)$ coincides with that of $S_j(t_2+t_1)$, we conclude that $R_j(t_2+\delta) \cap S_j(t_2+t_1) \neq \emptyset$ and that the length of $R_j(t_2+\delta)$ does not equal the length of $S_j(t_2+t_1)$. By geometry, the length of $R_j(t_2+\delta)$ is greater than the length of $S_j(t_2+t_1)$, so that the weighted curvature of $R_j(t_2+\delta)$ is negative and is greater than that of $S_j(t_2+t_1)$ (cf. [21], [10]). So the normal velocity of $R_j(t_2+\delta)$ is negative and is greater than that of $S_j(t_2+t_1)$, which contradicts the definition of t_2. $\square$

Remark. (i) The evolution equation (2.2) can be viewed as a crystalline curvature flow equation (1.1) (or (1.2)) with a suitable polygonal Wulff shape. Indeed, if $\theta_A > \theta_B$ for example, then there exists a convex polygon W such that the set $\mathcal{N}$ of the orientations of all facets in W includes all $\vec{n}_j$'s with $j \in \{1,\ldots,n\}$ and the length of facet with $\vec{n}_j$ equals Δ_j and that $\mathcal{N}$ does not include any $\vec{m} = (\cos\theta, \sin\theta)$ for $\theta \neq \theta_j, \theta \in (\theta_B, \theta_A)$. We may assume that W contains the origin as an interior point. The corresponding interfacial energy density γ is given as a support function: $\gamma(x) = \sup\{x \cdot p; p \in W\}$ for $x \in \mathbf{R}^2$. The case $\theta_A < \theta_B$ can be treated in a similar way.

(ii) If we dilate the Wulff shape so that the length of the jth facet equals $\lambda\Delta_j$, then $\sqrt{\lambda}S(t)$ is the corresponding self-similar expanding solution to (1.2) with Δ_j replaced by $\lambda\Delta_j$, where $S(t)$ is defined by (2.1) with (2.2).

(iii) Here is a numerical example of a profile S_* of the self-similar expanding solutions for given two different sectors with a fixed Wulff shape having many facets so that it looks a smooth curve. See Figures 5, 6, and 7. We use a Newton-type iteration which is closely related to our ODEs (3.4) to find numerical values of a_j's.

3. Existence of self-similar solution of ODE system. We shall show the existence theorem (Theorem 2.2). When $n = 0$, $\{S(t)\}_{t>0}$ with $S(t) = t^{1/2}(\partial C)$ is the desired self-similar expanding solution. In the following we suppose $n \geqq 1$. Let a family of polygon $\{S(t)\}_{t>0}$ belong to $\mathcal{S}$. Then we have a transport equation of

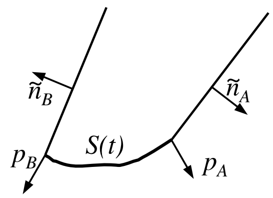


FIG. 7. *Self-similar expanding solution $S(t)$.*

$\{S(t)\}_{t>0}$:

$$\frac{dL_j(t)}{dt} = (\cot\varphi_j + \cot\varphi_{j+1})V_j(t) - \frac{1}{\sin\varphi_j}V_{j-1}(t) - \frac{1}{\sin\varphi_{j+1}}V_{j+1}(t), \quad t>0$$

for $j = 1,\dots,n$, where $\varphi_j := \theta_j - \theta_{j-1}$ for $j = 1,\dots,n+1$. Here we set $\theta_0 := \theta_A, \theta_{n+1} := \theta_B$ so that $V_0 := 0$ and $V_{n+1} := 0$ since $S_A(t)$ and $S_B(t)$ have infinite length. Plugging the governing law $V_j(t) = \Lambda_j(t)$ for $j = 1,\dots,n$, we have an ODE system:

$$\frac{dL_j(t)}{dt} = \frac{1}{2}\left\{\frac{p_j}{L_j(t)} + \frac{q_{j-1}}{L_{j-1}(t)} + \frac{r_{j+1}}{L_{j+1}(t)}\right\}, \quad t>0, \tag{3.1}$$

for $j = 1,\dots,n$, where

$$\begin{aligned} p_j &= 2\chi_j\Delta_j\frac{\sin(\varphi_j+\varphi_{j+1})}{\sin\varphi_j\sin\varphi_{j+1}} && \text{for } j = 1,\dots,n,\\ q_j &= -2\chi_j\Delta_j\frac{1}{\sin\varphi_{j+1}} && \text{for } j = 1,\dots,n-1,\\ r_j &= -2\chi_j\Delta_j\frac{1}{\sin\varphi_j} && \text{for } j = 2,\dots,n, \end{aligned}$$

and $q_0 = 0$ and $r_{n+1} = 0$. Here we used $\cot\varphi_j + \cot\varphi_{j+1} = \frac{\sin(\varphi_j+\varphi_{j+1})}{\sin\varphi_j\sin\varphi_{j+1}}$.

By the assumption on θ_A and θ_B in section 2, we note that $|\sum_{j=1}^{n+1}\varphi_j| = |\theta_A - \theta_B| < \pi$. When C is convex (resp., concave), i.e., $\theta_A > \theta_B$ (resp., $\theta_A < \theta_B$), then $\varphi_j < 0$ (resp., $\varphi_j > 0$) for $j = 1,\dots,n+1$ and $\chi_j < 0$ (resp., $\chi_j > 0$) for $j = 1,\dots,n$. Thus, we have $p_j > 0$ for $j = 1,\dots,n$, $q_j < 0$ for $j = 1,\dots,n-1$, and $r_j < 0$ for $j = 2,\dots,n$.

DEFINITION 3.1. *A family of functions $\{L_j(t)\}_{j=1}^n$ is called a* self-similar solution *of the ODE system* (3.1) *if $L_j(t)$ is of the form $L_j(t) = \alpha_j t^{1/2}$ with positive number α_j satisfying* (3.1) *for $j = 1,\dots,n$.*

Theorem 2.2 is obtained by showing the following proposition.

PROPOSITION 3.2. *Let n be a positive integer. There exists a self-similar solution $\{L_j(t)\}_{j=1}^n$ of the ODE system* (3.1).

When $n = 1$, (3.1) yields an ODE

$$\frac{dL_1(t)}{dt} = \frac{p_1}{2L_1(t)}, \quad t>0.$$

Since $(d/dt)\{L_1(t)^2\} = p_1$, $t>0$, we obtain $L_1(t) = t^{1/2}{p_1}^{1/2}$ for $t>0$, so that Proposition 3.2 holds for $n = 1$.

In the following we assume that $n \geqq 2$. Our strategy to prove Proposition 3.2 is as follows. Substituting $L_j(t) = \alpha_j t^{1/2}$ into (3.1), we have

$$\begin{pmatrix} 1/a_1 \\ 1/a_2 \\ \vdots \\ 1/a_n \end{pmatrix} = H \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \tag{3.2}$$

with unknowns $a_j = 1/\alpha_j$, where

$$H = \begin{pmatrix} p_1 & r_2 & & & \\ q_1 & p_2 & r_3 & & \mathbf{0} \\ & \ddots & \ddots & \ddots & \\ & & q_{n-2} & p_{n-1} & r_n \\ & \mathbf{0} & & q_{n-1} & p_n \end{pmatrix}.$$

In particular

$$H = \begin{pmatrix} p_1 & r_2 \\ q_1 & p_2 \end{pmatrix}$$

when $n = 2$. To show the existence of solutions to the system of nonlinear algebraic equations (3.2), we consider the following continuation method sometimes called Davidenko's method. Introducing an extra parameter $s \geq 0$ and the matrix $K(s)$:

$$K(s) = \begin{pmatrix} p_1 & sr_2 & & & \mathbf{0} \\ sq_1 & p_2 & sr_3 & & \\ & \ddots & \ddots & \ddots & \\ & & sq_{n-2} & p_{n-1} & sr_n \\ \mathbf{0} & & & sq_{n-1} & p_n \end{pmatrix},$$

we consider the system of nonlinear algebraic equations

$$\begin{pmatrix} 1/b_1(s) \\ 1/b_2(s) \\ \vdots \\ 1/b_n(s) \end{pmatrix} = K(s) \begin{pmatrix} b_1(s) \\ b_2(s) \\ \vdots \\ b_n(s) \end{pmatrix} \tag{3.3}$$

for $b(s) > 0$. Evidently $b_j(0) = 1/\sqrt{p_j}$, since $K(0)$ is a diagonal matrix. If the solution can be extended up to $s = 1$, then $b_j(1)$ is a solution of (3.2) since $K(1) = H$. Differentiating (3.3) formally with respect to parameter s, we have

$$-\begin{pmatrix} \vdots \\ b_j'(s)/b_j(s)^2 \\ \vdots \end{pmatrix} = K(s) \begin{pmatrix} \vdots \\ b_j'(s) \\ \vdots \end{pmatrix} - J \begin{pmatrix} \vdots \\ b_j(s) \\ \vdots \end{pmatrix}$$

with

$$J = -\begin{pmatrix} 0 & r_2 & & & & \\ q_1 & 0 & r_3 & & \text{\huge 0} & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & \text{\huge 0} & & q_{n-2} & 0 & r_n \\ & & & & q_{n-1} & 0 \end{pmatrix}.$$

Here and thereafter $'$ denotes d/ds, differentiation with respect to s. One can rewrite the above differential equation for b_j in the form

$$Q(s, \vec{b}(s)) \begin{pmatrix} \vdots \\ b_j'(s) \\ \vdots \end{pmatrix} = J \begin{pmatrix} \vdots \\ b_j(s) \\ \vdots \end{pmatrix}, \tag{3.4}$$

i.e.,

$$Q(s, \vec{b}(s))\ \vec{b}'(s) = J\vec{b}(s) \quad \text{with} \quad \vec{b}(s) = {}^{\mathrm{t}}(b_1(s), \ldots, b_n(s)),$$

if we define

$$Q(s, \vec{h}) = K(s) + \mathrm{diag}(1/h_1^2, \ldots, 1/h_n^2) \quad \text{with} \quad \vec{h} = {}^{\mathrm{t}}(h_1, \ldots, h_n).$$

If the inverse matrix of Q exists, (3.4) formally yields

$$\vec{b}'(s) = G(s, \vec{b}(s))\ \vec{b}(s). \tag{3.5}$$

Here we set $G(s, \vec{h}) = Q^{-1}(s, \vec{h})\ J$. We consider the system of ODEs (3.5) for $s > 0$ with initial condition

$$\vec{b}(0) = \vec{h}^*, \qquad \vec{h}^* := {}^{t}(1/\sqrt{p_1}, \ldots, 1/\sqrt{p_n}). \tag{3.6}$$

Local-in-time unique existence of a positive solution $\vec{b}(s)$ of (3.5) and (3.6) is guaranteed, since Q is smooth and det $Q \neq 0$ near $(0, \vec{h}^*)$, so that Q^{-1} is smooth near $(0, \vec{h}^*)$.

As we shall prove in Lemma 3.4, the local solution $\vec{b}(s)$ can be extended uniquely up to $s = 1 + \tau$ with some $\tau > 0$ (obtained in Theorem 4.1). Then $a_j := b_j(1)$ ($j = 1, \ldots, n$) is a solution of (3.2), so that $\{L_j(t)\}_{j=1}^n$, with $L_j(t) = t^{1/2}/b_j(1)$ for $t > 0,\ j = 1, \ldots, n$, is a self-similar solution of (3.1), which implies Proposition 3.2.

To prove the unique solvability up to $s = 1$ we need to prepare a priori estimate. We use the notation $\vec{x} < \vec{y}$ (resp., $\vec{x} \leqq \vec{y}$ for $\vec{x} = {}^{t}(x_1, \ldots, x_n), \vec{y} = {}^{t}(y_1, \ldots, y_n) \in \mathbf{R}^n$ if $x_j < y_j$ (resp., $x_j \leq y_j$) for $j = 1, \ldots, n$.

Lemma 3.3 (a priori estimate). *Let $S_0 > 0$ denote the maximal existence time of a positive solution of the system of ODEs* (3.5) *and* (3.6). *Set $S_1 := \min(S_0, 1 + \tau)$, where τ is a positive number obtained in Theorem* 4.1. *Let $\vec{b}(s)$ be the solution of* (3.5) *and* (3.6).

(I) *The derivative of each component of $\vec{b}(s)$ is positive, i.e., $\vec{b}'(s) > \vec{0}$ for $s \in [0, S_1)$.*

(II) *In particular,* $\vec{b}(s) > \vec{h}^*(> \vec{0})$ *for* $s \in (0, S_1)$.
(III) *There exists a constant* $C_3 > 0$ *independent of* s *such that*

$$0 \leq \{\text{each element of } Q^{-1}(s, \vec{b}(s))\} \leq C_3 \quad \textit{for } s \in [0, S_1).$$

Proof. The main steps of this lemma are proved in the next section, as summarized in Theorem 4.1. Since each element of the matrix J is nonnegative and $\vec{b}(s) > 0$ for $s \in [0, S_1)$, we have $J\vec{b}(s) > \vec{0}$. We now observe that (II) and (III) of Theorem 4.1 imply $\vec{b}'(s) = G(s, \vec{b}(s))\vec{b}(s) > 0$ for $s \in [0, S_1)$. Initial condition (3.6) and (I) yield (II). Theorem 4.1 and (II) implies (III). □

Lemma 3.4 (unique solvability up to $s = 1$). *There exists the unique positive solution* $\vec{b}(s)$ *(i.e.,* $\vec{b}(s) > \vec{0}$*) of the system of ODEs* (3.5) *and* (3.6) *for* $s \in [0, 1 + \tau]$, *where* τ *is a positive number obtained in Theorem* 4.1.

Proof. Let S_0 be the maximal existence time of (3.5) and (3.6), and set $S_1 := \min(S_0, 1 + \tau)$. We note that $G(s, \vec{b}(s))$ is well defined for $s \in [0, S_1)$ by Lemma 3.3. Integrating (3.5), we have

$$\vec{b}(s) - \vec{b}(0) = \int_0^s G(u, \vec{b}(u))\vec{b}(u)du \quad \text{for} \quad s \in [0, S_1),$$

which implies

$$|\vec{b}(s)| \;\leqq\; |\vec{b}(0)| + \int_0^s |G(u, \vec{b}(u))|_{\text{op}}\; |\vec{b}(u)|du \quad \text{for} \quad s \in [0, S_1).$$

Here $|\cdot|$ denotes the Euclidean norm and $|\cdot|_{\text{op}}$ denotes the operator norm from $\mathbf{R}^n$ to $\mathbf{R}^n$. Lemma 3.3 implies that there exist a constant C_1 independent of s such that

$$0 < \; |G(s, \vec{b}(s))|_{\text{op}} \; \leq C_1 \quad \text{for} \quad s \in [0, S_1),$$

since each element of J is nonnegative. So we have

$$|\vec{b}(s)| \;\leqq\; |\vec{b}(0)| + C_1 \int_0^s |\vec{b}(u)|du \quad \text{for} \quad s \in [0, S_1).$$

Gronwall's lemma implies

$$|\vec{b}(s)| \;\leqq\; |\vec{b}(0)| \exp(C_1\; s) \;\leqq\; |\vec{h}^*| \exp(C_1\; S_1) =: C_2 \quad \text{for} \quad s \in [0, S_1).$$

Suppose that $S_0 \leqq 1 + \tau$. Then Lemma 3.3(II) yields

$$1/\sqrt{p_j} \;\leq b_j(s) \;\leq C_2 \quad \text{for} \quad s \in [0, S_0),\; j = 1, \ldots, n. \tag{3.7}$$

Since S_0 is the maximal existence time, by a standard extension theorem for ODEs (e.g., [14, Chapter II, Theorem 3.1]), we have either $\lim_{s \to S_0} b_j(s) = \infty$ or $\lim_{s \to S_0} b_j(s) = 0$ for some $j = 1, \ldots, n$. This evidently contradicts (3.7). Thus we have $S_0 > 1 + \tau$. □

4. A priori estimates for matrices.

Theorem 4.1 (a priori estimates). (I) *There exist some positive constants* C_4 *and* τ *(independent of* s *and* $\vec{h}$*) such that*

$$\det Q(s, \vec{h}) > C_4 \prod_{j=1}^{n} p_j (> 0) \quad \textit{ for all } s \in [0, 1 + \tau] \textit{ and all } \vec{h} \in (\mathbf{R}_+)^n.$$

(II) *The matrix* $Q(s,\vec{h})$ *has its inverse for* $s \in [0,1+\tau]$ *and* $\vec{h} \in (\mathbf{R}_+)^n$. *Each element of* $Q^{-1}(s,\vec{h})$ *is smooth on* $[0,1+\tau] \times (\mathbf{R}_+)^n$.

(III) *Let* $\vec{h}^\sharp \in (\mathbf{R}_+)^n$ *with* $\vec{h}^\sharp > 0$. *There exists a constant* $C_5 > 0$ *(independent of* s *and* $\vec{h}$*) such that* $0 \leqq \{$*each element of* $Q^{-1}(s,\vec{h})\} \leqq C_5$ *for all* $s \in [0,1+\tau]$ *and all* $\vec{h} \in (\mathbf{R}_+)^n$ *with* $\vec{h} \geqq \vec{h}^\sharp$.

Here we use the notation $\mathbf{R}_+ := (0,\infty)$, so that $(\mathbf{R}_+)^n = \overbrace{(0,\infty)\times\cdots\times(0,\infty)}^{n}$. To prove the theorem we shall show the following lemmas. To show positiveness of the determinant of the matrix Q we consider matrix $M^{k\ell}(s)$:

$$M^{k\ell}(s) := \begin{pmatrix} 1 & sr_{k+1}/p_{k+1} & & & & & \\ sq_k/p_k & 1 & sr_{k+2}/p_{k+2} & & & \mathbf{0} & \\ & sq_{k+1}/p_{k+1} & 1 & sr_{k+3}/p_{k+3} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & \mathbf{0} & & & sq_{\ell-2}/p_{\ell-2} & 1 & sr_\ell/p_\ell \\ & & & & & sq_{\ell-1}/p_{\ell-1} & 1 \end{pmatrix}$$

for $s \geqq 0$ and $k,\ \ell = 1,\ldots,n$ with $k \leqq \ell$. In particular,

$$M^{kk}(s) = (1) \qquad \text{for}\quad s \geqq 0,\ k = 1,\ldots,n,$$

$$M^{kk+1}(s) = \begin{pmatrix} 1 & sr_{k+1}/p_{k+1} \\ sq_k/p_k & 1 \end{pmatrix} \qquad \text{for}\quad s \geqq 0,\ k = 1,\ldots,n-1.$$

We note that $M^{1n}(s)$ is obtained by dividing each jth column of $K(s)$ by p_j. We set $\tilde{M}^{k\ell} = M^{k\ell}(1)$. Fortunately, $\det\tilde{M}^{k\ell}$ is computable. Note that $\det H = \prod_{j=1}^n p_j \cdot \det\tilde{M}^{1n}$.

LEMMA 4.2. *For* $k,\ell = 1,\ldots,n$ *with* $k \leqq \ell$,

$$\det\tilde{M}^{k\ell} = \frac{\sin\left(\sum_{j=k}^{\ell+1}\varphi_j\right)}{\left(\prod_{j=k}^{\ell}\nu_j\right)\left(\prod_{j=k}^{\ell+1}\sin\varphi_j\right)}, \tag{4.1}$$

where

$$\nu_j := \cot\varphi_{j+1} + \cot\varphi_j \left(= \frac{\sin(\varphi_j+\varphi_{j+1})}{\sin\varphi_j\sin\varphi_{j+1}}\right). \tag{4.2}$$

Proof. The quantities q_j/p_j and r_j/p_j appearing in the matrix $M^{k\ell}(s)$ can be calculated as follows:

$$\frac{q_j}{p_j} = \frac{-1}{\nu_j\sin\varphi_{j+1}}\ ,\quad \frac{r_j}{p_j} = \frac{-1}{\nu_j\sin\varphi_j}.$$

We may assume that $k = 1$ without loss of generality. We shall prove (4.1) by induction. Let m^ℓ denote the right-hand side of (4.1) with $k = 1$.

(i) Using equality (4.2), we have

$$m^1 = \frac{\sin(\varphi_1 + \varphi_2)}{(\cot\varphi_2 + \cot\varphi_1)\sin\varphi_1\sin\varphi_2} = 1 = \det \tilde{M}^{11},$$

which implies (4.1) with $k = 1 = \ell$.

(ii) Next we shall show (4.1) with $k = 1,\ \ell = 2$. We have

$$\det \tilde{M}^{12} = 1 - \frac{-1}{\nu_1 \sin\varphi_2} \cdot \frac{-1}{\nu_2\sin\varphi_2} = \frac{d_2}{\nu_1\nu_2\sin\varphi_1\sin^2\varphi_2\sin\varphi_3} \tag{4.3}$$

with $d_2 = (\nu_1\nu_2\sin^2\varphi_2 - 1)\sin\varphi_1\sin\varphi_3$. Equality (4.2) yields $d_2 = \sin(\varphi_1 + \varphi_2)\sin(\varphi_2+\varphi_3) - \sin\varphi_1\sin\varphi_3$. Using the identity

$$\sin\alpha\sin\beta = -\{\cos(\alpha+\beta) - \cos(\alpha-\beta)\}/2, \tag{4.4}$$

we have

$$d_2 = -\{\cos(\varphi_1 + 2\varphi_2 + \varphi_3) - \cos(\varphi_1+\varphi_3)\}/2.$$

By the identity

$$\cos\alpha - \cos\beta = -2\sin\frac{\alpha+\beta}{2}\sin\frac{\alpha-\beta}{2}, \tag{4.5}$$

we have $d_2 = \sin(\varphi_1+\varphi_2+\varphi_3)\sin\varphi_2$. Substituting this into (4.3), we obtain

$$\det \tilde{M}^{12} = \frac{\sin(\varphi_1+\varphi_2+\varphi_3)}{\nu_1\nu_2\sin\varphi_1\sin\varphi_2\sin\varphi_3} = m^2,$$

which is (4.1) for $k = 1,\ \ell = 2$.

(iii) We assume that $\ell \geqq 3$ and (4.1) holds for $\tilde{M}^{11}, \tilde{M}^{12}, \ldots, \tilde{M}^{1\ \ell-1}$. We have

$$\det \tilde{M}^{1\ell} = \det \tilde{M}^{1\ \ell-1} - \frac{r_\ell}{p_\ell}\frac{q_{\ell-1}}{p_{\ell-1}} \det \tilde{M}^{1\ \ell-2},$$

since

$$\tilde{M}^{1i} = \left(\begin{array}{ccc|c} & & & 0 \\ & \tilde{M}^{1\ i-1} & & \vdots \\ & & & 0 \\ & & & r_i/p_i \\ \hline 0 & \cdots\ 0 & q_{i-1}/p_{i-1} & 1 \end{array}\right) \quad \text{for } i = 2, \ldots, \ell.$$

The assumption of the induction yields

$$\det \tilde{M}^{1\ell} = \frac{\sin\left(\sum_{j=1}^{\ell}\varphi_j\right)}{\left(\prod_{j=1}^{\ell-1}\nu_j\right)\left(\prod_{j=1}^{\ell}\sin\varphi_j\right)} - \frac{-1}{\nu_\ell \sin\varphi_\ell}\cdot\frac{-1}{\nu_{\ell-1}\sin\varphi_\ell}\cdot\frac{\sin\left(\sum_{j=1}^{\ell-1}\varphi_j\right)}{\left(\prod_{j=1}^{\ell-2}\nu_j\right)\left(\prod_{j=1}^{\ell-1}\sin\varphi_j\right)}.$$

An elementary calculation yields

$$\det \tilde{M}^{1\ell} = \frac{d_\ell}{\left(\prod_{j=1}^{\ell}\nu_j\right)\left(\prod_{j=1}^{\ell+1}\sin\varphi_j\right)\sin\varphi_\ell} \tag{4.6}$$

with

$$d_\ell = \sin\left(\sum_{j=1}^{\ell}\varphi_j\right)\sin(\varphi_\ell+\varphi_{\ell+1}) - \sin\left(\sum_{j=1}^{\ell-1}\varphi_j\right)\sin\varphi_{\ell+1}.$$

The identity (4.4) yields

$$d_\ell = \frac{-1}{2}\left\{\cos\left(\sum_{j=1}^{\ell+1}\varphi_j+\varphi_\ell\right) - \cos\left(\sum_{j=1}^{\ell-1}\varphi_j+\varphi_{\ell+1}\right)\right\},$$

and the identity (4.5) yields

$$d_\ell = \sin\left(\sum_{j=1}^{\ell+1}\varphi_j\right)\sin\varphi_\ell.$$

Substituting this into (4.6), we obtain (4.1) for $\det \tilde{M}^{1\ell}$. By induction the proof is now complete. $\square$

LEMMA 4.3. *The identities*

$$\begin{aligned}\frac{d}{ds}\left\{\det M^{k\ell}(s)\right\} = -s\Bigg\{&\sum_{j=k+1}^{\ell-2}\frac{q_j}{p_j}\frac{r_{j+1}}{p_{j+1}}\det M^{k\ j-1}(s)\cdot\det M^{j+2\ \ell}(s)\\ &+\sum_{j=k+2}^{\ell-1}\frac{q_{j-1}}{p_{j-1}}\frac{r_j}{p_j}\det M^{k\ j-2}(s)\cdot\det M^{j+1\ \ell}(s)\\ &+2\frac{q_k}{p_k}\frac{r_{k+1}}{p_{k+1}}\det M^{k+2\ \ell}(s)+2\frac{q_{\ell-1}}{p_{\ell-1}}\frac{r_\ell}{p_\ell}\det M^{k\ \ell-2}(s)\Bigg\}\end{aligned} \tag{4.7}$$

for $s \geqq 0$ and $k,\ell = 1,\dots,n$ with $k+2<\ell$ hold. Moreover,

$$\begin{aligned}&\frac{d}{ds}\left\{\det\ M^{kk}(s)\right\} = 0 && \text{for } s \geqq 0 \text{ and } k=1,\dots,n;\\ &\frac{d}{ds}\left\{\det M^{k\ k+1}(s)\right\} = -2s\frac{q_k}{p_k}\frac{r_{k+1}}{p_{k+1}} && \text{for } s \geqq 0 \text{ and } k=1,\dots,n-1;\\ &\frac{d}{ds}\left\{\det M^{k\ k+2}(s)\right\} = -2s\left(\frac{q_k}{p_k}\frac{r_{k+1}}{p_{k+1}}+\frac{q_{k+1}}{p_{k+1}}\frac{r_{k+2}}{p_{k+2}}\right) && \text{for } s \geqq 0 \text{ and } k=1,\dots,n-2.\end{aligned}$$

Proof. We observe that for $s \geqq 0$ and $k, \ell = 1, \ldots, n$ with $k+2<\ell$,

$$
\begin{aligned}
&\frac{d}{ds}\{\det M^{k\ell}(s)\} \\
&= \sum_{j=k+1}^{\ell-1} \det \left(\begin{array}{c|c|c}
 & 0 & \\
 & \vdots & \\
M^{k\ j-1}(s) & 0 & 0 \\
 & r_j/p_j & \\
\hline
0 \ \cdots \ 0 \ \ sq_{j-1}/p_{j-1} & 0 & sr_{j+1}/p_{j+1} \ \ 0 \ \cdots \ 0 \\
\hline
 & q_j/p_j & \\
0 & 0 & M^{j+1\ \ell}(s) \\
 & \vdots & \\
 & 0 &
\end{array}\right) \\
&\quad + \det \left(\begin{array}{c|c}
0 & sr_{k+1}/p_{k+1} \ \ 0 \ \cdots \ 0 \\
\hline
q_k/p_k & \\
0 & M^{k+1\ \ell}(s) \\
\vdots & \\
0 &
\end{array}\right)
+ \det \left(\begin{array}{c|c}
 & 0 \\
 & \vdots \\
M^{k\ \ell-1}(s) & 0 \\
 & r_\ell/p_\ell \\
\hline
0 \ \cdots \ 0 \ \ sq_{\ell-1}/p_{\ell-1} & 0
\end{array}\right) \\
&= -\sum_{j=k+1}^{\ell-2} \frac{q_j}{p_j} \det \left(\begin{array}{c|c|c}
 & 0 & \\
M^{k\ j-1}(s) & \vdots & 0 \\
 & 0 & \\
\hline
0 \ \cdots \ 0 \ \ sq_{j-1}/p_{j-1} & sr_{j+1}/p_{j+1} & 0 \ \cdots \ 0 \\
\hline
 & sq_{j+1}/p_{j+1} & \\
0 & 0 & M^{j+2\ \ell}(s) \\
 & \vdots & \\
 & 0 &
\end{array}\right) \\
&\quad - \frac{q_{\ell-1}}{p_{\ell-1}} \det \left(\begin{array}{c|c}
 & 0 \\
M^{k\ \ell-2}(s) & \vdots \\
 & 0 \\
\hline
0 \ \cdots \ 0 \ \ sq_{\ell-2}/p_{\ell-2} & sr_\ell/p_\ell
\end{array}\right) \\
&\quad - \sum_{j=k+2}^{\ell-1} \frac{r_j}{p_j} \det \left(\begin{array}{c|c|c}
 & 0 & \\
 & \vdots & \\
M^{k\ j-2}(s) & 0 & 0 \\
 & sr_{j-1}/p_{j-1} & \\
\hline
0 \ \cdots \ 0 & sq_{j-1}/p_{j-1} & sr_{j+1}/p_{j+1} \ \ 0 \ \cdots \ 0 \\
\hline
 & 0 & \\
0 & \vdots & M^{j+1\ \ell}(s) \\
 & 0 &
\end{array}\right) \\
&\quad - \frac{r_{k+1}}{p_{k+1}} \det \left(\begin{array}{c|c}
sq_k/p_k & sr_{k+2}/p_{k+2} \ \ 0 \cdots 0 \\
\hline
0 & \\
\vdots & M^{k+2\ \ell}(s) \\
0 &
\end{array}\right) \\
&\quad - s\frac{q_k}{p_k}\ \frac{r_{k+1}}{p_{k+1}} \det M^{k+2\ \ell}(s) - s\frac{r_\ell}{p_\ell}\ \frac{q_{\ell-1}}{p_{\ell-1}} \det M^{k\ \ell-2}(s),
\end{aligned}
$$

which yields (4.7). A direct calculation yields the desired formulas for $k+2 \geq \ell$. ☐

LEMMA 4.4.

(I) *For* $k,\ell = 1,\dots,n$ *with* $k \leqq \ell$, $\det M^{k\ell}(1) = \det \tilde{M}^{k\ell} > 0$.

(II) *There exists* $\tau > 0$ *such that*

$$\frac{d}{ds}\{\det M^{k\ell}(s)\} \leqq 0 \quad \textit{and} \quad \det M^{k\ell}(s) \geqq \det M^{k\ell}(1+\tau) > 0$$

for $s \in [0, 1+\tau]$ *and* $k,\ell = 1,\dots,n$ *with* $k \leqq \ell$.

Proof. (I) For all $j = 1,\dots,n+1$ we see that $\varphi_j < 0$ (resp., $\varphi_j > 0$) if the cone C is convex (resp., concave). Since the original assumption in section 2 yields $|\sum_{j=1}^{n+1} \varphi_j| = |\theta_A - \theta_B| < \pi$, identity (4.1) in Lemma 4.2 implies (I).

(II) First we note that $p_j > 0$ $(j = 1,\dots,n), q_j < 0$ $(j = 1,\dots,n-1)$, and $r_j < 0$ $(j = 2,\dots,n)$. We shall prove by induction.

(i) For $k = 1,\dots,n$, $\det M^{kk}(s) = 1$, so that $(d/ds)\{\det M^{kk}(s)\} = 0$ for $s \in \mathbf{R}$.

(ii) For $k = 1,\dots,n-1$, $(d/ds)\{\det M^{k\ k+1}(s)\} \leqq 0$ for $s \geqq 0$ by Lemma 4.3. By (I) with $\ell = k+1$, there exists $\tau^{k\,k+1} > 0$ such that $\det M^{k\,k+1}(s) > 0$ for $s \in [0, 1+\tau^{k\,k+1}]$.

(iii) For $k = 1,\dots,n-2, (d/ds)\{\det M^{k\ k+2}(s)\} \leq 0$ for $s \geq 0$ by Lemma 4.3. By (I) with $\ell = k+1$, there exists $\tau^{k\,k+2} > 0$ such that det $M^{k\,k+2}(s) > 0$ for $s \in [0, 1+\tau^{k\,k+2}]$.

(iv) Next we consider when $n \geqq 4$. We argue by induction on $\ell - k$. By (i)–(iii) we know (II) holds for $\ell - k \leqq 2$. Let m be $3,4,\dots,n-1$. Assume that (II) holds for $\ell - k \leqq m-1$. We shall prove (II) for $\ell - k = m$. By assumptions of the induction for $k = 1,2,\dots,\ n-(m-1)$, $i = 0,1,\dots,m-1$, there exists $\tau^{k\,k+i} > 0$ such that $\det M^{k\,k+i}(s) > 0$ for $s \in [0, 1+\tau^{k\,k+i}]$. For $k = 1,2,\dots,n-m$, identity (4.7) in Lemma 4.3 with $\ell = k+m$ yields $(d/ds)\{\det M^{k\,k+m}(s)\} \leqq 0$ for $s \in [0, 1+\hat{\tau}^{k\,k+m}]$ with $\hat{\tau}^{k\,k+m} = \min\{\min\{\tau^{kj}; j = k+1, k+2,\dots,k+m-2\}, \min\{\tau^{j\,k+m}; j = k+2, k+3,\dots,k+m-1\}\}$. By (I) with $\ell = k+m$, there exists $\tau^{k\,k+m} > 0$ such that $\det M^{k\,k+m}(s) > 0$ for $s \in [0, 1+\tau^{k\,k+m}]$. ☐

Next we shall show an inequality, which is used in the proof of Theorem 4.1. For $\mu_j > 0$ $(j = 1,\dots,n)$, $\xi_j \in \mathbf{R}$ $(j = 1,\dots,n-1)$, and $\eta_j \in \mathbf{R}$ $(j = 2,\dots,n)$, we set

$$A^{k\ell} := B^{k\ell} + \operatorname{diag}(\mu_k, \mu_{k+1},\dots,\mu_\ell),$$

$$B^{k\ell} := \begin{pmatrix} 1 & \eta_{k+1} & & & & & \mathbf{0} \\ \xi_k & 1 & \eta_{k+2} & & & & \\ & \xi_{k+1} & 1 & \eta_{k+3} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & \xi_{\ell-3} & 1 & \eta_{\ell-1} & \\ & \mathbf{0} & & & \xi_{\ell-2} & 1 & \eta_\ell \\ & & & & & \xi_{\ell-1} & 1 \end{pmatrix}$$

for $k,\ell = 1,\dots,n$ with $k \leqq \ell$. In particular,

$$B^{kk} := (1) \qquad \text{for } k = 1,\dots,n;$$

$$B^{k\ k+1} := \begin{pmatrix} 1 & \eta_{k+1} \\ \xi_k & 1 \end{pmatrix} \qquad \text{for } k = 1,\dots,n-1 \text{ when } n \geqq 2.$$

LEMMA 4.5. *Let* k *and* ℓ *be* $1,\dots,n$ *with* $k \leqq \ell$. *If* $\det B^{pq} > 0$ *for* $p,q = k,k+1,\dots,\ell$ *with* $p \leqq q$ *then* $\det A^{k\ell} > \det B^{k\ell}$.

Proof. For $k,\ \ell = 1, \ldots, n$ with $k \leq \ell$ we set

$$C^{k\ell}(r) := B^{k\ell} + r\ \mathrm{diag}(\mu_k, \ldots, \mu_\ell) \qquad \text{for } r > 0,$$

which yields $C^{k\ell}(0) = B^{k\ell}$ and $C^{k\ell}(1) = A^{k\ell}$. We shall prove by induction on $\ell - k = 0, 1, 2, \ldots, n-1$. If $0 \leq \ell - k \leq 1$, direct calculation yields $(d/dr)\{\det C^{k\ell}(r)\} > 0$ for $r > 0$, which implies $\det C^{k\ell}(r) > \det B^{k\ell} > 0$ for $r > 0$. We assume that $k, \ell = 1, \ldots, n$ with $k + 2 \leq \ell$. Suppose that $(d/dr)\{\det C^{pq}(r)\} > 0$ for $r > 0$ and for $p, q = k, k+1, \ldots, \ell$ with $p \leq q \leq p + \ell - k - 1$. An elementary calculation yields

$$\begin{aligned}\frac{d}{dr}\det C^{k\ell}(r) &= \mu_k \cdot \det C^{k+1\ \ell}(r) + \det C^{k\ \ell-1}(r) \cdot \mu_\ell \\ &\quad + \sum_{j=k+1}^{\ell-1} \det C^{k\ j-1}(r) \cdot \mu_j \cdot \det C^{j+1\ \ell}(r) \\ &> 0 \qquad\qquad \text{for} \quad r > 0,\end{aligned}$$

since $\det C^{pq}(r) > \det C^{pq}(0) = \det B^{pq} > 0$ for $r > 0$ and for $p, q = k, k+1, \ldots, \ell$ with $p \leq q \leq p + \ell - k - 1$ by the assumption of the induction. Thus we obtain

$$\det A^{k\ell} = \det C^{k\ell}(1) > \det C^{k\ell}(0) = \det B^{k\ell}$$

for $k \leq \ell$. $\square$

Finally, we shall show the identities on a cofactor of tridiagonal matrix, which are used in the proof of Theorem 4.1. For $\lambda_j \in \mathbf{R}$ $(j = 1, \ldots, n)$, $\xi_j \in \mathbf{R}$ $(j = 2, \ldots, n)$, and $\eta_j \in \mathbf{R}$ $(j = 1, \ldots, n-1)$, we set the matrix

$$E^{k\ell} := \begin{pmatrix} \lambda_k & \xi_{k+1} & & & & & & \\ \eta_k & \lambda_{k+1} & \xi_{k+2} & & & \mathbf{0} & & \\ & \eta_{k+1} & \lambda_{k+2} & \xi_{k+3} & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \ddots & \ddots & \ddots & & \\ & & & & \eta_{\ell-3} & \lambda_{\ell-2} & \xi_{\ell-1} & \\ & \mathbf{0} & & & & \eta_{\ell-2} & \lambda_{\ell-1} & \xi_\ell \\ & & & & & & \eta_{\ell-1} & \lambda_\ell \end{pmatrix}$$

for $k, \ell = 1, \ldots, n$ with $k \leqq \ell$. In particular,

$$\begin{aligned} E^{kk} &:= (\lambda_k) && \text{for } k = 1, \ldots, n; \\ E^{kk+1} &:= \begin{pmatrix} \lambda_k & \eta_{k+1} \\ \eta_k & \lambda_{k+1} \end{pmatrix} && \text{for } k = 1, \ldots, n-1 \text{ when } n \geqq 2. \end{aligned}$$

Let D_{pq} be the (p, q) cofactor of the matrix E^{1n} for $p, q = 1, \ldots, n$. The next lemma is obtained by an elementary calculation.

LEMMA 4.6. *For $p, q = 1, \ldots, n$ the following identities hold.*

(I) *When $p < q$,*

$$D_{pq} = (-1)^{p+q} \prod_{j=p}^{q-1} \eta_j \det E^{1\ p-1} \det E^{q+1\ n}.$$

(II) *When* $p > q$,

$$D_{pq} = (-1)^{p+q} \prod_{j=q+1}^{p} \xi_j \det E^{1\ q-1} \det E^{p+1\ n}.$$

(III)

$$D_{pp} = \det E^{1\ p-1} \det E^{p+1\ n}.$$

Here we use the convention $\det E^{10} = 1 = \det E^{n+1\ n}$.

Proof of Theorem 4.1. (I) We set

$$Q^{k\ell}\left(s, \vec{h}^{k\ell}\right) := K^{k\ell}(s) + \operatorname{diag}\left(1/h_k^2, 1/h_{k+1}^2, \ldots, 1/h_\ell^2\right)$$

and

$$K^{k\ell}(s) := \Big(K_{ij}(s)\Big)_{i,j=k,k+1,\ldots,\ell},$$

where $(K_{ij}(s))_{i,j=1,\ldots,n}$ is the matrix $K(s)$ defined in section 3. Note that $Q^{1\,n}$ equals Q defined in section 3. Setting

$$W^{k\ell}\left(s, \vec{h}^{k\ell}\right) := M^{k\ell}(s) + \operatorname{diag}\left(\frac{1}{p_k h_k^2}, \frac{1}{p_{k+1} h_{k+1}^2}, \ldots, \frac{1}{p_\ell h_\ell^2}\right)$$

for $s \geqq 0, \vec{h}^{k\ell} = {}^{\mathrm{t}}(h_k, h_{k+1}, \ldots, h_\ell) \in (\mathbf{R}_+)^{\ell-k+1}$, and $k, \ell = 1, \ldots, n$ with $k \leqq \ell$, we have

$$\det Q^{k\ell}\left(s, \vec{h}^{k\ell}\right) = \det W^{k\ell}\left(s, \vec{h}^{k\ell}\right) \cdot \prod_{j=k}^{\ell} p_j.$$

Lemma 4.4(II) yields that there exists $\tau > 0$ such that $\det M^{pq}(s) \geqq C_{pq}$ for $s \in [0, 1+\tau]$ and $p, q = 1, \ldots, n$ with $p \leqq q$, where $C_{pq} := \det M^{pq}(1+\tau) > 0$. By Lemma 4.5 we have $\det W^{k\ell}(s, \vec{h}^{k\ell}) > \det M^{k\ell}(s)$ for $s \in [0, 1+\tau]$ and $\vec{h}^{k\ell} \in (\mathbf{R}_+)^{\ell-k+1}$. Since $p_j > 0$ $(j = 1, \ldots, n)$,

$$\det Q^{k\ell}\left(s, \vec{h}^{k\ell}\right) > C_{k\ell} \prod_{j=k}^{\ell} p_j \quad (> 0) \tag{4.8}$$

for $s \in [0, 1+\tau]$ and $\vec{h}^{k\ell} \in (\mathbf{R}_+)^{\ell-k+1}$.

(II) Part (II) follows from (I) with $k = 1$ and $\ell = n$ and the definition of Q.

(III) When $n = 1, Q(s, \vec{h})$ is a scalar and equals $\left(p_1 + 1/h_1^2\right)$, so that

$$0 < Q^{-1}(s, \vec{h}) = \left(p_1 + 1/h_1^2\right)^{-1} < 1/p_1.$$

We may assume that $n \geqq 2$. Since $Q(s, \vec{h})$ is invertible for $s \in [0, 1+\tau]$ and $\vec{h} \in (\mathbf{R}_+)^n$ by (I), the (p, q) element of the inverse matrix of $Q(s, \vec{h})$ equals $\Delta_{qp}(s, \vec{h}) / \det Q(s, \vec{h})$ for $p, q = 1, \ldots, n$, where $\Delta_{pq}(s, \vec{h})$ denotes the (p, q) cofactor of $Q(s, \vec{h})$. By Lemma

4.6 we have

$$\Delta_{pq}(s,\vec{h}) = (-1)^{p+q}\left(\prod_{j=p}^{q-1} q_j\right) s^{q-p} \det Q^{1\ p-1}(s,\vec{h}) \det Q^{q+1\ n}(s,\vec{h}) \quad \text{for} \quad p<q;$$

$$\Delta_{pq}(s,\vec{h}) = (-1)^{p+q}\left(\prod_{j=q+1}^{p} r_j\right) s^{p-q} \det Q^{1\ q-1}(s,\vec{h}) \det Q^{p+1\ n}(s,\vec{h}) \quad \text{for} \quad p>q;$$

$$\Delta_{pp}(s,\vec{h}) = \det Q^{1\ p-1}(s,\vec{h}) \det Q^{p+1\ n}(s,\vec{h})$$

for $s \geqq 0$ and $\vec{h} \in (\mathbf{R}_+)^n$. Here we use the convention

$$\det Q^{10}(s,\vec{h}) = 1 = \det Q^{n+1\ n}(s,\vec{h}).$$

Since $q_j < 0$ $(j = 1,\ldots,n-1)$ and $r_j < 0$ $(j = 2,\ldots,n)$, inequality (4.8) yields $\Delta_{pq}(s,\vec{h}) \geqq 0$ for $s \in [0, 1+\tau]$, $\vec{h} \in (\mathbf{R}_+)^n$ and $p,q = 1,\ldots,n$, which implies $\{(p,q)$ element of $Q^{-1}(s,\vec{h})\} = \Delta_{qp}(s,\vec{h})/\det Q(s,\vec{h}) \geqq 0$ by (I). On the other hand, $\Delta_{pq}(s,\vec{h})$ is bounded from the above for $s \in [0,1+\tau]$ and $\vec{h} = (h_1,\ldots,h_n) \in (\mathbf{R}_+)^n$ with $\vec{h} \geq \vec{h}^\sharp$, since $\Delta_{pq}(s,\vec{h})$ is a polynomial of $1/h_1^2,\ldots,1/h_n^2$ and s. Thus (I) yields that there exists $C_5 > 0$ such that $\Delta_{pq}(s,\vec{h})/\det Q(s,\vec{h}) \leqq C_5$ for $s \in [0,1+\tau], \vec{h} \in (\mathbf{R}_+)^n$ with $\vec{h} \geq \vec{h}^\sharp$ and $p,q = 1,\ldots,n$. $\square$

Acknowledgment. The authors are grateful to the anonymous referees for their careful reading and valuable comments on the first version of this paper.

REFERENCES

[1] S. B. Angenent and M. E. Gurtin, *Multiphase thermomechanics with interfacial structure* 2. *Evolution of an isothermal interface*, Arch. Ration. Mech. Anal., 108 (1989), pp. 323–391.

[2] D. Campbell, *A First Glance at Crystal Motion*, Master's thesis, Rutgers University, New Brunswick, NJ, 2002.

[3] Y.-L. Chang, J.-S. Guo, and Y. Kohsaka, *On a two-point free boundary problem for a quasilinear parabolic equation*, Asymptot. Anal., 34 (2003), pp. 333–358.

[4] C. M. Elliott, A. R. Gardiner, and R. Schätzle, *Crystalline curvature flow of a graph in a variational setting*, Adv. Math. Sci. Appl., 8 (1998), pp. 425–460.

[5] T. Fukui and Y. Giga, *Motion of a graph by nonsmooth weighted curvature*, in World Congress of Nonlinear Analysts '92, Vol. I, V. Lakshmikantham, ed., Walter de Gruyter, Berlin, I 1996, pp. 47–56.

[6] M.-H. Giga and Y. Giga, *Evolving graphs by singular weighted curvature*, Arch. Ration. Mech. Anal., 141 (1998), pp. 117–198.

[7] M.-H. Giga and Y. Giga, *Stability for evolving graphs by nonlocal weighted curvature*, Comm. Partial Differential Equations, 24 (1999), pp. 109–184.

[8] M.-H. Giga and Y. Giga, *Crystalline and level-set flow—Convergence of a crystalline algorithm for a general anisotropic curvature flow in the plane*, in Free boundary problems: Theory and applications I, Gakuto Internat. Ser. Math. Sci. Appl. 13, N. Kenmochi, ed., Gakkotosho, Tokyo, 2000, pp. 64–79.

[9] M.-H. Giga and Y. Giga, *Generalized motion by nonlocal curvature in the place*, Arch. Ration. Mech. Anal., 159 (2001), pp. 295–333.

[10] Y. Giga and M. Gurtin, *A comparison theorem for crystalline evolution in the plane*, Quart. Appl. Math., 54 (1996), pp. 727–737.

[11] P. M. Girão, *Convergence of a crystalline algorithm for the motion of a simple closed convex curve by weighted curvature*, SIAM J. Numer. Anal., 32 (1995), pp. 886–899.

[12] P. M. Girão and R. V. Kohn, *Convergence of a crystalline algorithm for the heat equation in one dimension and for the motion of a graph by weighted curvature*, Numer. Math., 67 (1994), pp. 41-70.

[13] M. E. Gurtin, *Thermomechanics of Evolving Phase Boundaries in the Plane*, Oxford, Clarendon Press, New York, 1993.

[14] P. Hartman, *Ordinary Differential Equations*, 2nd ed., Birkhäuser, Boston, 1982.

[15] H. Hontani, M.-H. Giga, Y. Giga, and K. Deguchi, *A computation of a crystalline flow starting from non-admissible polygon using expanding selfsimilar solutions*, in 11th International Conference DGCI 2003, Lecture Notes in Comput. Sci. 2886, Springer, Naples, 2003, pp. 465–474.

[16] H. Hontani, M.-H. Giga, Y. Giga, and K. Deguchi, *Expanding selfsimilar solutions of a crystalline flow with applications to contour figure analysis*, Discrete Appl. Math., 147 (2005), pp. 265–285.

[17] Y. Kohsaka, *Free boundary problem for quasilinear parabolic equation with fixed angle of contact to a boundary*, Nonlinear Anal., 45 (2001), pp. 865–894.

[18] G. Sapiro, *Geometric Partial Differential Equations and Image Analysis*, Cambridge University Press, Cambridge, UK, 2001.

[19] A. Stancu, *Uniqueness of self-similar solutions for a crystalline flow*, Indiana Univ. Math. J., 45 (1996), pp. 1157–1174.

[20] A. Stancu, *Asymptotic behavior of solutions to a crystalline flow*, Hokkaido Math. J., 27 (1998), pp. 303–320.

[21] J. E. Taylor, *Constructions and conjectures in crystalline nondifferential geometry*, in Proceedings of the Conference on Differential Geometry, Pitman Monogr. Surveys, Pure Appl. Math. 52, Pitman, London, 1991, pp. 321–336.

[22] J. E. Taylor, *Mean curvature and weighted mean curvature*, Acta Metall., 40 (1992), pp. 1475–1485.

[23] J. E. Taylor, *Motion of curves by crystalline curvature, including triple junction and boundary points*, in Differential Geometry: Partial Differential Equations on Manifolds, Proc. Symp. Pure Math. 54 AMS, 1993, pp. 417–438.

SIAM J. MATH. ANAL.
Vol. 37, No. 4, pp. 1227–1255

EULERIAN CALCULUS FOR THE CONTRACTION IN THE WASSERSTEIN DISTANCE*

FELIX OTTO† AND MICHAEL WESTDICKENBERG†

Abstract. We consider the porous medium equation on a compact Riemannian manifold and give a new proof of the contraction of its semigroup in the Wasserstein distance. This proof is based on the insight that the porous medium equation does not increase the size of infinitesimal perturbations along gradient flow trajectories and on an Eulerian formulation for the Wasserstein distance using smooth curves. Our approach avoids the existence result for optimal transport maps on Riemannian manifolds.

Key words. porous medium equation, contraction, Wasserstein distance

AMS subject classifications. 58J65, 28D05, 60E15, 60G15

DOI. 10.1137/050622420

1. Introduction.

In this paper we consider the porous medium equation

$$\partial_t \rho - \Delta U(\rho) = 0 \tag{1.1}$$

and give an entirely "Eulerian" argument for the contraction of its semigroup in the Wasserstein distance. The argument is guided by the formal gradient flow structure of the porous medium equation proposed in [12].

More precisely, we choose as our state space $\mathcal{M}$ the space of probability measures $\rho(x)\,dx$, endowed with a suitable metric tensor g; see section 2.2. The metric tensor g induces a distance on $(\mathcal{M}, g)$ that coincides with the Wasserstein distance $\mathcal{W}(\rho_0, \rho_1)$. Loosely speaking, this equivalence is a consequence of the Benamou–Brenier Eulerian formulation of the optimal transportation problem defining $\mathcal{W}(\rho_0, \rho_1)$ [4]. Then the porous medium equation is the gradient flow on $(\mathcal{M}, g)$ of the functional

$$E(\rho) = \int e(\rho)\,dx, \tag{1.2}$$

where the "osmotic pressure" $U(\rho)$ is related to the energy density $e(\rho)$ via

$$U(\rho) = \rho e'(\rho) - e(\rho) \quad \text{for } \rho \geqslant 0. \tag{1.3}$$

We notice that U is (strictly) monotone if and only if e is (strictly) convex and that for strictly monotone U, (1.1) is of parabolic type. The contraction property for the porous medium semigroup then follows from the convexity of E on $(\mathcal{M}, g)$. The latter is a reformulation of McCann's displacement convexity [10].

This formal argument has been made rigorous in [12] using the fact that for any two points ρ_0, ρ_1 a shortest curve with respect to $\mathcal{W}$ exists. The existence of these shortest curves relies on Brenier's result [5] on the existence of a one-to-one optimal

*Received by the editors January 11, 2005; accepted for publication (in revised form) April 18, 2005; published electronically January 6, 2006. This work was partially supported by EU network *Hyperbolic and kinetic equations* HRPN-CT-2002-00282 and by Sonderforschungsbereich 611 *Singular phenomena and scaling in mathematical models* at Bonn University.
http://www.siam.org/journals/sima/37-4/62242.html

†Institut für Angewandte Mathematik, Universität Bonn, Wegelerstr. 10, 53115 Bonn, Germany (otto@iam.uni-bonn.de, mwest@iam.uni-bonn.de).

transport map $y = \Phi(x)$ between two measures $\rho_0(x)\,dx$ and $\rho_1(y)\,dy$. This can be seen as a Lagrangian approach. It is somewhat delicate since the optimal transport map Φ can be nonsmooth even if the densities ρ_0 and ρ_1 are smooth.

In this paper, we carry out a rigorous Eulerian approach based on the new insight

- that the porous medium equation does not increase the naturally defined action $\mathcal{A}(\rho)$ of *smooth* curves $[0,1] \ni s \mapsto \rho(s) \in \mathcal{M}$ (see Proposition 4.2);
- that the squared Wasserstein distance $\frac{1}{2}\mathcal{W}(\rho_0,\rho_1)^2$ is the infimum of $\mathcal{A}(\rho)$ over smooth curves connecting ρ_0 to ρ_1 (see Proposition 4.3).

Hence we can work in the "class of smooth objects." Alternatively, contraction estimates can also be derived in more elaborate frameworks based on metric space theory. We refer the reader to the recent publications [6, 1] for further information.

Our approach allows us to obtain the contraction property on a compact Riemannian manifold $\mathbb{M}^n$ (instead of $\mathbb{R}^n$) without additional effort. A sufficient condition is that the Ricci curvature of $\mathbb{M}^n$ be nonnegative. This is the well-known Bakry–Emery criterion for the logarithmic Sobolev inequality [3] (which can be refined using Γ_2-calculus [2]). It turns out that contractivity of the semigroup for certain nonlinear evolution equations is in fact equivalent to lower bounds for the Ricci curvature. This has been proved for the heat semigroup in [17] and more generally in [16]. Our Eulerian approach avoids the subtle existence result for optimal transport maps Φ on Riemannian manifolds by McCann [11].

2. Gradient flows. It is instructive to discuss our approach in the language of gradient flows. This heuristics will serve as a guideline for the rigorous argument.

2.1. Abstract framework. Let us quickly recall the mathematical structure required for a gradient flow. One first needs a smooth function $\mathcal{M} \ni \rho \mapsto E(\rho)$ on a differentiable manifold $\mathcal{M}$. The differential $\operatorname{diff} E$ of E is a cotangent vector field:

$$\mathcal{M} \ni \rho \mapsto \operatorname{diff} E_{|\rho} \in T_\rho\mathcal{M}^*.$$

Therefore one also needs a metric tensor g on $\mathcal{M}$, i.e., a scalar product g_ρ on $T_\rho\mathcal{M}$ in every point $\rho \in \mathcal{M}$. This scalar product allows one to identify cotangent with tangent vectors, yielding the gradient vector field $\operatorname{grad} E$. The gradient flow of E on the Riemannian manifold $(\mathcal{M}, g)$ is then given by the dynamical system

$$\frac{d\rho}{dt} = -\operatorname{grad} E_{|\rho}. \tag{2.1}$$

For subsequent use, we shall reformulate (2.1). We recall that the differential $\operatorname{diff} E$ can be inferred from differentiating E along a curve $[0,1] \ni s \mapsto \rho(s) \in \mathcal{M}$:

$$\frac{d}{ds}E\big(\rho(s)\big) = \left\langle \operatorname{diff} E_{|\rho(s)},\, \frac{d\rho}{ds}(s)\right\rangle.$$

Then the gradient $\operatorname{grad} E$ is defined by the requirement that for any tangent vector field $[0,1] \ni s \mapsto \delta\rho(s) \in T_{\rho(s)}\mathcal{M}$ along the above curve we have

$$g_{\rho(s)}\big(\operatorname{grad} E_{|\rho(s)},\, \delta\rho(s)\big) = \big\langle \operatorname{diff} E_{|\rho(s)},\, \delta\rho(s)\big\rangle.$$

Now a trajectory $[0,\infty) \ni t \mapsto \rho(t) \in \mathcal{M}$ of (2.1) is characterized by the fact that for any tangent vector field $[0,\infty) \ni t \mapsto \delta\rho(t) \in T_{\rho(t)}\mathcal{M}$ one has

$$g_{\rho(t)}\left(\frac{d\rho}{dt}(t),\, \delta\rho(t)\right) + \big\langle \operatorname{diff} E_{|\rho(t)},\, \delta\rho(t)\big\rangle = 0 \quad \forall\, t. \tag{2.2}$$

2.2. Heuristics: The porous medium equation as gradient flow. We are interested in the porous medium equation on a compact, connected Riemannian manifold $\mathbb{M}^n$ without boundary. We denote by $\cdot$ the metric tensor on $\mathbb{M}^n$ and by ∇, $\nabla\cdot$, and $\Delta = \nabla\cdot\nabla$ the gradient, divergence, and Laplacian on $\mathbb{M}^n$. Finally, dx denotes the volume form on $\mathbb{M}^n$; without loss of generality we assume $\int_{\mathbb{M}^n} 1\,dx = 1$. The porous medium equation describes the evolution of a nonnegative density $\rho(t,x)$ on $\mathbb{M}^n$. It is given by the nonlinear diffusion equation

$$\partial_t \rho - \Delta U(\rho) = 0. \tag{2.3}$$

The porous medium equation preserves the total mass, and we assume $\int_{\mathbb{M}^n} \rho\,dx = 1$ for definiteness. In view of this, our state space $\mathcal{M}$ is the space of all nonnegative functions $\rho\colon \mathbb{M}^n \to [0,\infty)$ with unit integral:

$$\int_{\mathbb{M}^n} \rho\,dx = 1. \tag{2.4}$$

We also may think of $\mathcal{M}$ as the space of probability measures $\rho(x)\,dx$ on $\mathbb{M}^n$. For convenience we will not distinguish in the following between functions and the measures they induce via the volume element dx defined on $\mathbb{M}^n$.

Following [12], we now introduce the metric tensor g on $\mathcal{M}$. Notice that in view of (2.4) we may think of infinitesimal perturbations $\delta\rho \in T_\rho\mathcal{M}$ of a state $\rho\in\mathcal{M}$ as functions $\delta\rho\colon \mathbb{M}^n \to \mathbb{R}$ with

$$\int_{\mathbb{M}^n} \delta\rho\,dx = 0. \tag{2.5}$$

For given $\rho\in\mathcal{M}$ we define the scalar product g_ρ on $T_\rho\mathcal{M}$ as

$$g_\rho\big(\delta\rho_0,\, \delta\rho_1\big) = \int_{\mathbb{M}^n} \nabla\phi_0\cdot\nabla\phi_1\,\rho\,dx, \tag{2.6}$$

where, up to additive constants, the functions $\phi_i\colon \mathbb{M}^n \to \mathbb{R}$ are defined by

$$\delta\rho_i - \nabla\cdot(\rho\nabla\phi_i) = 0. \tag{2.7}$$

Notice that (2.7) constitutes an elliptic equation with variable coefficient $\rho \geqslant 0$ for ϕ_i; (2.5) is necessary for the existence. If ρ is strictly positive and $\rho, \delta\rho_i$ are smooth, then (2.5) is also sufficient for the existence of a smooth solution ϕ_i. For later use we notice that $g_\rho(\delta\rho_0, \delta\rho_1)$ can be rewritten as

$$g_\rho\big(\delta\rho_0,\, \delta\rho_1\big) = -\int_{\mathbb{M}^n} \delta\rho_0\,\phi_1\,dx. \tag{2.8}$$

The quadratic part of the metric tensor can also be characterized variationally:

$$\tfrac12 g_\rho\big(\delta\rho,\, \delta\rho\big) = \sup_\phi \Big\{ -\int_{\mathbb{M}^n} \tfrac12|\nabla\phi|^2\rho\,dx - \int_{\mathbb{M}^n} \delta\rho\,\phi\,dx \Big\}, \tag{2.9}$$

where the sup is taken over all smooth functions $\phi\colon \mathbb{M}^n \longrightarrow \mathbb{R}$. In view of (2.7), we may think of ϕ_i as the "velocity potential" that generates the infinitesimal change $\delta\rho_i$ of the density ρ.

We now formally argue that (2.3) is indeed the gradient flow of (1.2) on $(\mathcal{M}, g)$, reproducing the argument in [12]. We are given a nonnegative function $\rho = \rho(t,x)$

satisfying (2.4); we fix a time t. Let the function $\delta\rho$ of x be given with (2.5), and let ϕ be related to $\delta\rho$ by (2.7). Then we have, on the one hand, that

$$\begin{aligned} \left\langle \operatorname{diff} E_{|\rho}, \delta\rho \right\rangle &= \int_{\mathbb{M}^n} e'(\rho)\, \delta\rho\, dx \\ &\overset{(2.7)}{=} -\int_{\mathbb{M}^n} e''(\rho)\nabla\rho\cdot\rho\nabla\phi\, dx \\ &\overset{(1.3)}{=} -\int_{\mathbb{M}^n} \nabla U(\rho)\cdot\nabla\phi\, dx \\ &= \int_{\mathbb{M}^n} \Delta U(\rho)\,\phi\, dx. \end{aligned} \tag{2.10}$$

On the other hand, we have according to (2.8)

$$g_\rho\big(\partial_t\rho,\, \delta\rho\big) = -\int_{\mathbb{M}^n} \partial_t\rho\,\phi\, dx.$$

The combination of the last two identities gives, for any $\delta\rho$ satisfying (2.5),

$$g_\rho\big(\partial_t\rho,\, \delta\rho\big) + \left\langle \operatorname{diff} E_{|\rho}, \delta\rho \right\rangle = -\int_{\mathbb{M}^n} \big(\partial_t\rho - \Delta U(\rho)\big)\,\phi\, dx.$$

In view of (2.2), this proves that indeed (2.3) is the gradient flow of (1.2) with respect to the metric tensor (2.6) defined on $\mathcal{M}$.

3. Convexity and contraction. In this section we discuss heuristically how the convexity of E on $(\mathcal{M}, g)$ implies contraction for the gradient flow.

3.1. Abstract framework. Recall that a function E on a Riemannian manifold $(\mathcal{M}, g)$ is convex if its Hessian $\operatorname{Hess} E$ is positive definite in any point $\rho \in \mathcal{M}$, i.e.,

$$g_\rho\Big(\delta\rho,\, \operatorname{Hess} E_{|\rho}\,\delta\rho\Big) \geqslant 0 \quad \forall\ \delta\rho \in T_\rho\mathcal{M} \text{ and } \rho\in\mathcal{M}.$$

In an infinite-dimensional context, it is convenient to have alternative ways of probing convexity. We mention two possibilities:

- The standard way to probe convexity is by geodesics: If $[0,1] \ni s \mapsto \rho(s) \in \mathcal{M}$ is a geodesic, i.e., any curve for which

$$\frac{D}{ds}\frac{d\rho}{ds} = 0,$$

  where $\frac{D}{ds}$ denotes the covariant derivative along $s \mapsto \rho(s)$, then we have

$$\frac{d^2}{ds^2}\, E\big(\rho(s)\big) \geqslant 0.$$

  Indeed, this follows from the chain rule

$$\begin{aligned} \frac{d^2}{ds^2}\, E\big(\rho(s)\big) &= \frac{d}{ds}\, g_\rho\left(\frac{d\rho}{ds},\, \operatorname{grad} E_{|\rho}\right) \\ &= g_\rho\left(\frac{d\rho}{ds},\, \operatorname{Hess} E_{|\rho}\,\frac{d\rho}{ds}\right) + g_\rho\left(\frac{D}{ds}\frac{d\rho}{ds},\, \operatorname{grad} E_{|\rho}\right) \\ &= g_\rho\left(\frac{d\rho}{ds},\, \operatorname{Hess} E_{|\rho}\,\frac{d\rho}{ds}\right). \end{aligned} \tag{3.1}$$

- There is another way to probe convexity of E: For any gradient flow trajectory $[0,\infty) \ni t \mapsto \rho(t) \in \mathcal{M}$, i.e., any curve for which

  $$\frac{d\rho}{dt} = -\operatorname{grad} E_{|\rho},$$

  and any infinitesimal perturbation $[0,\infty) \ni t \mapsto \delta\rho(t) \in T_{\rho(t)}\mathcal{M}$ along this curve for which by the chain rule

  $$\frac{D}{dt}\,\delta\rho = -\operatorname{Hess} E_{|\rho}\,\delta\rho, \tag{3.2}$$

  we have that the size of this perturbation does not increase over time:

  $$\frac{d}{dt}\,\tfrac{1}{2} g_\rho\big(\delta\rho,\,\delta\rho\big) \leqslant 0. \tag{3.3}$$

  Indeed, this follows from

  $$\frac{d}{dt}\,\tfrac{1}{2} g_\rho\big(\delta\rho,\,\delta\rho\big) = g_\rho\left(\delta\rho,\,\frac{D}{dt}\,\delta\rho\right) = -g_\rho\Big(\delta\rho,\,\operatorname{Hess} E_{|\rho}\,\delta\rho\Big). \tag{3.4}$$

The property (3.3) has a finite counterpart: Recall that the distance $\operatorname{dist}(\rho_0,\rho_1)$ between $\rho_0,\rho_1 \in \mathcal{M}$ induced by the metric tensor g is defined by

$$\tfrac{1}{2}\operatorname{dist}(\rho_0,\rho_1)^2 = \inf\left\{\mathcal{A}(\rho)\,\middle|\,[0,1]\ni s\mapsto \rho(s)\in\mathcal{M},\ \left\{\begin{array}{l}\rho(0,\cdot)=\rho_0\\ \rho(1,\cdot)=\rho_1\end{array}\right\}\right\}, \tag{3.5}$$

where $\mathcal{A}(\rho)$ is the natural action of a curve, i.e.,

$$\mathcal{A}(\rho) := \int_0^1 \tfrac{1}{2} g_\rho\left(\frac{d\rho}{ds},\frac{d\rho}{ds}\right) ds. \tag{3.6}$$

We now argue that (3.3) easily yields a global consequence of the convexity of E: The gradient flow of E is a contraction in dist. This means that for any two gradient flow trajectories $[0,\infty) \ni t \mapsto \rho_i(t) \in \mathcal{M}$, $i=0,1$, i.e., any curves with

$$\frac{d\rho_i}{dt} = -\operatorname{grad} E_{|\rho_i},$$

we have

$$\operatorname{dist}(\rho_0,\rho_1) \quad \text{is nonincreasing in } t.$$

Indeed, by translational invariance in time, it is enough to show that

$$\tfrac{1}{2}\operatorname{dist}\big(\rho_0(t),\rho_1(t)\big)^2 \leqslant \tfrac{1}{2}\operatorname{dist}\big(\rho_0(0),\rho_1(0)\big)^2 \quad \forall\, t \geqslant 0. \tag{3.7}$$

According to (3.5), for given $\epsilon > 0$, there exists a curve $[0,1] \ni s \mapsto \bar\rho(s) \in \mathcal{M}$ such that $\bar\rho(s=0) = \rho_0(t=0)$ and $\bar\rho(s=1) = \rho_1(t=0)$, with

$$\tfrac{1}{2}\operatorname{dist}\big(\rho_0(0),\rho_1(0)\big)^2 = \tfrac{1}{2}\operatorname{dist}\big(\bar\rho(0),\bar\rho(1)\big)^2 \geqslant \mathcal{A}(\bar\rho) - \epsilon. \tag{3.8}$$

Now for every $s \in [0,1]$ let $[0,\infty) \ni t \mapsto \rho(s,t) \in \mathcal{M}$ denote the solution of

$$\frac{d\rho(s,\cdot)}{dt} = -\operatorname{grad} E_{|\rho(s,\cdot)}, \tag{3.9}$$

with $\rho(s,0)=\bar\rho(s)$. Notice that then $\rho(0,t)=\rho_0(t)$ and $\rho(1,t)=\rho_1(t)$ so that

$$\tfrac12 \operatorname{dist}\big(\rho_0(t),\rho_1(t)\big)^2 \leqslant \mathcal{A}\big(\rho(\cdot,t)\big). \tag{3.10}$$

Taking the covariant derivative of (3.9) with respect to s yields

$$\frac{D}{\partial t}\frac{\partial\rho}{\partial s}=\frac{D}{\partial s}\frac{\partial\rho}{\partial t}=-\operatorname{Hess}E_{|\rho}\frac{\partial\rho}{\partial s}.$$

Thus we obtain from (3.3) applied to $\delta\rho=\frac{\partial\rho}{\partial s}$

$$\frac{\partial}{\partial t}\,\tfrac12 g_\rho\left(\frac{\partial\rho}{\partial s},\frac{\partial\rho}{\partial s}\right)\leqslant 0.$$

Integration over s yields

$$\frac{d}{dt}\mathcal{A}\big(\rho(\cdot,t)\big)=\int_0^1\frac{\partial}{\partial t}\,\tfrac12 g_\rho\left(\frac{\partial\rho}{\partial s},\frac{\partial\rho}{\partial s}\right)ds\leqslant 0.$$

Integration over t yields

$$\mathcal{A}\big(\rho(\cdot,t)\big)\leqslant\mathcal{A}\big(\rho(\cdot,0)\big).$$

Together with (3.10) and (3.8) we therefore end up with

$$\tfrac12 \operatorname{dist}\big(\rho_0(t),\rho_1(t)\big)^2\leqslant\tfrac12 \operatorname{dist}\big(\rho_0(0),\rho_1(0)\big)^2+\epsilon,$$

and since $\epsilon>0$ was arbitrary, (3.7) is proved.

Remark 3.1 (added in proof). It is possible to give an argument in favor of

$$\frac{\partial}{\partial t}\,\tfrac12 g_\rho\left(\frac{\partial\rho}{\partial s},\frac{\partial\rho}{\partial s}\right)=-g_\rho\left(\frac{\partial\rho}{\partial s},\operatorname{Hess}E_{|\rho}\frac{\partial\rho}{\partial s}\right) \tag{3.11}$$

that avoids using the covariant derivative altogether: Consider first a family of curves $[0,1]\ni s\mapsto\tilde\rho(s,t)\in\mathcal{M}$ for $t\in[0,\infty)$ such that $s\mapsto\tilde\rho(s,0)$ is a geodesic. Then

$$\frac{\partial}{\partial t}\,\tfrac12 g_{\tilde\rho}\left(\frac{\partial\tilde\rho}{\partial s},\frac{\partial\tilde\rho}{\partial s}\right)=\frac{\partial}{\partial s}\,g_{\tilde\rho}\left(\frac{\partial\tilde\rho}{\partial s},\frac{\partial\tilde\rho}{\partial t}\right)\quad\text{for } t=0. \tag{3.12}$$

Indeed, given any function $[0,1]\ni s\mapsto\alpha(s)\in\mathbb{R}$ with $\alpha(0)=\alpha(1)=0$ let

$$\hat\rho(s,t):=\tilde\rho\big(s,\alpha(s)t\big)\quad\forall\ s,t.$$

Since $\hat\rho(0,t)=\tilde\rho(0,0)$ and $\hat\rho(1,t)=\tilde\rho(1,0)$, the definition of geodesic yields

$$0=\frac{d}{dt}_{|t=0}\int_0^1\tfrac12 g_{\hat\rho}\left(\frac{\partial\hat\rho}{\partial s},\frac{\partial\hat\rho}{\partial s}\right)ds=\int_0^1\frac{\partial}{\partial t}_{|t=0}\tfrac12 g_{\hat\rho}\left(\frac{\partial\hat\rho}{\partial s},\frac{\partial\hat\rho}{\partial s}\right)ds. \tag{3.13}$$

On the other hand, we have $\frac{\partial\hat\rho}{\partial s}(s,t)=\frac{\partial\tilde\rho}{\partial s}(s,\alpha(s)t)+\alpha'(s)t\,\frac{\partial\tilde\rho}{\partial t}(s,\alpha(s)t)$, and therefore

$$\begin{aligned}\left[\tfrac12 g_{\hat\rho}\left(\frac{\partial\hat\rho}{\partial s},\frac{\partial\hat\rho}{\partial s}\right)\right](s,t)&=\left[\tfrac12 g_{\tilde\rho}\left(\frac{\partial\tilde\rho}{\partial s},\frac{\partial\tilde\rho}{\partial s}\right)\right]\big(s,\alpha(s)t\big)\\&\quad+\alpha'(s)t\left[g_{\tilde\rho}\left(\frac{\partial\tilde\rho}{\partial s},\frac{\partial\tilde\rho}{\partial t}\right)\right]\big(s,\alpha(s)t\big)\\&\quad+\tfrac12\big(\alpha'(s)t\big)^2\left[g_{\tilde\rho}\left(\frac{\partial\tilde\rho}{\partial t},\frac{\partial\tilde\rho}{\partial t}\right)\right]\big(s,\alpha(s)t\big).\end{aligned}$$

Using this identity in (3.13) then gives

$$\begin{aligned} 0 &= \int_0^1 \alpha \,\frac{\partial}{\partial t}_{|t=0} \tfrac{1}{2} g_{\tilde\rho}\left(\frac{\partial\tilde\rho}{\partial s}, \frac{\partial\tilde\rho}{\partial s}\right) ds + \int_0^1 \alpha' \, g_{\tilde\rho}\left(\frac{\partial\tilde\rho}{\partial s}, \frac{\partial\tilde\rho}{\partial t}\right)_{|t=0} ds \\ &= \int_0^1 \alpha \left\{ \frac{\partial}{\partial t} \,\tfrac{1}{2} g_{\tilde\rho}\left(\frac{\partial\tilde\rho}{\partial s}, \frac{\partial\tilde\rho}{\partial s}\right) - \frac{\partial}{\partial s}\, g_{\tilde\rho}\left(\frac{\partial\tilde\rho}{\partial s}, \frac{\partial\tilde\rho}{\partial t}\right)\right\}_{|t=0} ds. \end{aligned}$$

This proves (3.12) because α was arbitrary. Consider now the family of gradient flows $\rho = \rho(s,\cdot)$ satisfying (3.9). For any $s_0 \in [0,1]$ there exists a map $\tilde\rho$ such that

$$\left\{ \begin{array}{c} [0,1] \ni s \mapsto \tilde\rho(s,t) \text{ is a geodesic} \\ \tilde\rho(s_0,t) = \rho(s_0,t) \\ \frac{\partial\tilde\rho}{\partial s}(s_0,t) = \frac{\partial\rho}{\partial s}(s_0,t) \end{array} \right\} \quad \forall\, t \in [0,\infty). \tag{3.14}$$

At $s = s_0$ we then find

$$\begin{aligned} \frac{\partial}{\partial t}\,\tfrac{1}{2} g_\rho\left(\frac{\partial\rho}{\partial s}, \frac{\partial\rho}{\partial s}\right) \overset{(3.14)}{=} \frac{\partial}{\partial t}\,\tfrac{1}{2} g_{\tilde\rho}\left(\frac{\partial\tilde\rho}{\partial s}, \frac{\partial\tilde\rho}{\partial s}\right) &\overset{(3.12)}{=} \frac{\partial}{\partial s}\, g_{\tilde\rho}\left(\frac{\partial\tilde\rho}{\partial s}, \frac{\partial\tilde\rho}{\partial t}\right) \\ &\overset{(3.14)}{=} \frac{\partial}{\partial s}\, g_{\tilde\rho}\left(\frac{\partial\tilde\rho}{\partial s}, \frac{\partial\rho}{\partial t}\right) \\ &\overset{(2.2)}{=} \frac{\partial}{\partial s}\left[- \left\langle \operatorname{diff} E_{|\rho}, \frac{\partial\tilde\rho}{\partial s}\right\rangle\right] \\ &\overset{(3.14)}{=} \frac{\partial}{\partial s}\left[- \left\langle \operatorname{diff} E_{|\tilde\rho}, \frac{\partial\tilde\rho}{\partial s}\right\rangle\right] = -\frac{\partial^2}{\partial s^2} E(\tilde\rho). \end{aligned}$$

By definition of the Hessian, we have

$$\frac{\partial^2}{\partial s^2} E(\tilde\rho) = g_{\tilde\rho}\left(\frac{\partial\tilde\rho}{\partial s}, \operatorname{Hess} E_{|\tilde\rho} \frac{\partial\tilde\rho}{\partial s}\right) \overset{(3.14)}{=} g_\rho\left(\frac{\partial\rho}{\partial s}, \operatorname{Hess} E_{|\rho} \frac{\partial\rho}{\partial s}\right),$$

and (3.11) follows.

3.2. Heuristics: Convexity and induced metric. In this section we show heuristically how the abstract framework of the previous section yields a contraction property in the Wasserstein distance for the porous medium equation. This argument will be made rigorous in the remainder of the paper.

We recall the heuristic argument for the convexity of E on $(\mathcal{M}, g)$ for which we probe the convexity along geodesics. Therefore we start by heuristically deriving the equation for geodesics, essentially reproducing [13]. An alternative heuristic derivation can be found in [12]. Notice first that within the abstract framework, the geodesic equation is the Euler–Lagrange equation (i.e., the first variation) of the action functional (3.6). In view of (2.6), our action functional for a curve in $\mathcal{M}$, i.e., for a function $\rho\colon [0,1] \times \mathbb{M}^n \to [0,\infty)$ with $\int_{\mathbb{M}^n} \rho(s,x)\,dx = 1$ for all $s \in [0,1]$, takes the form

$$\mathcal{A}(\rho) = \iint_{[0,1]\times\mathbb{M}^n} \tfrac{1}{2} |\nabla\phi|^2 \,\rho \,dx\,ds, \tag{3.15}$$

where the function $\phi\colon [0,1] \times \mathbb{M}^n \to \mathbb{R}$ is determined by

$$\partial_s \rho - \nabla \cdot (\rho \nabla \phi) = 0 \tag{3.16}$$

and plays the role of the tangent vector field along the curve. Like for the metric tensor itself (cf. (2.9)), the action functional can be written variationally:

$$\begin{aligned}\mathcal{A}(\rho) &= \sup_\phi \bigg\{ -\iint_{[0,1]\times\mathbb{M}^n} \tfrac{1}{2}|\nabla\phi|^2 \,\rho\, dx\, ds - \iint_{[0,1]\times\mathbb{M}^n} \phi\, \partial_s \rho\, dx\, ds \bigg\} \\ &= \sup_\phi \bigg\{ -\iint_{[0,1]\times\mathbb{M}^n} \tfrac{1}{2}|\nabla\phi|^2 \,\rho\, dx\, ds + \iint_{[0,1]\times\mathbb{M}^n} \partial_s\phi\, \rho\, dx\, ds \\ &\qquad + \int_{\mathbb{M}^n} \phi(0,x)\, \rho_0(x)\, dx - \int_{\mathbb{M}^n} \phi(1,x)\, \rho_1(x)\, dx \bigg\},\end{aligned} \tag{3.17}$$

where the sup is taken over all smooth functions $\phi\colon [0,1]\times\mathbb{M}^n \to \mathbb{R}$. Here ρ_0, ρ_1 are the fixed endpoints of the curve; i.e., we have

$$\rho(0,\cdot) = \rho_0 \quad \text{and} \quad \rho(1,\cdot) = \rho_1. \tag{3.18}$$

To obtain the induced distance in $\mathcal{M}$, the expression (3.17) needs to be minimized over all functions $\rho\colon [0,1]\times\mathbb{M}^n \to [0,\infty)$ with $\int_{\mathbb{M}^n} \rho(\cdot,x)\,dx = 1$; see (3.5). In fact, we may think of minimizing (3.17) over all functions $\rho\colon [0,1]\times\mathbb{M}^n \to \mathbb{R}$ because (3.17) is $+\infty$ if (3.16) or (3.18) is violated. Maximizing in ϕ and minimizing in ρ amounts to a saddle-point problem. The first variation in ϕ is given by (3.16) and (3.18). The first variation in ρ is given by the Hamilton–Jacobi equation

$$\partial_s \phi - \tfrac{1}{2}|\nabla\phi|^2 = 0. \tag{3.19}$$

Hence the combination of the transport equation (3.16) and the Hamilton–Jacobi equation (3.19) forms the geodesic equation. Note that the system (3.16) and (3.19) is of hyperbolic nature as a partial differential equation. The velocity $u = -\nabla\phi$ satisfies the "pressureless Euler equation"

$$\frac{D}{ds} u + Du\, u = 0,$$

and thus the flow $\partial_s \Phi = u \circ \Phi$ consists of geodesic trajectories, i.e., $\frac{D}{ds}\frac{\partial}{\partial s}\Phi = 0$. Notice that (3.16) states that $\rho(s,\cdot)$ is the push-forward of $\rho(s=0)$ under $\Phi(s,\cdot)$. This is what we call the Lagrangian approach. Geodesics in the sense of shortest curves were given a rigorous meaning for a Riemannian manifold $\mathbb{M}^n$ in [11].

Having identified the geodesic equation, we can probe the convexity of (1.2) along geodesics. This was first done in the Lagrangian framework in [10] and gave rise to the notion of displacement convexity. We reproduce the heuristic Eulerian argument from [13]. Let $\rho\colon [0,1]\times\mathbb{M}^n \to [0,\infty)$ be a geodesic with tangent field $\phi\colon [0,1]\times\mathbb{M}^n \to \mathbb{R}$; i.e., let (3.16) and (3.19) be satisfied. As in (2.10), we find for the first derivative

$$\frac{dE}{ds} = \int_{\mathbb{M}^n} \Delta U(\rho)\, \phi\, dx = \int_{\mathbb{M}^n} U(\rho)\, \Delta\phi\, dx.$$

For the second derivative, we obtain

$$\begin{aligned}
\frac{d^2E}{ds^2} &= \int_{\mathbb{M}^n} \Big(U'(\rho)\,\partial_s\rho\,\Delta\phi + U(\rho)\,\Delta\partial_s\phi\Big)\,dx \\
&\overset{(3.16)}{\underset{(3.19)}{=}} \int_{\mathbb{M}^n} \Big(U'(\rho)\,\nabla\cdot(\rho\nabla\phi)\,\Delta\phi + U(\rho)\,\Delta\tfrac{1}{2}|\nabla\phi|^2\Big)\,dx \\
&= \int_{\mathbb{M}^n} \Big(\rho U'(\rho)\,(\Delta\phi)^2 + \nabla U(\rho)\cdot\nabla\phi\,\Delta\phi + U(\rho)\,\Delta\tfrac{1}{2}|\nabla\phi|^2\Big)\,dx \\
&= \int_{\mathbb{M}^n} \Big(\rho U'(\rho)\,(\Delta\phi)^2 + U(\rho)\Big(-\nabla\cdot(\nabla\phi\,\Delta\phi) + \Delta\tfrac{1}{2}|\nabla\phi|^2\Big)\Big)\,dx \\
&= \int_{\mathbb{M}^n} \Big(\Big(\rho U'(\rho) - U(\rho)\Big)(\Delta\phi)^2 + U(\rho)\Big(-\nabla\phi\cdot\nabla\Delta\phi + \Delta\tfrac{1}{2}|\nabla\phi|^2\Big)\Big)\,dx.
\end{aligned}$$

We appeal to Bochner's formula (see [14]):

$$-\nabla\phi\cdot\nabla\Delta\phi + \Delta\tfrac{1}{2}|\nabla\phi|^2 = |\mathrm{D}^2\phi|^2 + \nabla\phi\cdot\mathrm{Ric}\,\nabla\phi,$$

where $\mathrm{D}^2\phi$ denotes the Hessian of ϕ, $|A|^2$ stands for the trace of A^tA, and Ric denotes the Ricci curvature of $\mathbb{M}^n$. We thus obtain the formula

$$\frac{d^2E}{ds^2} = \int_{\mathbb{M}^n} \Big(\Big(\rho U'(\rho) - U(\rho)\Big)(\Delta\phi)^2 + U(\rho)\Big(|\mathrm{D}^2\phi|^2 + \nabla\phi\cdot\mathrm{Ric}\,\nabla\phi\Big)\Big)\,dx. \tag{3.20}$$

In view of (3.1), the right-hand side of (3.20) can be understood as the quadratic part of the Hessian of E in the direction of the infinitesimal variation $\delta\rho = \nabla\cdot(\rho\nabla\phi)$. We notice that it is nonnegative for all functions $\rho \geqslant 0$ and ϕ if and only if

$$\rho U'(\rho) \geqslant \big(1 - \tfrac{1}{n}\big)U(\rho) \geqslant 0 \quad \text{and} \quad \mathrm{Ric}(x) \geqslant 0 \quad \forall\ x \in \mathbb{M}^n$$

because $(\Delta\phi)^2 \leqslant n|\mathrm{D}^2\phi|^2$. The convexity of E along geodesics in the Riemannian case $\mathbb{M}^n$ was given a rigorous meaning in [7].

To conclude, it remains only to prove that (3.5) with (3.15) and (3.16) coincides with $\frac{1}{2}\mathcal{W}^2(\rho_0, \rho_1)$. Recall that for $\rho_0, \rho_1 \in \mathrm{Prob}(\mathbb{M}^n)$, $\mathcal{W}^2(\rho_0, \rho_1)$ is defined as

$$\inf\left\{ \iint_{\mathbb{M}^n\times\mathbb{M}^n} d(x,y)^2\,d\pi(x,y) \;\middle|\; \begin{array}{l} \pi \in \mathrm{Prob}(\mathbb{M}^n\times\mathbb{M}^n), \\ \displaystyle\int_{\mathbb{M}^n} d\pi(\cdot,y) = \rho_0, \int_{\mathbb{M}^n} d\pi(x,\cdot) = \rho_1 \end{array} \right\};$$

cf. [18]. Several heuristic arguments are possible here (cf. [12] and [13]). However, the rigorous proof we provide in the next section is no more difficult than a heuristic one; therefore we refer the reader to Proposition 4.3.

4. Rigorous result: Contraction. We recall that $\mathbb{M}^n$ is a compact connected Riemannian manifold without boundary, with geodesic distance d and $\int_{\mathbb{M}^n} 1\,dx = 1$.

Here is our main result.

THEOREM 4.1 (contraction estimate). *Assume that $\rho U'(\rho) \geqslant \big(1 - \frac{1}{n}\big)U(\rho) \geqslant 0$ for all $\rho \geqslant 0$ and that $\mathrm{Ric}(x) \geqslant 0$ for all $x \in \mathbb{M}^n$. For nonnegative initial data $\bar\rho_0, \bar\rho_1$ with $\int_{\mathbb{M}^n} \bar\rho_i\,dx = 1$ consider solutions ρ_i of the porous medium equation*

$$\left.\begin{array}{r} \partial_t\rho_i - \Delta U(\rho_i) = 0 \\ \rho_i(t=0) = \bar\rho_i \end{array}\right\} \quad \textit{for } i = 0, 1.$$

Then the Wasserstein distance of ρ_0 and ρ_1 is nonincreasing in time, i.e.,

$$\frac{d^+}{dt}\mathcal{W}^2(\rho_0,\rho_1) \leqslant 0. \tag{4.1}$$

Remark 4.1. We have seen in section 3.2 that heuristically the convexity of E is equivalent to the conditions on U and Ric required in Theorem 4.1. We have seen in section 3.1 that convexity of E is equivalent to the contractivity of the corresponding gradient flow. Hence we expect that the conditions on U and Ric are also necessary. This has been rigorously proven in [16]. Also in [16], the sufficiency of these conditions has been established using the Lagrangian approach mentioned in section 1 which relies on [11].

The theorem will be a consequence of the following two propositions.

PROPOSITION 4.2. *Assume that $\rho U'(\rho) \geqslant \big(1-\frac{1}{n}\big)U(\rho) \geqslant 0$ for all $\rho \geqslant 0$ and that $\mathrm{Ric}(x) \geqslant 0$ for all $x \in \mathbb{M}^n$. Consider a family of smooth positive solutions of*

$$\partial_t\rho - \Delta U(\rho) = 0, \tag{4.2}$$

depending smoothly on the parameter $s \in [0,1]$. For any (s,t) let ϕ be defined by

$$\partial_s\rho - \nabla\cdot(\rho\nabla\phi) = 0.$$

Then the following holds:

$$\frac{d}{dt}\iint_{[0,1]\times\mathbb{M}^n} |\nabla\phi|^2\rho\,dx\,ds \leqslant 0.$$

Remark 4.2. Proposition 4.2 is guided by the abstract observation of section 3.1: Convexity can be probed by the gradient flow. More precisely, convexity expresses itself by the fact that the action of curves is reduced when the points along the curve are evolved by the gradient flow.

PROPOSITION 4.3. *Consider $\rho_0\,dx, \rho_1\,dx \in \mathrm{Prob}(\mathbb{M}^n)$, where ρ_0,ρ_1 are smooth and positive functions. Then the Wasserstein distance squared $\frac{1}{2}\mathcal{W}^2(\rho_0,\rho_1)$ equals*

$$\inf\left\{\iint_{[0,1]\times\mathbb{M}^n} \tfrac{1}{2}|\nabla\phi|^2\rho\,dx\,ds \;\middle|\; \begin{array}{l}(\rho>0,\phi) \text{ smooth functions on } [0,1]\times\mathbb{M}^n,\\ \partial_s\rho - \nabla\cdot(\rho\nabla\phi) = 0,\ \left\{\begin{array}{l}\rho(0,\cdot)=\rho_0\\ \rho(1,\cdot)=\rho_1\end{array}\right\}\end{array}\right\}. \tag{4.3}$$

Proof of Theorem 4.1. Assume first that the initial data are smooth and positive and that U is linear for $\rho \notin [\alpha, 1/\alpha]$ with $\alpha > 0$ small. Then standard parabolic theory yields that solutions of the porous medium equation for smooth and positive initial data are also smooth and positive. By Proposition 4.3 we can, for any $\varepsilon > 0$, find smooth functions $\big(\bar\rho > 0, \bar\phi\big)$ on $[0,1]\times\mathbb{M}^n$, with

$$\partial_s\bar\rho - \nabla\cdot\big(\bar\rho\nabla\bar\phi\big) = 0,\ \left\{\begin{array}{l}\bar\rho(0,\cdot)=\bar\rho_0\\ \bar\rho(1,\cdot)=\bar\rho_1\end{array}\right\}$$

such that

$$\iint_{[0,1]\times\mathbb{M}^n} \big|\nabla\bar\phi\big|^2\bar\rho\,dx\,ds \leqslant \mathcal{W}^2(\bar\rho_0,\bar\rho_1) + \varepsilon.$$

For any $s \in [0,1]$, let $\bar{\rho}(\cdot, s)$ evolve according to the porous medium equation. This yields a family ρ of solutions depending smoothly on s for which Proposition 4.2 applies. Using again the characterization of Proposition 4.3 then yields

$$\begin{aligned} \mathcal{W}^2\big(\rho_0(t), \rho_1(t)\big) &\leqslant \iint_{[0,1]\times\mathbb{M}^n} |\nabla\phi(t)|^2 \rho(t)\, dx\, ds \\ &\leqslant \iint_{[0,1]\times\mathbb{M}^n} \big|\nabla\bar{\phi}\big|^2 \bar{\rho}\, dx\, ds \leqslant \mathcal{W}^2(\bar{\rho}_0, \bar{\rho}_1) + \varepsilon \quad \forall\, t > 0. \end{aligned} \tag{4.4}$$

Since $\varepsilon > 0$ was arbitrary, we obtain (4.1) in this case.

The general case follows by an approximation argument that we do not discuss in detail here. For general nonnegative initial data one can find sequences of smooth positive functions, converging strongly to the given $\bar{\rho}_0, \bar{\rho}_1$. Then standard theory for the porous medium equation yields that the solutions converge strongly in $L^1(\mathbb{M}^n)$, hence a posteriori also in the Wasserstein distance which metrizes the weak* topology of measures. Therefore the contraction estimate generalizes to this setting. Similarly, one can approximate a given U with $\rho U'(\rho) \geqslant \big(1 - \frac{1}{n}\big) U(\rho) \geqslant 0$ for all $\rho \geqslant 0$ by a sequence of functions that have the same property and are linear for small and large ρ and that converge uniformly. Then standard theory applies and allows one to conclude. We refer the reader to [12], where this program has been carried out in $\mathbb{R}^n$. □

Proof of Proposition 4.2. The following remark is at the core of Proposition 4.2.

LEMMA 4.4. *Consider smooth functions* $\big(\rho > 0, \delta\rho\big)$ *on* $[0,\infty) \times \mathbb{M}^n$ *solving*

$$\left\{ \begin{aligned} &\partial_t \rho \quad\;\; - \Delta U(\rho) &&= 0, \\ &\partial_t(\delta\rho) - \Delta\Big(U'(\rho)\, \delta\rho\Big) &&= 0. \end{aligned} \right. \tag{4.5}$$

For any t let ϕ be defined by

$$\delta\rho - \nabla \cdot \big(\rho \nabla\phi\big) = 0. \tag{4.6}$$

Then we have

$$\begin{aligned} &\frac{d}{dt} \int_{\mathbb{M}^n} \tfrac{1}{2} |\nabla\phi|^2 \rho\, dx \\ &\quad = - \int_{\mathbb{M}^n} \bigg(\Big(\rho U'(\rho) - U(\rho)\Big)(\Delta\phi)^2 + U(\rho)\Big(|\mathrm{D}^2\phi|^2 + \nabla\phi \cdot \mathrm{Ric}\, \nabla\phi\Big)\bigg)\, dx. \end{aligned} \tag{4.7}$$

Remark 4.3. Observe that the second equation in (4.5) describes the evolution of an infinitesimal perturbation $\delta\rho$ of ρ; see (3.2). Notice further that in view of (4.6), the left-hand side of (4.7) measures how the squared norm of $\delta\rho$ changes in time; cf. (3.3). Observe finally that the right-hand side expression of (4.7) coincides with what we expect to be—up to the sign—the Hessian; see (3.20). In this sense the formula (4.7) reproduces (3.4).

Proof. The left-hand side of (4.7) equals, after an integration by parts,

$$\frac{d}{dt} \int_{\mathbb{M}^n} \tfrac{1}{2} |\nabla\phi|^2 \rho\, dx = \int_{\mathbb{M}^n} \Big(- \phi\, \nabla \cdot \big(\rho\, \partial_t \nabla\phi\big) + \tfrac{1}{2} |\nabla\phi|^2\, \partial_t \rho\Big)\, dx. \tag{4.8}$$

We express $-\nabla\cdot(\rho\,\partial_t\nabla\phi)$ in terms of ρ and ϕ. We find by differentiating (4.6)

$$\begin{aligned} &-\nabla\cdot\big(\rho\,\partial_t\nabla\phi\big)\\ &\qquad = \; -\partial_t(\delta\rho)+\nabla\cdot\big(\partial_t\rho\,\nabla\phi\big)\\ &\qquad \overset{(4.5)}{=} -\Delta\Big(U'(\rho)\,\delta\rho\Big)+\nabla\cdot\Big(\Delta U(\rho)\,\nabla\phi\Big)\\ &\qquad \overset{(4.6)}{=} -\Delta\Big(U'(\rho)\,\nabla\cdot(\rho\nabla\phi)\Big)+\nabla\cdot\Big(\Delta U(\rho)\,\nabla\phi\Big)\\ &\qquad = \; -\Delta\Big(\big(\rho U'(\rho)-U(\rho)\big)\,\Delta\phi\Big)-\Delta\nabla\cdot\Big(U(\rho)\,\nabla\phi\Big)+\nabla\cdot\Big(\Delta U(\rho)\,\nabla\phi\Big). \end{aligned}$$

Using this identity and (4.5) in (4.8) gives, after throwing all derivatives onto ϕ,

$$\begin{aligned} &\frac{d}{dt}\int_{\mathbb{M}^n}\tfrac12|\nabla\phi|^2\rho\,dx\\ &\quad = -\int_{\mathbb{M}^n}\bigg(\Big(\rho U'(\rho)-U(\rho)\Big)(\Delta\phi)^2+U(\rho)\Big(-\nabla\Delta\phi\cdot\nabla\phi+\Delta\tfrac12|\nabla\phi|^2\Big)\bigg)\,dx. \end{aligned}$$

Then we use Bochner's formula

$$-\nabla\Delta\phi\cdot\nabla\phi+\Delta\tfrac12|\nabla\phi|^2=|\mathrm{D}^2\phi|^2+\nabla\phi\cdot\mathrm{Ric}\,\nabla\phi$$

(see Proposition 3.3 of [14]) to conclude. ◻

Fix $s\in[0,1]$, and let $\delta\rho=\partial_s\rho(\cdot,s)$. Differentiating (4.2) with respect to s gives

$$\partial_t\big(\partial_s\rho\big)-\Delta\Big(U'(\rho)\,\partial_s\rho\Big)=0.$$

Then Lemma 4.4 applies and yields

$$\begin{aligned} &\frac{d}{dt}\int_{\mathbb{M}^n}\tfrac12|\nabla\phi|^2\rho\,dx\\ &\quad = -\int_{\mathbb{M}^n}\bigg(\Big(\rho U'(\rho)-U(\rho)\Big)(\Delta\phi)^2+U(\rho)\Big(|\mathrm{D}^2\phi|^2+\nabla\phi\cdot\mathrm{Ric}\,\nabla\phi\Big)\bigg)\,dx. \end{aligned}\tag{4.9}$$

Notice that $(\Delta\phi)^2\leqslant n|\mathrm{D}^2\phi|^2$. By the assumption on U, we therefore get

$$\Big(\rho U'(\rho)-U(\rho)\Big)(\Delta\phi)^2+U(\rho)|\mathrm{D}^2\phi|^2\geqslant U(\rho)\Big(-\tfrac1n(\Delta\phi)^2+|\mathrm{D}^2\phi|^2\Big)\geqslant 0.$$

Furthermore, we have $\nabla\phi\cdot\mathrm{Ric}\,\nabla\phi\geqslant 0$. This proves the proposition. ◻

Remark 4.4. The same reasoning also yields convergence rates: In fact, if

$$U(\rho)\;\xi\cdot\mathrm{Ric}(x)\xi\geqslant\lambda\rho|\xi|^2\quad\forall\rho\geqslant 0\text{ and }(x,\xi)\in T\mathbb{M}^n\tag{4.10}$$

for a suitable constant $\lambda\in\mathbb{R}$, then (4.9) gives

$$\frac{d}{dt}\int_{\mathbb{M}^n}\tfrac12|\nabla\phi|^2\rho\,dx=-\int_{\mathbb{M}^n}U(\rho)\Big(\nabla\phi\cdot\mathrm{Ric}\,\nabla\phi\Big)\,dx\leqslant-2\lambda\int_{\mathbb{M}^n}\tfrac12|\nabla\phi|^2\rho\,dx.$$

We obtain exponential decay of $\int_{\mathbb{M}^n}\frac12|\nabla\phi|^2\rho\,dx$ with rate 2λ, thus of $\mathcal{W}^2(\rho_1,\rho_0)$, by (4.4). For the heat equation on the unit sphere, for example, condition (4.10) is satisfied with constant $\lambda=1$.

Proof of Proposition 4.3. We proceed in five steps.

Step 1. We first prove that $\frac{1}{2}\mathcal{W}^2(\rho_0,\rho_1) \leqslant (4.3)$. Therefore assume that (ρ,ϕ) is admissible in (4.3). For abbreviation we introduce the velocity field $u := -\nabla\phi$, such that $\partial_s\rho + \nabla\cdot(\rho u) = 0$, and consider the flow induced by u:

$$\Phi\colon [0,1]\times\mathbb{M}^n \longrightarrow \mathbb{M}^n \quad\text{with}\quad \partial_s\Phi(s,x) = u\big(s,\Phi(s,x)\big),\ \Phi(0,x) = x, \tag{4.11}$$

for all $(s,x)\in[0,1]\times\mathbb{M}^n$. Then the measure $\rho(s,x)\,dx$ is the push-forward of the measure $\rho_0(x)\,dx$ under $\Phi(s,\cdot)$; i.e., we have for all smooth functions ζ on $\mathbb{M}^n$

$$\int_{\mathbb{M}^n}\zeta(x)\rho(s,x)\,dx = \int_{\mathbb{M}^n}\zeta\big(\Phi(s,x)\big)\rho_0(x)\,dx \quad \forall\, s\in[0,1]. \tag{4.12}$$

Moreover, by definition of the geodesic distance d we have

$$d\big(x,\Phi(1,x)\big)^2 \leqslant \int_0^1 |\partial_s\Phi(s,x)|^2\,ds. \tag{4.13}$$

Let π be the nonnegative measure defined by

$$\iint_{\mathbb{M}^n\times\mathbb{M}^n}\zeta(x,y)\,d\pi(x,y) = \int_{\mathbb{M}^n}\zeta\big(x,\Phi(1,x)\big)\rho_0(x)\,dx \tag{4.14}$$

for all smooth functions ζ on $\mathbb{M}^n\times\mathbb{M}^n$. Thanks to (4.12), π is admissible in the definition of the Wasserstein distance $\mathcal{W}^2(\rho_0,\rho_1)$. Furthermore, we have

$$\begin{aligned}
&\iint_{\mathbb{M}^n\times\mathbb{M}^n} d(x,y)^2\,d\pi(x,y)\\
&\overset{(4.14)}{=} \int_{\mathbb{M}^n} d\big(x,\Phi(1,x)\big)^2\rho_0(x)\,dx\\
&\overset{(4.13)}{\leqslant} \int_{\mathbb{M}^n}\left(\int_0^1 |\partial_s\Phi(s,x)|^2\,ds\right)\rho_0(x)\,dx\\
&\overset{(4.11)}{=} \int_0^1\int_{\mathbb{M}^n}\big|u\big(s,\Phi(s,x)\big)\big|^2\rho_0(x)\,dx\,ds\\
&\overset{(4.12)}{=} \int_0^1\int_{\mathbb{M}^n}|u(s,x)|^2\rho(s,x)\,dx\,ds = \iint_{[0,1]\times\mathbb{M}^n}|\nabla\phi|^2\rho\,dx\,ds.
\end{aligned}$$

This proves our claim.

Step 2. Notice that any smooth vector field on $[0,1]\times\mathbb{M}^n$ can be identified with a pair (ρ,m), where ρ is a function on $[0,1]\times\mathbb{M}^n$ and m is an s-dependent vector field on $\mathbb{M}^n$ (such as $m=-\rho\nabla\phi$). We will now show that (4.3) equals

$$\inf\left\{\iint_{[0,1]\times\mathbb{M}^n}\tfrac{1}{2}\rho^{-1}|m|^2\,dx\,ds \;\middle|\; \begin{array}{l}(\rho>0,m)\text{ smooth vector field on }[0,1]\times\mathbb{M}^n,\\ \partial_s\rho+\nabla\cdot m = 0,\ \left\{\begin{array}{l}\rho(0,\cdot)=\rho_0\\ \rho(1,\cdot)=\rho_1\end{array}\right\}\end{array}\right\}. \tag{4.15}$$

That (4.15) does not exceed (4.3) is obvious. To prove the converse consider an admissible pair (ρ,m) in the sense of (4.15). By positivity of ρ we then find, for any $s\in[0,1]$, a smooth function ϕ on $\mathbb{M}^n$ solving the elliptic equation

$$\nabla\cdot\big(m+\rho\nabla\phi\big) = 0 \quad\text{on }\mathbb{M}^n. \tag{4.16}$$

This ϕ depends smoothly on s because (ρ, m) does. Since $\mathbb{M}^n$ has no boundary,

$$\int_{\mathbb{M}^n} \big(m + \rho\nabla\phi\big) \cdot \nabla\phi \, dx = 0.$$

Therefore by Cauchy–Schwarz

$$\begin{aligned}\int_{\mathbb{M}^n} \rho|\nabla\phi|^2 \, dx &\overset{(4.16)}{=} \int_{\mathbb{M}^n} -m \cdot \nabla\phi \, dx \\ &\leqslant \left(\int_{\mathbb{M}^n} \rho^{-1}|m|^2 \, dx\right)^{1/2} \left(\int_{\mathbb{M}^n} \rho|\nabla\phi|^2 \, dx\right)^{1/2},\end{aligned}$$

and thus

$$\int_{\mathbb{M}^n} \tfrac{1}{2}\rho|\nabla\phi|^2 \, dx \leqslant \int_{\mathbb{M}^n} \tfrac{1}{2}\rho^{-1}|m|^2 \, dx.$$

Step 3. Now we generalize the functional (4.15) to a certain class of distributions and prove that then the inf is bounded by the Wasserstein distance for any measures $\rho_0, \rho_1 \in \mathrm{Prob}(\mathbb{M}^n)$. To achieve this, notice first that thanks to the Riemannian metric on $\mathbb{M}^n$, any smooth 1-form ω on $[0,1] \times \mathbb{M}^n$ can be identified with a pair (σ, ξ), where σ is a function on $[0,1] \times \mathbb{M}^n$ and ξ is an s-dependent vector field on $\mathbb{M}^n$, via

$$\big\langle \omega,\ (\rho, m)\big\rangle = \sigma\rho + \xi \cdot m$$

for all (smooth) vector fields (ρ, m). We write

$$\omega = \sigma \, ds + \xi \cdot dx.$$

The space of 1-forms can be topologized as usual in the theory of distributions, but we do not want to go into details and refer the reader to [8, 9] instead. A linear functional on the space of smooth 1-forms is called a current. Any smooth vector field (ρ, m) defined on $[0,1] \times \mathbb{M}^n$ gives rise to a current T via

$$\Big\langle T,\ \sigma \, ds + \xi \cdot dx\Big\rangle := \iint_{[0,1]\times\mathbb{M}^n} \rho\sigma + m \cdot \xi \, dx \, ds. \tag{4.17}$$

But of course not all currents T can be represented in this form.

We consider currents defined on $[0,1] \times \mathbb{M}^n$ that satisfy

$$\Big\langle T,\ \partial_s\zeta \, ds + \nabla\zeta \cdot dx\Big\rangle = \int_{\mathbb{M}^n} \zeta(1,x) \, d\rho_1(x) - \int_{\mathbb{M}^n} \zeta(0,x) \, d\rho_0(x) \tag{4.18}$$

for all test functions ζ for given $\rho_0, \rho_1 \in \mathrm{Prob}(\mathbb{M}^n)$. If now T is of the form (4.17), then (4.18) is just the weak formulation of the continuity equation $\partial_s\rho + \nabla \cdot m = 0$ with initial and final data ρ_0 and ρ_1. Following [4], we can generalize the action

$$\iint_{[0,1]\times\mathbb{M}^n} \tfrac{1}{2}\rho^{-1}|m|^2 \, dx \, ds \tag{4.19}$$

as follows. For any current T with (4.18) we consider

$$\mathcal{A}(T) := \sup\left\{ \Big\langle T,\ \sigma \, ds + \xi \cdot dx\Big\rangle \;\middle|\; \begin{array}{l}(\sigma, \xi) \text{ smooth vector field on } [0,1] \times \mathbb{M}^n \\ \text{with } \sigma + \tfrac{1}{2}|\xi|^2 \leqslant 0\end{array} \right\}. \tag{4.20}$$

We claim that this $\mathcal{A}(T)$ coincides with (4.19) if T is of the form (4.17): Indeed, setting $\xi = \rho^{-1}m$ and $\sigma = -\frac{1}{2}|\xi|^2 = -\frac{1}{2}\rho^{-2}|m|^2$ shows that (4.19) $\leqslant \mathcal{A}(T)$; and

$$\rho\sigma + m\cdot\xi \leqslant -\rho\,\tfrac{1}{2}|\xi|^2 + m\cdot\xi \leqslant \tfrac{1}{2}\rho^{-1}|m|^2$$

for all admissible (σ,ξ) implies that $\mathcal{A}(T) \leqslant$ (4.19).

Step 4. Now we prove that $\frac{1}{2}\mathcal{W}^2(\rho_0,\rho_1)$ is bigger than or equal to

$$\inf\Big\{\mathcal{A}(T)\,\Big|\,T \text{ current on } [0,1]\times\mathbb{M}^n \text{ satisfying (4.18)}\Big\}. \tag{4.21}$$

Consider any transference plan $\pi \in \operatorname{Prob}(\mathbb{M}^n\times\mathbb{M}^n)$ that is admissible in the definition of $\mathcal{W}^2(\rho_0,\rho_1)$, and let $\Phi\colon [0,1]\times\mathbb{M}^n\times\mathbb{M}^n \longrightarrow \mathbb{M}^n$ be defined by

$$[0,1]\ni s\mapsto \Phi(s,x,y)\quad\text{is the shortest geodesic between } x \text{ and } y.$$

Then we have in particular

$$\int_0^1 |\partial_s\Phi(s,x,y)|^2\,ds = d(x,y)^2 \quad\text{and}\quad \left\{\begin{array}{l}\Phi(0,x,y)=x\\ \Phi(1,x,y)=y\end{array}\right\}. \tag{4.22}$$

We define a current T on $[0,1]\times\mathbb{M}^n$ as follows: For all 1-forms $\sigma\,ds + \xi\cdot dx$ let

$$\begin{aligned}&\Big\langle T,\ \sigma\,ds + \xi\cdot dx\Big\rangle\\ &\quad := \iint\int_0^1 \Big\{\sigma\big(s,\Phi(s,x,y)\big) + \xi\big(s,\Phi(s,x,x)\big)\cdot\partial_s\Phi(s,x,y)\Big\}\,ds\,d\pi(x,y).\end{aligned}$$

This current satisfies the admissibility condition (4.18). Indeed, we have

$$\begin{aligned}&\Big\langle T,\ \partial_s\zeta\,ds + \nabla\zeta\cdot dx\Big\rangle\\ &\quad= \iint\int_0^1 \Big\{\partial_s\zeta\big(s,\Phi(s,x,y)\big) + \nabla\zeta\big(s,\Phi(s,x,y)\big)\cdot\partial_s\Phi(s,x,y)\Big\}\,ds\,d\pi(x,y)\\ &\quad= \iint\int_0^1 \frac{d}{ds}\Big\{\zeta\big(s,\Phi(s,x,y)\big)\Big\}\,ds\,d\pi(x,y)\\ &\quad= \iint \zeta\big(1,\Phi(1,x,y)\big)\,d\pi(x,y) - \iint \zeta\big(0,\Phi(0,x,y)\big)\,d\pi(x,y)\\ &\quad\overset{(4.22)}{=} \iint \zeta(1,y)\,d\pi(x,y) - \iint \zeta(0,x)\,d\pi(x,y)\\ &\quad= \int \zeta(1,y)\,d\rho_1(y) - \int \zeta(0,x)\,d\rho_0(x)\end{aligned}$$

for all test functions ζ. Now we argue that $\mathcal{A}(T) \leqslant \iint \frac{1}{2}d(x,y)^2\,d\pi(x,y)$. Indeed, we have for any vector field (σ, ξ) admissible in (4.20) that

$$\begin{aligned}
&\Big\langle T,\ \sigma\, ds + \xi \cdot dx \Big\rangle \\
&\quad = \iint \int_0^1 \Big\{ \sigma\big(s, \Phi(s,x,y)\big) + \xi\big(s, \Phi(s,x,y)\big) \cdot \partial_s \Phi(s,x,y) \Big\}\, ds\, d\pi(x,y) \\
&\quad \leqslant \iint \int_0^1 \Big\{ -\tfrac{1}{2}\big|\xi\big(s, \Phi(s,x,y)\big)\big|^2 + \xi\big(s, \Phi(s,x,y)\big) \cdot \partial_s \Phi(s,x,y) \Big\}\, ds\, d\pi(x,y) \\
&\quad \leqslant \iint \int_0^1 \tfrac{1}{2} |\partial_s \Phi(s,x,y)|^2\, ds\, d\pi(x,y) \\
&\quad \overset{(4.22)}{=} \iint \tfrac{1}{2} d(x,y)^2\, d\pi(x,y).
\end{aligned}$$

Step 5. To conclude the proof of the proposition it is then sufficient to show that the two inf in (4.15) and (4.21) coincide. This will follow from Proposition 5.1 below, which shows that any current T satisfying the admissibility condition (4.18) for smooth and positive data ρ_0, ρ_1 can in fact be approximated by a current T_ε that is representable by a smooth vector field $(\rho_\varepsilon \geqslant 0, m_\varepsilon)$ in such a way that (4.18) still holds with T_ε in place of T and $\limsup_{\varepsilon \to 0} \mathcal{A}(T_\varepsilon) \leqslant \mathcal{A}(T)$.

The only detail that needs to be settled is (strict) positivity of ρ_ε. We argue as follows. Since $\rho_0, \rho_1 > 0$ and $\mathbb{M}^n$ is compact, there exists $0 < \delta < 1$ with $\rho_0, \rho_1 \geqslant \delta$. Recall that by assumption $\int_{\mathbb{M}^n} 1\, dx = 1$. Then we consider

$$\tilde{\rho}_0 := \frac{\rho_0 - \delta}{1 - \delta} \quad \text{and} \quad \tilde{\rho}_1 := \frac{\rho_1 - \delta}{1 - \delta},$$

which are in $\mathrm{Prob}(\mathbb{M}^n)$. Let $\tilde{T}$ be the current constructed in Step 4, based on an admissible transference plan π in the definition of $\mathcal{W}^2(\tilde{\rho}_0, \tilde{\rho}_1)$. As shown there,

$$\mathcal{A}(\tilde{T}) \leqslant \tfrac{1}{2} \mathcal{W}^2(\tilde{\rho}_0, \tilde{\rho}_1). \tag{4.23}$$

We apply Proposition 5.1 to $\tilde{T}$. This gives an approximation $\tilde{T}_\varepsilon$ that satisfies (4.18) and is representable by smooth vector fields $(\tilde{\rho}_\varepsilon \geqslant 0, \tilde{m}_\varepsilon)$, such that

$$\limsup_{\varepsilon \to 0} \mathcal{A}(\tilde{T}_\varepsilon) \leqslant \mathcal{A}(\tilde{T}). \tag{4.24}$$

In view of the remark in Step 3, the admissibility condition (4.18) amounts to

$$\partial_s \tilde{\rho}_\varepsilon + \nabla \cdot \tilde{m}_\varepsilon = 0, \ \left\{ \begin{array}{l} \tilde{\rho}_\varepsilon(0, \cdot) = \tilde{\rho}_0 \\ \tilde{\rho}_\varepsilon(1, \cdot) = \tilde{\rho}_1 \end{array} \right\}.$$

Now notice that $\tilde{\rho}_0$ and $\tilde{\rho}_1$ are constructed in such a way that

$$(\rho_\varepsilon, m_\varepsilon) := \big((1-\delta)\tilde{\rho}_\varepsilon + \delta, (1-\delta)\tilde{m}_\varepsilon \big)$$

is admissible in (4.15) because $\rho_\varepsilon \geqslant \delta > 0$. We have $\rho_\varepsilon^{-1} |m_\varepsilon|^2 \leqslant (1-\delta)\, \tilde{\rho}_\varepsilon^{-1} |\tilde{m}_\varepsilon|^2$, and thus by the remark in Step 3

$$\iint_{[0,1] \times \mathbb{M}^n} \tfrac{1}{2} \rho_\varepsilon^{-1} |m_\varepsilon|^2\, dx\, ds \leqslant (1-\delta) \iint_{[0,1] \times \mathbb{M}^n} \tfrac{1}{2} \tilde{\rho}_\varepsilon^{-1} |\tilde{m}_\varepsilon|^2 = (1-\delta)\, \mathcal{A}(\tilde{T}_\varepsilon). \tag{4.25}$$

In view of (4.23), (4.24), and (4.25) it remains to argue that

$$\mathcal{W}(\tilde{\rho}_0,\tilde{\rho}_1) \leqslant \mathcal{W}(\rho_0,\rho_1) + o(1) \quad \text{as } \delta \to 0.$$

By the triangle inequality for the Wasserstein distance (see Theorem 7.3 of [18]),

$$\mathcal{W}(\tilde{\rho}_0,\tilde{\rho}_1) \leqslant \mathcal{W}(\rho_0,\rho_1) + \mathcal{W}(\rho_0,\tilde{\rho}_0) + \mathcal{W}(\rho_1,\tilde{\rho}_1). \tag{4.26}$$

In order to conclude, it suffices therefore to prove that the last two terms on the right-hand side of (4.26) can be made small by choosing δ appropriately. We consider the transference plan $\pi \in \mathrm{Prob}(\mathbb{M}^n \times \mathbb{M}^n)$ defined by

$$\begin{aligned} &\iint_{\mathbb{M}^n\times\mathbb{M}^n} \zeta(x,y)\,d\pi(x,y) \\ &:= \int_{\mathbb{M}^n} \zeta(x,x)\,\big(\rho_0(x)-\delta\big)\,dx + \tfrac{\delta}{1-\delta} \iint_{\mathbb{M}^n\times\mathbb{M}^n} \zeta(x,y)\,\big(\rho_0(y)-\delta\big)\,dx\,dy \end{aligned} \tag{4.27}$$

for all ζ. This π is admissible in the definition of $\mathcal{W}^2(\rho_0,\tilde{\rho}_0)$ because

$$\begin{aligned} &\iint_{\mathbb{M}^n\times\mathbb{M}^n} \zeta(x)\,d\pi(x,y) \\ &\quad= \int_{\mathbb{M}^n} \zeta(x)\,\big(\rho_0(x)-\delta\big)\,dx + \tfrac{\delta}{1-\delta}\int_{\mathbb{M}^n}\zeta(x)\,dx \int_{\mathbb{M}^n}\big(\rho_0(y)-\delta\big)\,dy \\ &\quad= \int_{\mathbb{M}^n} \zeta(x)\,\rho_0(x)\,dx \end{aligned}$$

and similarly

$$\begin{aligned} &\iint_{\mathbb{M}^n\times\mathbb{M}^n} \zeta(y)\,d\pi(x,y) \\ &\quad= \int_{\mathbb{M}^n} \zeta(y)\,\big(\rho_0(y)-\delta\big)\,dy + \tfrac{\delta}{1-\delta}\int_{\mathbb{M}^n} dx \int_{\mathbb{M}^n}\zeta(y)\,\big(\rho_0(y)-\delta\big)\,dy \\ &\quad= \int_{\mathbb{M}^n} \zeta(y)\,\frac{\rho_0(y)-\delta}{1-\delta}\,dy. \end{aligned}$$

Using $\zeta(x,y) := d(x,y)^2$ in (4.27) then yields

$$\mathcal{W}^2(\rho_0,\tilde{\rho}_0) \leqslant \delta\,\mathrm{diam}(\mathbb{M}^n)^2.$$

The same argument applies to $\mathcal{W}^2(\rho_1,\tilde{\rho}_1)$, thereby finishing the proof. ∎

5. Approximation of currents. In this section we prove the approximation result for currents used in the proof of Proposition 4.3. Notice that the regularization of currents is well understood; see, e.g., [15, 9]. Here we need to adopt the standard arguments somewhat in order to obtain convergence of the action functional (4.20).

PROPOSITION 5.1. *Let $\mathbb{M}^n$ be a compact connected Riemannian manifold without boundary. For given measures $\rho_0,\rho_1 \in \mathrm{Prob}(\mathbb{M}^n)$ consider a current T on $[0,1]\times\mathbb{M}^n$ with $\mathcal{A}(T)<\infty$ which satisfies the admissibility condition*

$$\Big\langle T,\ \partial_s\zeta\,ds + \nabla\zeta\cdot dx \Big\rangle = \int_{\mathbb{M}^n} \zeta(1,x)\,d\rho_1(x) - \int_{\mathbb{M}^n}\zeta(0,x)\,d\rho_0(x) \tag{5.1}$$

for all test functions ζ defined on $[0,1]\times\mathbb{M}^n$. Then we have the following:

(1) *There exists a family of currents* $\{T_\varepsilon\}_{\varepsilon>0}$ *representable in the form*

$$\Big\langle T_\varepsilon,\ \sigma\, ds + \xi \cdot dx \Big\rangle = \iint_{[0,1]\times\mathbb{M}^n} \rho_\varepsilon \sigma + m_\varepsilon \cdot \xi \, dx\, ds \tag{5.2}$$

for suitable vector fields $(\rho_\varepsilon \geqslant 0, m_\varepsilon)$ *which are smooth inside* $(0,1)\times\mathbb{M}^n$*. The admissibility condition* (5.1) *still holds with* T_ε *in place of* T *and*

$$\limsup_{\varepsilon\to 0} \mathcal{A}(T_\varepsilon) \leqslant \mathcal{A}(T).$$

(2) *If* ρ_0 *and* ρ_1 *are smooth functions, we may assume that the fields* $(\rho_\varepsilon, m_\varepsilon)$ *are smooth up to the boundary, and thus* $\rho_\varepsilon(0,\cdot) = \rho_0$ *and* $\rho_\varepsilon(1,\cdot) = \rho_1$*.*

Remark 5.1. Since the action functional $\mathcal{A}(T)$ is lower semicontinuous in the usual weak* topology of currents (see [9]), it even holds that $\lim_{\varepsilon\to 0}\mathcal{A}(T_\varepsilon) = \mathcal{A}(T)$.

Proof. We start with a remark on notation. Because of the action and the admissibility condition, the s- and x-variables have to be treated differently. However, it will often be convenient to lump s- and x-variables together; therefore we will write $\mathbf{x} = (s,x)$, $\quad = (\sigma,\xi)$, and $\mathbf{m} = (\rho, m)$. As a rule, bold symbols always denote $(n+1)$-dimensional objects (vector fields, parameters, operators, sets).

The approximating currents T_ε are obtained by regularization of T. We proceed as usual (see [15, 9]): Since a current is a linear form on 1-forms, we regularize T by duality, i.e., by constructing a linear operator that regularizes 1-forms $\ \cdot d\mathbf{x}$. This must be done in such a way that exact 1-forms $\ \cdot d\mathbf{x} = \boldsymbol{\nabla}\zeta \cdot d\mathbf{x}$ turn into exact 1-forms since, by assumption (5.1), T vanishes on exact 1-forms that are compactly supported in $(0,1)\times\mathbb{M}^n$. Recall that pulling back a 1-form under a smooth map preserves exactness. Therefore we regularize $\ \cdot d\mathbf{x}$ as follows: We construct a family of diffeomorphisms $\{\boldsymbol{\Phi}(\mathbf{z},\cdot)\}_{\mathbf{z}}$ of $\mathbb{R}\times\mathbb{M}^n$, parametrized by $\mathbf{z}\in\mathbb{R}\times\mathbb{R}^n$, and then consider its pull-back $\boldsymbol{\Phi}(\mathbf{z},\cdot)^{\#}(\ \cdot d\mathbf{x})$ and average over $\mathbf{z}$.

In order to preserve the boundary condition (5.1), it is necessary that $\boldsymbol{\Phi}(\mathbf{z},\cdot)$ leaves the complement of $(0,1)\times\mathbb{M}^n$ invariant. On the other hand, in order to achieve the regularizing effect, it is important that $\mathbf{z}\in\mathbb{R}\times\mathbb{R}^n$ "acts transitively" on $(0,1)\times\mathbb{M}^n$. Because of topological reasons, this cannot be achieved globally by a single map $\boldsymbol{\Phi}$ in general. We have to work locally with several maps $\boldsymbol{\Phi}$, each of which is attached to some open set U of a suitable covering of $\mathbb{M}^n$.

More precisely, we consider a finite covering $\{U_i\}_{i=1}^N$ of $\mathbb{M}^n$ subordinate to some atlas, with $U_i\subset\mathbb{M}^n$ homeomorphic to the unit ball $B_1(0)\subset\mathbb{R}^n$ and $\phi_i\colon U_i \longrightarrow \mathbb{R}^n$ the corresponding smooth coordinate map. We may assume that each ϕ_i extends to a neighborhood of U_i and $B_1(0)$. Based on this map, we shall construct an operator $T\mapsto T^{U_i}_{\varepsilon_i}$ for $\varepsilon_i>0$ with the following properties:

(a) The operator $T\mapsto T^{U_i}_{\varepsilon_i}$ regularizes in $(0,1)\times U_i$; i.e., $T^{U_i}_{\varepsilon_i}$ is representable in $(0,1)\times U_i$ by a smooth vector field $(\rho_{\varepsilon_i}\geqslant 0, m_{\varepsilon_i})$ as in (5.2).

(b) The operator $T\mapsto T^{U_i}_{\varepsilon_i}$ does not destroy smoothness; i.e., if T is representable by a smooth vector field in $(0,1)\times V$ with $V\subset\mathbb{M}^n$ open, then also $T^{U_i}_{\varepsilon_i}$ is representable by a smooth vector field in $(0,1)\times V$.

(c) The new current is admissible in the sense that (5.1) still holds with $T^{U_i}_{\varepsilon_i}$ in place of T, and we have upper semicontinuity of the action

$$\limsup_{\varepsilon_i\to 0}\mathcal{A}\big(T^{U_i}_{\varepsilon_i}\big) \leqslant \mathcal{A}(T).$$

Then the composition

$$T \;\mapsto\; T^{U_1}_{\varepsilon_1} \;\mapsto\; \big(T^{U_1}_{\varepsilon_1}\big)^{U_2}_{\varepsilon_2} \;\mapsto\; \cdots \;\mapsto\; \big(\cdots\big(T^{U_1}_{\varepsilon_1}\big)^{U_2}_{\varepsilon_2}\cdots\big)^{U_N}_{\varepsilon_N}$$

yields an approximate current T_ε with all the properties required by the proposition. In particular, we obtain a vector field $(\rho_\varepsilon \geqslant 0, m_\varepsilon)$ that represents T_ε in the sense of (5.2) and is smooth throughout $(0,1)\times\mathbb{M}^n$. It suffices to consider each operator $T \mapsto T^{U_i}_{\varepsilon_i}$ separately and check that (a)–(c) are satisfied. To simplify the notation, we will suppress the index i and do not indicate the dependence on U_i. The idea of regularizing a current defined on a manifold by composing several operators attached to local coordinate maps already appeared in [15].

We proceed in eight steps.

Step 1. As mentioned before, the regularization $T \mapsto T_\varepsilon$ is based on a family of diffeomorphisms $\{\mathbf{\Phi}(\mathbf{z},\cdot)\}_{\mathbf{z}}$ of $\mathbb{R}\times\mathbb{M}^n$, parametrized by $\mathbf{z}\in\mathbb{R}\times\mathbb{R}^n$ and attached to the open set $U\subset\mathbb{M}^n$. We would like these diffeomorphisms to leave the complement of $(0,1)\times U$ invariant, but we cannot impose this since in order to control the action we need that the first component of $\mathbf{\Phi}$ does not depend on x.

We shall use

- a smooth map $\mathbf{\Phi} = (\Phi_0,\Phi)\colon (\mathbb{R}\times\mathbb{R}^n)\times(\mathbb{R}\times\mathbb{M}^n) \longrightarrow \mathbb{R}\times\mathbb{M}^n$ (whose construction is postponed until Step 7) with the following properties:

$$\forall\, \mathbf{z}\in\mathbb{R}\times\mathbb{R}^n \quad \mathbf{\Phi}(\mathbf{z},\cdot) \text{ is a diffeomorphism of } \mathbb{R}\times\mathbb{M}^n \text{ onto } \mathbb{R}\times\mathbb{M}^n, \tag{5.3a}$$

$$\forall\, \mathbf{x}\in\mathbb{R}\times\mathbb{M}^n \quad \mathbf{\Phi}(0,\mathbf{x}) = \mathbf{x}, \tag{5.3b}$$

$$\forall\, \mathbf{x}\in(0,1)\times U \quad \mathbf{\Phi}(\cdot,\mathbf{x}) \text{ is a diffeomorphism of } \mathbb{R}\times\mathbb{R}^n \text{ onto } (0,1)\times U, \tag{5.3c}$$

$$\Phi_0 \text{ does not depend on } x, \tag{5.3d}$$

$$\forall\, (\mathbf{z},\mathbf{x})\in(\mathbb{R}\times\mathbb{R}^n)\times\big((0,1)\times(\mathbb{M}^n - U)\big) \quad \Phi(\mathbf{z},\mathbf{x}) = x, \tag{5.3e}$$

$$\forall\, (\mathbf{z},\mathbf{x})\in(\mathbb{R}\times\mathbb{R}^n)\times\big(\big(\mathbb{R}-(0,1)\big)\times\mathbb{M}^n\big) \quad \mathbf{\Phi}(\mathbf{z},\mathbf{x}) = \mathbf{x}. \tag{5.3f}$$

We shall also need the following maps which exist by (5.3a) and (5.3c):

- the right inverse $\mathbf{\Theta} = (\Theta_0,\Theta)\colon (\mathbb{R}\times\mathbb{R}^n)\times(\mathbb{R}\times\mathbb{M}^n) \longrightarrow \mathbb{R}\times\mathbb{M}^n$ of $\mathbf{\Phi}$ which is characterized by

$$\forall\, (\mathbf{z},\mathbf{x})\in(\mathbb{R}\times\mathbb{R}^n)\times(\mathbb{R}\times\mathbb{M}^n) \quad \mathbf{\Phi}\big(\mathbf{z},\mathbf{\Theta}(\mathbf{z},\mathbf{x})\big) = \mathbf{x}; \tag{5.4}$$

- the left inverse $\mathbf{\Psi} = (\Psi_0,\Psi)\colon \big((0,1)\times U\big)\times\big((0,1)\times U\big) \longrightarrow \mathbb{R}\times\mathbb{R}^n$ of $\mathbf{\Phi}$ which is characterized by

$$\forall\, (\mathbf{x},\mathbf{y})\in\big((0,1)\times U\big)\times\big((0,1)\times U\big) \quad \mathbf{\Phi}\big(\mathbf{\Psi}(\mathbf{y},\mathbf{x}),\mathbf{x}\big) = \mathbf{y}. \tag{5.5}$$

For later reference we collect some properties: Let $D_1\mathbf{\Phi}$ and $D_2\mathbf{\Phi}$ denote the derivatives of $\mathbf{\Phi}$ with respect to the first (resp., second) variable. Then

$$\forall\, (\mathbf{z},\mathbf{x})\in(\mathbb{R}\times\mathbb{R}^n)\times(\mathbb{R}\times\mathbb{M}^n) \quad D_2\mathbf{\Phi}(\mathbf{z},\mathbf{x}) \text{ has full rank}, \tag{5.6}$$

$$\forall\, (\mathbf{z},\mathbf{x})\in(\mathbb{R}\times\mathbb{R}^n)\times\big((0,1)\times U\big) \quad D_1\mathbf{\Phi}(\mathbf{z},\mathbf{x}) \text{ has full rank}, \tag{5.7}$$

as a consequence of (5.3a) and (5.3c). From (5.4) we obtain

$$\begin{gathered} D_{\mathbf{z}}\mathbf{\Theta}(\mathbf{z},\mathbf{x}) = -\Big(D_2\mathbf{\Phi}\big(\mathbf{z},\mathbf{\Theta}(\mathbf{z},\mathbf{x})\big)\Big)^{-1} D_1\mathbf{\Phi}\big(\mathbf{z},\mathbf{\Theta}(\mathbf{z},\mathbf{x})\big), \\ D_{\mathbf{x}}\mathbf{\Theta}(\mathbf{z},\mathbf{x}) = \Big(D_2\mathbf{\Phi}\big(\mathbf{z},\mathbf{\Theta}(\mathbf{z},\mathbf{x})\big)\Big)^{-1}, \end{gathered} \tag{5.8}$$

which together with (5.6) implies that $\mathbf{\Theta}$ is smooth. Similarly, (5.5) yields

$$\begin{aligned} D_{\mathbf{y}}\mathbf{\Psi}(\mathbf{y},\mathbf{x}) &= \Big(D_1\mathbf{\Phi}\big(\mathbf{\Psi}(\mathbf{y},\mathbf{x}),\mathbf{x}\big)\Big)^{-1}, \\ D_{\mathbf{x}}\mathbf{\Psi}(\mathbf{y},\mathbf{x}) &= -\Big(D_1\mathbf{\Phi}\big(\mathbf{\Psi}(\mathbf{y},\mathbf{x}),\mathbf{x}\big)\Big)^{-1} D_2\mathbf{\Phi}\big(\mathbf{\Psi}(\mathbf{y},\mathbf{x}),\mathbf{x}\big), \end{aligned} \tag{5.9}$$

so $\mathbf{\Psi}$ is smooth by (5.7). Moreover, we gather from (5.3b) that

$$D_2\mathbf{\Phi}(\mathbf{z},\cdot) = \mathrm{Id} + \mathcal{O}(|\mathbf{z}|) \quad \text{as } |\mathbf{z}| \to 0. \tag{5.10}$$

Finally, we notice that the properties above entail that

$$\forall\, \mathbf{y} \in (0,1)\times U \qquad \lim_{\substack{\mathbf{x}\in(0,1)\times U \\ \mathbf{x}\to\partial((0,1)\times U)}} |\mathbf{\Psi}(\mathbf{y},\mathbf{x})| = +\infty. \tag{5.11}$$

We argue by contradiction. Indeed, suppose that (5.11) fails. Then there exist a sequence $\{\mathbf{x}_\nu\}_\nu \subset (0,1)\times U$ and $\mathbf{x} \in \partial\big((0,1)\times U\big)$, $\mathbf{z} \in \mathbb{R}\times\mathbb{R}^n$ with

$$\lim_{\nu\to\infty} \mathbf{x}_\nu = \mathbf{x} \quad \text{and} \quad \lim_{\nu\to\infty} \mathbf{\Psi}(\mathbf{y},\mathbf{x}_\nu) = \mathbf{z}.$$

Passing to the limit in (5.5) yields by continuity of $\mathbf{\Phi}$ that

$$\mathbf{\Phi}(\mathbf{z},\mathbf{x}) = \mathbf{y} \;\in\; (0,1)\times U. \tag{5.12}$$

Now recall that $\mathbf{x} \in \partial\big((0,1)\times U\big) = \big((0,1)\times\partial U\big) \cup \big(\{0,1\}\times\bar{U}\big)$. If $\mathbf{x} \in (0,1)\times\partial U$, then (5.12) contradicts (5.3e); if $\mathbf{x} \in \{0,1\}\times\bar{U}$, then (5.12) contradicts (5.3f).

We now introduce our T_ε. We select a smooth nonnegative function k on $\mathbb{R}\times\mathbb{R}^n$ with compact support in $B_1(0)$ and $\iint_{\mathbb{R}\times\mathbb{R}^n} k(\mathbf{z})\,d\mathbf{z} = 1$. For $\varepsilon > 0$, we denote by $k_\varepsilon(\mathbf{z}) = k(\mathbf{z}/\varepsilon)/\varepsilon^{n+1}$ the rescaled kernel. Given a smooth 1-form $\;\cdot\, d\mathbf{x}$ on $\mathbb{R}\times\mathbb{M}^n$ and $\mathbf{z} \in \mathbb{R}\times\mathbb{R}^n$ we consider its pull-back $\mathbf{\Phi}(\mathbf{z},\cdot)^{\#}(\;\cdot\, d\mathbf{x}) =: \;(\mathbf{z},\cdot)\cdot d\mathbf{x}$. Observe that in terms of the vector fields this means

$$\forall\, \mathbf{x} \in \mathbb{R}\times\mathbb{M}^n \qquad (\mathbf{z},\mathbf{x}) = \big(D_{\mathbf{x}}\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)^t \;\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big), \tag{5.13}$$

where A^t denotes the transpose of A with respect to the metric on $\mathbb{R}\times\mathbb{M}^n$. Then we define the smeared out 1-form $\;_\varepsilon \cdot d\mathbf{x}$ by averaging $\;(\mathbf{z},\cdot)\cdot d\mathbf{x}$ over $\mathbf{z}$ with respect to k_ε. On the level of the vector fields this means

$$\forall\, \mathbf{x} \in \mathbb{R}\times\mathbb{M}^n \qquad {}_\varepsilon(\mathbf{x}) = \iint_{\mathbb{R}\times\mathbb{R}^n} \;(\mathbf{z},\mathbf{x})\;\; k_\varepsilon(\mathbf{z})\,d\mathbf{z}. \tag{5.14}$$

Finally, we introduce T_ε by duality; i.e., for all 1-forms $\;\cdot\, d\mathbf{x}$ we put

$$\Big\langle T_\varepsilon, \;\cdot\, d\mathbf{x}\Big\rangle := \Big\langle T, \;_\varepsilon \cdot d\mathbf{x}\Big\rangle. \tag{5.15}$$

Step 2. We first argue that T_ε has a smooth representative in $(0,1)\times U$. In order to see this, we write (5.14) in the form of

$$ {}_{\varepsilon}(\mathbf{x}) = \iint_{(0,1)\times U} K_\varepsilon(\mathbf{y},\mathbf{x}) \quad (\mathbf{y})\, d\mathbf{y} \quad \text{for } \mathbf{x}\in(0,1)\times U. \tag{5.16}$$

Indeed, we shall see that (5.16) holds for the tensor field

$$ K_\varepsilon(\mathbf{y},\mathbf{x}) = \left\{ \begin{array}{ll} \Big(-\big(D_{\mathbf{y}}\mathbf{\Psi}(\mathbf{y},\mathbf{x})\big)^{-1} D_{\mathbf{x}}\mathbf{\Psi}(\mathbf{y},\mathbf{x})\Big)^t & k_\varepsilon\big(\mathbf{\Psi}(\mathbf{y},\mathbf{x})\big) \det D_{\mathbf{y}}\mathbf{\Psi}(\mathbf{y},\mathbf{x}) \\ & \text{for } \mathbf{x}\in(0,1)\times U \\ 0 & \text{otherwise} \end{array} \right\} $$

$$ \forall\ \mathbf{y}\in\big(0,1)\times U. \tag{5.17}$$

Notice that $K_\varepsilon(\mathbf{y},\mathbf{x})$ is an endomorphism from the tangent space $T_{\mathbf{y}}(\mathbb{R}\times\mathbb{M}^n)$ into $T_{\mathbf{x}}(\mathbb{R}\times\mathbb{M}^n)$ and that $K_\varepsilon(\mathbf{y},\mathbf{x})$ is smooth in $(\mathbf{y},\mathbf{x})\in\big((0,1)\times U\big)\times(\mathbb{R}\times\mathbb{M}^n)$. Indeed, if $\mathbf{y}$ varies in a compact subset of $(0,1)\times U$ and $\mathbf{x}\in(0,1)\times U$ is close to $\partial\big((0,1)\times U\big)$, we learn from (5.11) that $k_\varepsilon\big(\mathbf{\Psi}(\mathbf{y},\mathbf{x})\big)=0$ and thus $K_\varepsilon(\mathbf{y},\mathbf{x})=0$, because k_ε has bounded support. We check (5.16): For all $\mathbf{x}\in(0,1)\times U$

$$\begin{aligned}
{}_{\varepsilon}(\mathbf{x}) &\overset{(5.14)}{=} \iint_{\mathbb{R}\times\mathbb{R}^n} \quad (\mathbf{z},\mathbf{x}) \ \ k_\varepsilon(\mathbf{z})\, d\mathbf{z} \\
&\overset{(5.13)}{=} \iint_{\mathbb{R}\times\mathbb{R}^n} \big(D_{\mathbf{x}}\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)^t \ \ \big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big) \ \ k_\varepsilon(\mathbf{z})\, d\mathbf{z} \\
&\overset{(5.5)}{=} \iint_{(0,1)\times U} \Big(D_2\mathbf{\Phi}\big(\mathbf{\Psi}(\mathbf{y},\mathbf{x}),\mathbf{x}\big)\Big)^t \ \ (\mathbf{y}) \ \ k_\varepsilon\big(\mathbf{\Psi}(\mathbf{y},\mathbf{x})\big) \det D_{\mathbf{y}}\mathbf{\Psi}(\mathbf{y},\mathbf{x})\, d\mathbf{y} \\
&\overset{(5.9)}{=} \iint_{(0,1)\times U} \Big(-\big(D_{\mathbf{y}}\mathbf{\Psi}(\mathbf{y},\mathbf{x})\big)^{-1} D_{\mathbf{x}}\mathbf{\Psi}(\mathbf{y},\mathbf{x})\Big)^t \ \ (\mathbf{y}) \ \ k_\varepsilon\big(\mathbf{\Psi}(\mathbf{y},\mathbf{x})\big) \\
&\qquad\qquad \times \det D_{\mathbf{y}}\mathbf{\Psi}(\mathbf{y},\mathbf{x})\, d\mathbf{y} \\
&\overset{(5.17)}{=} \iint_{(0,1)\times U} K_\varepsilon(\mathbf{y},\mathbf{x}) \quad (\mathbf{y})\, d\mathbf{y}.
\end{aligned}$$

We now argue that in $(0,1)\times U$, T_ε is represented by $\mathbf{m}_\varepsilon$ defined through

$$ \mathbf{m}_\varepsilon(\mathbf{y})\cdot \quad := \Big\langle T,\ \big(K_\varepsilon(\mathbf{y},\cdot) \quad \big)\cdot d\mathbf{x}\Big\rangle \quad \text{for } \mathbf{y}\in(0,1)\times U, \quad \in T_{\mathbf{y}}(\mathbb{R}\times\mathbb{M}^n). \tag{5.18}$$

Since $K_\varepsilon(\mathbf{y},\mathbf{x})$ is smooth in $(\mathbf{y},\mathbf{x})\in\big((0,1)\times U\big)\times(\mathbb{R}\times\mathbb{M}^n)$, $\mathbf{m}_\varepsilon$ is smooth in $\mathbf{y}\in(0,1)\times U$. We check that $\mathbf{m}_\varepsilon$ is indeed the representative of T_ε in $(0,1)\times U$. Let be a smooth vector field compactly supported in $(0,1)\times U$: Then

$$\begin{aligned}
\Big\langle T_\varepsilon, \quad \cdot\, d\mathbf{x}\Big\rangle &\overset{(5.15)}{=} \Big\langle T, \quad {}_{\varepsilon}\cdot d\mathbf{x}\Big\rangle \\
&\overset{(5.16)}{=} \iint_{(0,1)\times U} \Big\langle T,\ \big(K_\varepsilon(\mathbf{y},\cdot) \quad (\mathbf{y})\big)\cdot d\mathbf{x}\Big\rangle\, d\mathbf{y} \\
&\overset{(5.18)}{=} \iint_{(0,1)\times U} \mathbf{m}_\varepsilon(\mathbf{y})\cdot \quad (\mathbf{y})\, d\mathbf{y}.
\end{aligned}$$

Step 3. We now prove that the operator $T \mapsto T_\varepsilon$ does not destroy smoothness. More precisely, we shall argue that for $V \subset \mathbb{M}^n$ open

$$\begin{aligned} &T \text{ has a smooth representative in } (0,1) \times V \\ &\quad \Longrightarrow \quad T_\varepsilon \text{ has a smooth representative in } (0,1) \times (U \cup V) \end{aligned} \tag{5.19}$$

and that

$$\begin{aligned} &T \text{ has a smooth representative in a neighborhood of } \{0,1\} \times \mathbb{M}^n \\ &\quad \Longrightarrow \quad T_\varepsilon \text{ has the same property.} \end{aligned} \tag{5.20}$$

To treat both situations simultaneously, we consider a set $\mathbf{V}$ that is relatively open in $[0,1] \times \mathbb{M}^n$ and in which T is represented by a smooth vector field $\mathbf{m}$ in the sense that for all smooth vector fields compactly supported in $\mathbf{V}$:

$$\Big\langle T, \quad \cdot\, d\mathbf{x} \Big\rangle = \iint_{\mathbb{R}\times\mathbb{M}^n} \mathbf{m} \cdot \quad d\mathbf{x}. \tag{5.21}$$

Then we claim that in the set

$$\mathbf{V}_\varepsilon := \bigcap_{\mathbf{z} \in \overline{B_\varepsilon(0)}} \mathbf{\Phi}(\mathbf{z}, \mathbf{V}) \tag{5.22}$$

the regularized current T_ε is represented by

(5.23)

$$\mathbf{m}_\varepsilon(\mathbf{x}) := \left\{ \begin{array}{ll} \displaystyle\iint_{\mathbb{R}\times\mathbb{R}^n} & \big(D_\mathbf{x}\mathbf{\Theta}(\mathbf{z},\mathbf{x})\big)^{-1} \mathbf{m}\big(\mathbf{\Theta}(\mathbf{z},\mathbf{x})\big) \;\; k_\varepsilon(\mathbf{z}) \, \det D_\mathbf{x}\mathbf{\Theta}(\mathbf{z},\mathbf{x}) \, d\mathbf{z} \\ & \qquad \text{for } \mathbf{x} \in \mathbf{V}_\varepsilon \cap \big((0,1) \times \mathbb{M}^n\big) \\ \mathbf{m}(\mathbf{x}) & \qquad \text{for } \mathbf{x} \in \mathbf{V}_\varepsilon \cap \big(\{0,1\} \times \mathbb{M}^n\big) \end{array} \right\}.$$

Notice first that $\mathbf{V}_\varepsilon$ is relatively open in $[0,1] \times \mathbb{M}^n$ since (5.3a) and (5.3f) give

$$\mathbf{V}_\varepsilon = \big([0,1] \times \mathbb{M}^n\big) - \mathbf{\Phi}\Big(\overline{B_\varepsilon(0)},\ \big([0,1] \times \mathbb{M}^n\big) - \mathbf{V}\Big).$$

According to (5.21), (5.22), and (5.4), the vector field $\mathbf{m}_\varepsilon$ is well defined. Moreover, $\mathbf{m}_\varepsilon$ inherits the smoothness of $\mathbf{m}$ separately in both subsets of $\mathbf{V}_\varepsilon$. Hence we only need to check that $\mathbf{m}_\varepsilon$ is regular throughout $\mathbf{V}_\varepsilon$. By smoothness of $\mathbf{\Theta}$ and (5.3f), the function $(s,x) \mapsto \mathbf{\Theta}\big(\mathbf{z},(s,x)\big)$ approaches the identity map as $s \to \{0,1\}$, uniformly in all derivatives and in both $\mathbf{z} \in \overline{B_\varepsilon(0)}$ and $x \in \mathbb{M}^n$. This implies in particular that $D_\mathbf{x}\mathbf{\Theta} \to \mathrm{Id}$ and $\det D_\mathbf{x}\mathbf{\Theta} \to 1$. Since, by assumption, $\mathbf{m}$ is smooth in $\mathbf{V}$, regularity of $\mathbf{m}_\varepsilon$ then follows easily by standard arguments. Therefore the operator $T \mapsto T_\varepsilon$ does not destroy smoothness in the above sense. We now check that $\mathbf{m}_\varepsilon$ is indeed the representative. Let a smooth vector field be given that is compactly supported in $\mathbf{V}_\varepsilon$. Because of (5.22), $(\mathbf{z}, \cdot)$ defined in (5.13) is compactly supported in $\mathbf{V}$ for all

$\mathbf{z} \in \overline{B_\varepsilon(0)}$, and so $\quad_\varepsilon$ is compactly supported in $\mathbf{V}$ by definition (5.14). We obtain

$$
\begin{aligned}
&\Big\langle T_\varepsilon, \quad \cdot\, d\mathbf{x} \Big\rangle \\
&\quad \overset{(5.15)}{=} \Big\langle T, \quad_\varepsilon \cdot d\mathbf{x} \Big\rangle \\
&\quad \overset{(5.21)}{=} \iint_{\mathbb{R}\times\mathbb{M}^n} \mathbf{m} \cdot \quad_\varepsilon \, d\mathbf{x} \\
&\quad \overset{(5.14)}{=} \iint_{\mathbb{R}\times\mathbb{R}^n} \iint_{\mathbb{R}\times\mathbb{M}^n} \mathbf{m}(\mathbf{x}) \cdot \quad (\mathbf{z},\mathbf{x}) \, d\mathbf{x} \;\; k_\varepsilon(\mathbf{z}) \, d\mathbf{z} \\
&\quad \overset{(5.13)}{=} \iint_{\mathbb{R}\times\mathbb{R}^n} \iint_{\mathbb{R}\times\mathbb{M}^n} D_\mathbf{x}\boldsymbol{\Phi}(\mathbf{z},\mathbf{x}) \, \mathbf{m}(\mathbf{x}) \cdot \quad \big(\boldsymbol{\Phi}(\mathbf{z},\mathbf{x})\big) \, d\mathbf{x} \;\; k_\varepsilon(\mathbf{z}) \, d\mathbf{z} \\
&\quad \overset{(5.4)}{=} \iint_{\mathbb{R}\times\mathbb{R}^n} d\mathbf{z} \; k_\varepsilon(\mathbf{z}) \\
&\qquad\qquad \times \iint_{\mathbb{R}\times\mathbb{M}^n} D_\mathbf{x}\boldsymbol{\Phi}\big(\mathbf{z},\boldsymbol{\Theta}(\mathbf{z},\mathbf{y})\big) \, \mathbf{m}\big(\boldsymbol{\Theta}(\mathbf{z},\mathbf{y})\big) \cdot \quad (\mathbf{y}) \;\; \det D_\mathbf{y}\boldsymbol{\Theta}(\mathbf{z},\mathbf{y}) \, d\mathbf{y} \\
&\quad \overset{(5.8)}{=} \iint_{\mathbb{R}\times\mathbb{R}^n} d\mathbf{z} \; k_\varepsilon(\mathbf{z}) \\
&\qquad\qquad \times \iint_{\mathbb{R}\times\mathbb{M}^n} \big(D_\mathbf{y}\boldsymbol{\Theta}(\mathbf{z},\mathbf{y})\big)^{-1} \mathbf{m}\big(\boldsymbol{\Theta}(\mathbf{z},\mathbf{y})\big) \cdot \quad (\mathbf{y}) \;\; \det D_\mathbf{y}\boldsymbol{\Theta}(\mathbf{z},\mathbf{y}) \, d\mathbf{y} \\
&\quad \overset{(5.23)}{=} \iint_{\mathbf{V}_\varepsilon} \mathbf{m}_\varepsilon(\mathbf{y}) \cdot \quad (\mathbf{y}) \, d\mathbf{y}.
\end{aligned}
$$

If now $\mathbf{V} = (0,1) \times V$ with $V \subset \mathbb{M}^n$ open, then (5.3d)–(5.3f) entail

$$
\mathbf{V}_\varepsilon - \big((0,1)\times U\big) = \mathbf{V} - \big((0,1)\times U\big).
$$

In particular, $\mathbf{V}_\varepsilon$ contains an open neighborhood of $\mathbf{V} \cap \big((0,1)\times \partial U\big)$. Therefore T_ε is smooth in $\mathbf{V} \cup \big((0,1)\times U\big)$. This establishes (5.19). Similarly, if $\mathbf{V} = [0,\alpha) \times \mathbb{M}^n$ for some $0 < \alpha < 1$, then $\mathbf{V}_\varepsilon = [0,\alpha') \times \mathbb{M}^n$ for some $0 < \alpha' \leqslant \alpha$ by (5.3d) and (5.3f). Therefore T_ε is smooth up to the boundary $\{0\} \times \mathbb{M}^n$. The same argument applies to $\mathbf{V} = (1-\alpha, 1] \times \mathbb{M}^n$. This establishes (5.20).

Step 4. We now argue that T_ε is admissible if T is, i.e., if

$$
\Big\langle T, \; \boldsymbol{\nabla}\zeta \cdot d\mathbf{x} \Big\rangle = \int_{\mathbb{M}^n} \zeta(1,x) \, d\rho_1(x) - \int_{\mathbb{M}^n} \zeta(0,x) \, d\rho_0(x) \tag{5.24}
$$

for all smooth functions ζ on $\mathbb{R}\times\mathbb{M}^n$. Consider the gradient field $\quad := \boldsymbol{\nabla}\zeta$. We gather from (5.13) and the chain rule that

$$
\quad (\mathbf{z},\mathbf{x}) = \boldsymbol{\nabla}_\mathbf{x} \zeta(\mathbf{z},\mathbf{x}), \quad \text{where} \quad \zeta(\mathbf{z},\mathbf{x}) := \zeta\big(\boldsymbol{\Phi}(\mathbf{z},\mathbf{x})\big).
$$

We thus infer from (5.14) that

$$
\quad_\varepsilon(\mathbf{x}) = \boldsymbol{\nabla}\zeta_\varepsilon(\mathbf{x}), \quad \text{where} \quad \zeta_\varepsilon(\mathbf{x}) := \iint_{\mathbb{R}\times\mathbb{R}^n} \zeta(\mathbf{z},\mathbf{x}) \;\; k_\varepsilon(\mathbf{z}) \, d\mathbf{z}. \tag{5.25}
$$

Then (5.3f) implies that

$$
\forall\, (\mathbf{z},\mathbf{x}) \in (\mathbb{R}\times\mathbb{R}^n) \times \big(\big(\mathbb{R} - (0,1)\big) \times \mathbb{M}^n\big) \quad \left\{ \begin{array}{c} \zeta(\mathbf{z},\mathbf{x}) = \zeta(\mathbf{x}) \\ \text{and thus} \\ \zeta_\varepsilon(\mathbf{x}) = \zeta(\mathbf{x}) \end{array} \right\}. \tag{5.26}
$$

Hence we obtain as desired

$$\begin{aligned}
\Big\langle T_\varepsilon,\ \boldsymbol{\nabla}\zeta\cdot d\mathbf{x}\Big\rangle &\overset{\substack{(5.15)\\(5.25)}}{=} \Big\langle T,\ \boldsymbol{\nabla}\zeta_\varepsilon\cdot d\mathbf{x}\Big\rangle\\
&\overset{(5.24)}{=} \int_{\mathbb{M}^n}\zeta_\varepsilon(1,x)\,d\rho_1(x) - \int_{\mathbb{M}^n}\zeta_\varepsilon(0,x)\,d\rho_0(x)\\
&\overset{(5.26)}{=} \int_{\mathbb{M}^n}\zeta(1,x)\,d\rho_1(x) - \int_{\mathbb{M}^n}\zeta(0,x)\,d\rho_0(x).
\end{aligned}$$

Step 5. Now we address the action estimate. We claim that for small ε

$$\mathcal{A}(T_\varepsilon) \leqslant \big(1+\mathcal{O}(\varepsilon)\big)\,\mathcal{A}(T)+\mathcal{O}(\varepsilon) \tag{5.27}$$

with the modulus $\mathcal{O}(\varepsilon)$ depending only on $\mathbf{\Phi}$. Let $\quad = (\sigma,\xi)$ be an admissible vector field in the definition of $\mathcal{A}$, i.e., for which

$$\sigma+\tfrac{1}{2}|\xi|^2 \leqslant 0. \tag{5.28}$$

Consider $\quad(\mathbf{z},\mathbf{x}) = \big(\sigma(\mathbf{z},\mathbf{x}),\xi(\mathbf{z},\mathbf{x})\big)$ defined in (5.13). We will then show that the modified vector field

$$\begin{gathered}
\Big(\lambda(\mathbf{z})\,\big(\sigma(\mathbf{z},\cdot)-\mu(\mathbf{z})\big),\ \lambda(\mathbf{z})\,\xi(\mathbf{z},\cdot)\Big)\ \text{ is admissible}\\
\text{for suitable constants } \lambda(\mathbf{z}) = 1-\mathcal{O}(|\mathbf{z}|),\ \mu(\mathbf{z}) = O(|\mathbf{z}|).
\end{gathered} \tag{5.29}$$

Indeed, the anisotropy condition (5.3d) on $\mathbf{\Phi} = (\Phi_0,\Phi)$ and (5.13) give

$$\begin{aligned}
\sigma(\mathbf{z},\mathbf{x}) &= \partial_s\Phi_0(\mathbf{z},s)\,\sigma\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)+\partial_s\Phi(\mathbf{z},\mathbf{x})\cdot\xi\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big),\\
\xi(\mathbf{z},\mathbf{x}) &= \big(D_x\Phi(\mathbf{z},\mathbf{x})\big)^t\,\xi\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big).
\end{aligned}$$

Because of (5.10) this yields the estimates

$$\begin{aligned}
\sigma(\mathbf{z},\mathbf{x}) &\leqslant \big(1-\mathcal{O}(|\mathbf{z}|)\big)\,\sigma\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)+\mathcal{O}(|\mathbf{z}|)\,\big|\xi\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)\big|,\\
|\xi(\mathbf{z},\mathbf{x})| &\leqslant \big(1+\mathcal{O}(|\mathbf{z}|)\big)\,\big|\xi\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)\big|.
\end{aligned} \tag{5.30}$$

Using Young's inequality

$$2\lambda(\mathbf{z})\,\big|\xi\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)\big| \leqslant 1+\lambda(\mathbf{z})^2\,\big|\xi\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)\big|^2, \tag{5.31}$$

we notice that the latter implies

$$\begin{aligned}
&\lambda(\mathbf{z})\,\big(\sigma(\mathbf{z},\mathbf{x})-\mu(\mathbf{z})\big)+\tfrac{1}{2}|\lambda(\mathbf{z})\,\xi(\mathbf{z},\mathbf{x})|^2\\
&\quad\overset{(5.30)}{\leqslant} \lambda(\mathbf{z})\,\Big(\big(1-\mathcal{O}(|\mathbf{z}|)\big)\,\sigma\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)+\mathcal{O}(|\mathbf{z}|)\,\big|\xi\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)\big|-\mu(\mathbf{z})\Big)\\
&\qquad\quad + \lambda(\mathbf{z})^2\,\big(1+\mathcal{O}(|\mathbf{z}|)\big)^2\,\tfrac{1}{2}\big|\xi\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)\big|^2\\
&\quad\overset{(5.31)}{\leqslant} \lambda(\mathbf{z})\,\big(1-\mathcal{O}(|\mathbf{z}|)\big)\,\sigma\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)+\lambda(\mathbf{z})^2\,\big(1+\mathcal{O}(|\mathbf{z}|)\big)\,\tfrac{1}{2}\big|\xi\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)\big|^2\\
&\qquad\quad + \mathcal{O}(|\mathbf{z}|)-\lambda(\mathbf{z})\,\mu(\mathbf{z})\\
&\quad\overset{(5.28)}{\leqslant} \lambda(\mathbf{z})\,\Big(-\big(1-\mathcal{O}(|\mathbf{z}|)\big)+\big(1+\mathcal{O}(|\mathbf{z}|)\big)\,\lambda(\mathbf{z})\Big)\,\tfrac{1}{2}\big|\xi\big(\mathbf{\Phi}(\mathbf{z},\mathbf{x})\big)\big|^2\\
&\qquad\quad + \mathcal{O}(|\mathbf{z}|)-\lambda(\mathbf{z})\,\mu(\mathbf{z}),
\end{aligned}$$

which in turn yields (5.29).

Choosing $\zeta(s,x) = s$ in the admissibility condition (5.24) yields

$$\Big\langle T,\ 1\,ds \Big\rangle = 1. \tag{5.32}$$

Thus we have by definition of $\mathcal{A}(T)$

$$\begin{aligned}
&\Big\langle T,\ \sigma(\mathbf{z},\cdot)\,ds + \xi(\mathbf{z},\cdot)\cdot dx \Big\rangle \\
&\quad \overset{(5.29)}{=} \frac{1}{\lambda(\mathbf{z})} \Big\langle T,\ \lambda(\mathbf{z})\big(\sigma(\mathbf{z},\cdot) - \mu(\mathbf{z})\big)\,ds + \lambda(\mathbf{z})\,\xi(\mathbf{z},\cdot)\cdot dx \Big\rangle + \mu(\mathbf{z}) \Big\langle T,\ 1\,ds \Big\rangle \\
&\quad \overset{(5.32)}{\leqslant} \frac{1}{\lambda(\mathbf{z})}\,\mathcal{A}(T) + \mu(\mathbf{z}) \\
&\quad \overset{(5.29)}{=} \big(1 + \mathcal{O}(|\mathbf{z}|)\big)\,\mathcal{A}(T) + \mathcal{O}(|\mathbf{z}|).
\end{aligned} \tag{5.33}$$

We therefore obtain as desired

$$\begin{aligned}
\Big\langle T_\varepsilon,\ \sigma\,ds + \xi\cdot dx \Big\rangle &\overset{(5.15)}{=} \Big\langle T,\ \sigma_\varepsilon\,ds + \xi_\varepsilon\cdot dx \Big\rangle \\
&\overset{(5.14)}{=} \iint_{\mathbb{R}\times\mathbb{R}^n} \Big\langle T,\ \sigma(\mathbf{z},\cdot)\,ds + \xi(\mathbf{z},\cdot)\cdot dx \Big\rangle\ k_\varepsilon(\mathbf{z})\,d\mathbf{z} \\
&\overset{(5.33)}{=} \big(1 + \mathcal{O}(\varepsilon)\big)\mathcal{A}(T) + \mathcal{O}(\varepsilon).
\end{aligned}$$

Since (σ,ξ) was arbitrary with (5.28), this yields (5.27) by definition of $\mathcal{A}(T_\varepsilon)$.

Step 6. Let T_ε have a smooth representative $(\rho_\varepsilon, m_\varepsilon)$ in $(0,1)\times\mathbb{M}^n$ and satisfy $\mathcal{A}(T_\varepsilon) < \infty$. We now argue that $\rho_\varepsilon \geqslant 0$. More precisely, we shall show that

$$\Big\langle T_\varepsilon,\ \zeta\,ds \Big\rangle \geqslant 0 \quad \forall \text{ smooth test function } \zeta \text{ with } \zeta \geqslant 0. \tag{5.34}$$

Indeed, for $n \in \mathbb{N}$ the vector field $\;= (-n\zeta, 0)$ is admissible and yields

$$-n\Big\langle T_\varepsilon,\ \zeta\,ds \Big\rangle = \Big\langle T_\varepsilon,\ \ \cdot d\mathbf{x} \Big\rangle \leqslant \mathcal{A}(T_\varepsilon),$$

which gives (5.34) in the limit $n \to \infty$. By (5.2), this proves that $\rho_\varepsilon \geqslant 0$.

Step 7. It remains to construct the map $\mathbf{\Phi} = (\Phi_0, \Phi)$. This is done in a series of short steps. The starting point is the diffeomorphism $h_0\colon (0,1) \longrightarrow \mathbb{R}$ defined by

$$h_0(s) = \big(s - \tfrac{1}{2}\big)\,\exp\Big(\frac{1}{s(1-s)}\Big).$$

Next we introduce the map $\Phi_0\colon \mathbb{R}\times\mathbb{R} \longrightarrow \mathbb{R}$:

$$\Phi_0(u,s) = \left\{ \begin{array}{ll} h_0^{-1}\big(h_0(s) + u\big) & \text{for } s \in (0,1) \\ s & \text{otherwise} \end{array} \right\}. \tag{5.35}$$

The properties of the exponential function imply that Φ_0 is smooth. In particular, we have by the inverse function theorem that

$$\partial_u \Phi_0(u,s) = \left\{ \begin{array}{ll} \Big(h_0'\big(\Phi_0(u,s)\big)\Big)^{-1} & \text{for } s \in (0,1) \\ 0 & \text{otherwise} \end{array} \right\},$$

$$\partial_s \Phi_0(u,s) = \left\{ \begin{array}{ll} \Big(h_0'\big(\Phi_0(u,s)\big)\Big)^{-1} h_0'(s) & \text{for } s \in (0,1) \\ 1 & \text{otherwise} \end{array} \right\}.$$

The map Φ_0 obviously has the homomorphism property

$$\begin{aligned} \forall\, u, u' \in \mathbb{R},\ s \in \mathbb{R} \quad \Phi_0(u+u', s) &= \Phi_0\big(u, \Phi_0(u', s)\big), \\ \forall\, s \in \mathbb{R} \quad \Phi_0(0, s) &= s. \end{aligned} \tag{5.36}$$

In a similar way, we introduce the diffeomorphism $\tilde{h}\colon B_1(0) \to \mathbb{R}^n$ defined by

$$\tilde{h}(\tilde{x}) = \tilde{x}\, \exp\bigg(\frac{1}{1-|\tilde{x}|^2}\bigg)$$

and the map $\tilde{\Phi}\colon \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ through

$$\tilde{\Phi}(z, \tilde{x}) = \left\{ \begin{array}{ll} \tilde{h}^{-1}\big(\tilde{h}(\tilde{x}) + z\big) & \text{for } \tilde{x} \in B_1(0) \\ \tilde{x} & \text{otherwise} \end{array} \right\}. \tag{5.37}$$

Again, the properties of the exponential function imply that $\tilde{\Phi}$ is smooth, and

$$D_z\tilde{\Phi}(z, \tilde{x}) = \left\{ \begin{array}{ll} \Big(D\tilde{h}\big(\tilde{\Phi}(z, \tilde{x})\big)\Big)^{-1} & \text{for } \tilde{x} \in B_1(0) \\ 0 & \text{otherwise} \end{array} \right\},$$

$$D_{\tilde{x}}\tilde{\Phi}(z, \tilde{x}) = \left\{ \begin{array}{ll} \Big(D\tilde{h}\big(\tilde{\Phi}(z, \tilde{x})\big)\Big)^{-1} D\tilde{h}(\tilde{x}) & \text{for } \tilde{x} \in B_1(0) \\ \mathrm{Id} & \text{otherwise} \end{array} \right\}.$$

As for Φ_0 above, the map $\tilde{\Phi}$ has the homomorphism property

$$\begin{aligned} \forall\, z, z' \in \mathbb{R}^n,\ \tilde{x} \in \mathbb{R}^n \quad \tilde{\Phi}(z+z', \tilde{x}) &= \tilde{\Phi}\big(z, \tilde{\Phi}(z', \tilde{x})\big), \\ \forall\, \tilde{x} \in \mathbb{R}^n \quad \tilde{\Phi}(0, \tilde{x}) &= \tilde{x}. \end{aligned} \tag{5.38}$$

Recall that $U \subset \mathbb{M}^n$ is an open subset homeomorphic to the ball $B_1(0) \subset \mathbb{R}^n$, with $\phi\colon U \longrightarrow \mathbb{R}^n$ a coordinate map. We may assume that ϕ extends to some neighborhood of U and $B_1(0)$. Then the composition

$$h(x) := \tilde{h}\big(\phi(x)\big) \quad \forall\, x \in U$$

defines a diffeomorphism $h\colon U \longrightarrow \mathbb{R}^n$. In view of (5.37),

$$\Phi(z, x) = \left\{ \begin{array}{ll} \phi^{-1}\Big(\tilde{\Phi}\big(z, \phi(x)\big)\Big) & \text{for } x \in U \\ x & \text{otherwise} \end{array} \right\}$$

defines a smooth map $\Phi\colon \mathbb{R}^n \times \mathbb{M}^n \longrightarrow \mathbb{M}^n$, and it is immediately clear that the properties (5.37) and (5.38) are conserved; i.e., we have

$$\Phi(z, x) = \left\{ \begin{array}{ll} h^{-1}\big(h(x) + z\big) & \text{for } x \in U \\ x & \text{otherwise} \end{array} \right\} \tag{5.39}$$

and

$$\begin{aligned} \forall\, z, z' \in \mathbb{R}^n,\ x \in \mathbb{M}^n \quad \Phi(z+z', x) &= \Phi\big(z, \Phi(z', x)\big), \\ \forall\, x \in \mathbb{M}^n \quad \Phi(0, x) &= x. \end{aligned} \tag{5.40}$$

Finally, we consider the smooth function $\delta\colon \mathbb{R} \longrightarrow \mathbb{R}$ defined through

$$\delta(s) = \left\{ \begin{array}{ll} \exp\Big(-\dfrac{1}{s(1-s)}\Big) & \text{for } s \in (0,1) \\ 0 & \text{otherwise} \end{array} \right\}. \tag{5.41}$$

We now introduce the smooth map $\mathbf{\Phi}\colon (\mathbb{R}\times\mathbb{R}^n)\times(\mathbb{R}\times\mathbb{M}^n) \longrightarrow (\mathbb{R}\times\mathbb{M}^n)$:

$$\mathbf{\Phi}\big((u,z),(s,x)\big) := \Big(\Phi_0(u,s), \Phi\big(\delta(s)z,x\big)\Big).$$

Let us quickly check that $\mathbf{\Phi}$ has the required properties. We establish the diffeomorphism properties (5.3a) by explicitly giving the right inverse

$$\mathbf{\Theta}\big((u,z),(s,x)\big) = \Big(\Phi_0(-u,s), \Phi\big(-\delta\big(\Phi_0(-u,s)\big)z,x\big)\Big)$$

which defines a smooth map $\mathbf{\Theta}\colon (\mathbb{R}\times\mathbb{R}^n)\times(\mathbb{R}\times\mathbb{M}^n) \longrightarrow (\mathbb{R}\times\mathbb{M}^n)$. Then (5.36) and (5.40) imply (5.4) and (5.3b). The anisotropy property (5.3d) is clear by construction. The first invariance property (5.3e) can be read off from (5.39); the second invariance property (5.3f) follows from (5.35) for the Φ_0-component and from (5.41) combined with (5.40) for the Φ-component. The second diffeomorphism property (5.3c) follows from an explicit formula for the left inverse

$$\mathbf{\Psi}\big((t,y),(s,x)\big) = \left(h_0(t)-h_0(s), \frac{h(y)-h(x)}{\delta(s)}\right)$$

which defines a smooth map $\mathbf{\Psi}\colon \big((0,1)\times U\big)\times\big((0,1)\times U\big) \longrightarrow \mathbb{R}\times\mathbb{R}^n$ (recall that δ is positive on $(0,1)$). The identities (5.35) and (5.39) imply (5.5) as desired.

Step 8. In this last step, we prove that if ρ_0 and ρ_1 are smooth functions and if T is a current defined on $[0,1]\times\mathbb{M}^n$ satisfying the admissibility condition (5.1), then the regularization T_ε of T is representable by a vector field $(\rho_\varepsilon \geqslant 0, m_\varepsilon)$ that is smooth up to the boundary $\{0,1\}\times\mathbb{M}^n$. This can be achieved by first approximating T by an admissible current T_α that is representable by a smooth vector field in $[0,\alpha)\times\mathbb{M}^n$ and $(1-\alpha,1]\times\mathbb{M}^n$ for suitable α and then applying to T_α the regularization procedure described above. Since the operators $T\mapsto T^{U_i}_{\varepsilon_i}$ do not destroy smoothness in stripes around the boundary $\{0,1\}\times\mathbb{M}^n$ (see Step 3), the regularized current $(T_\alpha)_\varepsilon$ can be represented by a vector field that attains the data ρ_0 and ρ_1 smoothly as desired.

We proceed as follows: For $0<\alpha<\frac{1}{2}$ we consider the map

$$\mathbf{\Phi}_\alpha(s,x) := \big(\alpha+(1-2\alpha)s,x\big) \quad \text{for } (s,x)\in[0,1]\times\mathbb{M}^n.$$

Given a smooth 1-form $\cdot\, d\mathbf{x}$, let $\mathbf{\Phi}_\alpha^{\#}(\ \cdot\, d\mathbf{x}) =: {}_\alpha \cdot d\mathbf{x}$ be its pull-back under the map $\mathbf{\Phi}_\alpha$. In terms of the vector field $= (\sigma,\xi)$ this means

$$\forall\, \mathbf{x}\in[0,1]\times\mathbb{M}^n \qquad {}_\alpha(\mathbf{x}) = \big(D_{\mathbf{x}}\mathbf{\Phi}_\alpha(\mathbf{x})\big)^t\ \big(\mathbf{\Phi}_\alpha(\mathbf{x})\big), \tag{5.42}$$

and hence ${}_\alpha = (\sigma_\alpha,\xi_\alpha)$ with

$$\sigma_\alpha(\mathbf{x}) = (1-2\alpha)\,\sigma\big(\mathbf{\Phi}_\alpha(\mathbf{x})\big) \quad \text{and} \quad \xi_\alpha(\mathbf{x}) = \xi\big(\mathbf{\Phi}_\alpha(\mathbf{x})\big). \tag{5.43}$$

We define the approximate current T_α by duality as

$$\begin{aligned} \Big\langle T_\alpha,\ \cdot\, d\mathbf{x}\Big\rangle := \Big\langle T,\ {}_\alpha\cdot d\mathbf{x}\Big\rangle &+ \iint_{[0,\alpha]\times\mathbb{M}^n} \sigma(s,x)\,\rho_0(x)\,dx\,ds \\ &+ \iint_{[1-\alpha,1]\times\mathbb{M}^n} \sigma(s,x)\,\rho_1(x)\,dx\,ds \end{aligned} \tag{5.44}$$

for all smooth 1-forms $\cdot\, d\mathbf{x} = \sigma\, ds + \xi \cdot dx$. This T_α is admissible: If ζ is a smooth function and $:= \boldsymbol{\nabla}\zeta$ the gradient field, then (5.42) implies that

$$ {}_\alpha(\mathbf{x}) = \boldsymbol{\nabla}\zeta_\alpha(\mathbf{x}), \quad \text{where} \quad \zeta_\alpha(\mathbf{x}) := \zeta\big(\boldsymbol{\Phi}_\alpha(\mathbf{x})\big), $$

and therefore

$$ \begin{aligned} \Big\langle T_\alpha,\ \boldsymbol{\nabla}\zeta \cdot d\mathbf{x} \Big\rangle &= \int_{\mathbb{M}^n} \zeta(1-\alpha, x)\, \rho_1(x)\, dx + \iint_{[1-\alpha,1]\times\mathbb{M}^n} \partial_s\zeta(s,x)\, \rho_1(x)\, dx\, ds \\ &\quad - \int_{\mathbb{M}^n} \zeta(\alpha, x)\, \rho_0(x)\, dx + \iint_{[0,\alpha]\times\mathbb{M}^n} \partial_s\zeta(s,x)\, \rho_0(x)\, dx\, ds \\ &= \int_{\mathbb{M}^n} \zeta(1,x)\, \rho_1(x)\, dx - \int_{\mathbb{M}^n} \zeta(0,x)\, \rho_0(x)\, dx. \end{aligned} $$

It follows easily from (5.44) that T_α is represented by the smooth vector field $(\rho_0, 0)$ in the stripe $[0,\alpha)\times\mathbb{M}^n$ and by $(\rho_1, 0)$ in $(1-\alpha, 1]\times\mathbb{M}^n$. Consider the action functional. Notice first that if $= (\sigma, \xi)$ is admissible in the definition of $\mathcal{A}$, then also the modified vector field $(1-2\alpha)\ {}_\alpha$ is admissible since by (5.43)

$$ (1-2\alpha)\sigma_\alpha + \tfrac{1}{2}|(1-2\alpha)\xi_\alpha|^2 = (1-2\alpha)^2\Big(\sigma \circ \boldsymbol{\Phi}_\alpha + \tfrac{1}{2}|\xi \circ \boldsymbol{\Phi}_\alpha|^2\Big) \leqslant 0. $$

The extra integrals in (5.44) do not contribute to $\mathcal{A}(T_\alpha)$ because $\sigma \leqslant -\frac{1}{2}|\xi|^2 \leqslant 0$ and $\rho_0, \rho_1 \geqslant 0$. This yields as above $\mathcal{A}(T_\alpha) \leqslant (1-2\alpha)^{-1}\mathcal{A}(T)$, and thus

$$ \limsup_{\alpha\to 0} \mathcal{A}(T_\alpha) \leqslant \mathcal{A}(T). $$

Now we regularize as before to conclude. $\square$

REFERENCES

[1] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2005.

[2] C. Ané, S. Blachère, D. Chafaï, P. Fougères, I. Gentil, F. Malrieu, C. Roberto, and G. Scheffer, *Sur les inégalités de Sobolev logarithmiques*, Panor. Syntheses 10, Société Mathématique de France, Paris, 2000.

[3] D. Bakry and M. Émery, *Diffusions hypercontractives*, in Séminaire de probabilités XIX, Lecture Notes in Math. 1123, Springer, Berlin, 1985, pp. 177–206.

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

[5] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.

[6] J. A. Carrillo, R. J. McCann, and C. Villani, *Kinetic equilibration rates for granular media and related equations: Entropy dissipation and mass transportation estimates*, Rev. Mat. Iberoamericana, 19 (2003), pp. 971–1018.

[7] D. Cordero-Erausquin, R. J. McCann, and M. Schmuckenschläger, *A Riemannian interpolation inequality à la Borell, Brascamp and Lieb*, Invent. Math., 146 (2001), pp. 219–257.

[8] H. Federer, *Geometric Measure Theory*, Springer, Berlin, 1969.

[9] G. Giaquinta, A. Modica, and J. Souček, *Cartesian Currents in the Calculus of Variations*, Springer, Berlin, 1998.

[10] R. J. McCann, *A convexity principle for interacting gases*, Adv. Math., 128 (1997), pp. 153–179.

[11] R. J. McCann, *Polar factorization of maps on Riemannian manifolds*, Geom. Funct. Anal., 11 (2001), pp. 589–608.

[12] F. Otto, *The geometry of dissipative evolution equations: The porous medium equation*, Comm. Partial Differential Equations, 26 (2001), pp. 101–174.

[13] F. Otto and C. Villani, *Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality*, J. Funct. Anal., 173 (2000), pp. 361–400.
[14] G. Petersen, *Riemannian Geometry*, Springer, New York, 1998.
[15] G. de Rham, *Variétés différentiables*, Hermann et Cie, Paris, 1955.
[16] K.-Th. Sturm, *Convex functionals of probability measures and nonlinear diffusions on manifolds*, J. Math. Pures Appl. (9), 84 (2005), pp. 149–168.
[17] K.-Th. Sturm and M.-K. von Renesse, *Transport inequalities, gradient estimate, entropy and Ricci curvature*, Comm. Pure Appl. Math., 58 (2005), pp. 923–940
[18] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, AMS, Providence, RI, 2003.

SIAM J. MATH. ANAL.
Vol. 37, No. 4, pp. 1256–1298

ZERO-VISCOSITY LIMIT OF THE LINEARIZED COMPRESSIBLE NAVIER-STOKES EQUATIONS WITH HIGHLY OSCILLATORY FORCES IN THE HALF-PLANE*

YA-GUANG WANG† AND ZHOUPING XIN‡

Abstract. We study the asymptotic behavior of the solution to the linearized compressible Navier–Stokes equations with highly oscillatory forces in the half-plane with nonslip boundary conditions for small viscosity. The wavelength of oscillation is assumed to be proportional to the square root of the viscosity. By means of asymptotic analysis, we deduce that the leading profiles of the solution have four terms: the first one is the outflow satisfying the linearized Euler equations, the second one is an oscillatory wave propagated along the characteristic field tangential to the boundary associated with the linearized Euler operator in the half-plane, the third one is a boundary layer satisfying a linearized Prandtl equation, the fourth one represents the oscillation propagated in the boundary layer, and it is described by an initial-boundary value problem for an Poisson–Prandtl coupled system. By using the energy method and mode analysis, we obtain the well-posedness of this Poisson–Prandtl coupled problem, and a rigorous theory on the asymptotic analysis of the zero-viscosity limit. Finally, we have briefly discussed the case that the wavelength of the oscillatoy force is shorter than the square root of the viscosity.

Key words. linearized compressible Navier–Stokes equations, boundary layers, oscillatory waves

AMS subject classifications. 35Q30, 76N20, 35B05

DOI. 10.1137/040614967

1. Introduction.

Consider the following initial-boundary value problem for the two-dimensional isentropic compressible Navier–Stokes equations with nonslip boundary conditions in $\{t, x_1 > 0, x_2 \in \mathbb{R}\}$:

$$
\begin{cases}
\partial_t \rho + (v \cdot \nabla)\rho + \rho \nabla \cdot v = f(t,x), \\
\rho(\partial_t v + (v \cdot \nabla)v) + \nabla p = \nabla \cdot (2\mu P + \lambda I_2 \nabla \cdot v) + g(t,x), \\
v|_{x_1=0} = 0, \\
(\rho, v)|_{t=0} = (\rho_0, v_0)(x),
\end{cases}
\tag{1.1}
$$

where f and g represent the source and force terms, $P = \frac{1}{2}\{\partial_{x_j} v_i + \partial_{x_i} v_j\}_{i \times j}$ is a 2×2 matrix with $v = (v_1, v_2)^T$, $p = p(\rho)$ is the equation of state, μ and λ denote the coefficient and the second coefficient of viscosity, respectively, with $\mu > 0$ and $\mu' = \mu + \lambda \geq 0$. Corresponding to (1.1), the motion of an inviscid compressible fluid

*Received by the editors September 14, 2004; accepted for publication (in revised form) April 26, 2005; published electronically January 6, 2006. This research was supported by the Zheng Ge Ru Foundation when Ya-Guang Wang was visiting the Institute of Mathematical Sciences in the Chinese University of Hong Kong.

http://www.siam.org/journals/sima/37-4/61496.html

†Department of Mathematics, Shanghai Jiao Tong University, 200030 Shanghai (ygwang@sjtu.edu.cn). This research was partially supported in part by the NSFC grants 10131050 and 10531020, the Educational Ministry of China and the Shanghai Science and Technology Committee grant 03QMH1407.

‡The Institute of Mathematical Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong (zpxin@ims.cuhk.edu.hk). This research was supported in part by grants from RGC of HKSAR CUHK4279/00P, CUHK4040/02P, CUHK4181/03E and CUHK4028/04P.

is governed by the following Euler equations:

$$
\begin{cases}
\partial_t \rho + (v \cdot \nabla)\rho + \rho \nabla \cdot v = f(t,x), \\
\rho(\partial_t v + (v \cdot \nabla) v) + \nabla p = g(t,x), \\
v_1|_{x_1=0} = 0, \\
(\rho, v)|_{t=0} = (\rho_0, v_0)(x).
\end{cases} \tag{1.2}
$$

For simplicity, we assume that μ and μ' are proportional to a parameter, say ϵ^2 with $\epsilon > 0$.

One interesting problem is to study the convergence of the solution of the Navier–Stokes equations (1.1) to that of the Euler equations (1.2) in the limit of small viscosity. It is expected that uniform convergence can take place only away from the physical boundary $\{x_1 = 0\}$ even for smooth solutions of (1.2) due to the disparity of the boundary conditions in (1.1) and (1.2), and a thin region comes out near the boundary $\{x_1 = 0\}$ (the boundary layer) in which values of unknowns change drastically in the process of this limit.

It is a challenging problem to analyze rigorously this boundary layer phenomena displayed by the actual Navier–Stokes solutions. For the problem of incompressible Navier–Stokes equations, Prandtl carried out a formal analysis in his speech [6] at the International Congress of Mathematicians in 1904, and derived a nonlinear degenerate parabolic-elliptic coupled system for the velocity fields in the boundary layer, which is now called the Prandtl system. Under the monotonic assumption on the velocity of the outflow, Oleinik and Samokhin established the local existence of smooth solutions for boundary value problems of the Prandtl system since the 1960's, and their works were surveyed recently in the monography [5]. The existence and uniqueness of global weak solutions to the Prandtl system are recently established by Xin and Zhang [13] and Xin, Zhang, and Zhao [14], respectively. In [7, 8], Sammartino and Caflisch obtained the local existence of analytic solutions to the Prandtl system, and a rigorous theory on the boundary layer in incompressible fluids with analytic data in the frame of the abstract Cauchy–Kowaleskaya theory. Grenier [2, 3] investigated the stability of boundary layer type solutions to the Euler equations and the instability of solutions to the incompressible Navier–Stokes equations. Till now, there existed no general rigorous theory of viscous boundary layer in the case of nonslip boundary conditions; this is reviewed in [1, 11]. The problem of the viscous boundary layer in the case of slip boundary conditions was studied rigorously by Temam and Wang in [10].

To study the theory of the viscous boundary layer for compressible fluids with nonslip boundary conditions, recently, Xin and Yanagisawa [12] obtained a rigorous justification of the Prandtl boundary layer theory for the linearized compressible fluids when the viscosity goes to zero.

The purpose of this paper is to study the asymptotic behavior of solutions to the linearized compressible Navier–Stokes equations on the half-plane with nonslip boundary conditions perturbed by high frequency oscillatory force terms, and to investigate the interaction between the linearized boundary layer and rapidly oscillatory waves.

In the case that the oscillation of the force term is propagated along the tangential characteristic field of the boundary with respect to the linearized Euler operator, see (2.6)–(2.9), and the wavelength is proportional to the square root of the viscosity, we establish a rigorous theory on the boundary layer and its oscillatory behavior. Roughly speaking, it is shown that the leading profiles of the solution to the linearized compressible Navier–Stokes equations can be divided into four terms: the first one is the outflow satisfying the linearized Euler equations, the second one is an oscillatory wave propagated along the characteristic field tangential to the boundary associated with the linearized Euler operator in the half-plane, and its amplitude satisfies a linear degenerate parabolic equation with the second order term coming from the viscous term in the linearized Navier–Stokes equations, the third one is the classical linearized Prandtl boundary layer supported in a thin neighborhood of the boundary, and the fourth one is an oscillatory wave propagated in the boundary layer, this term together with its vorticity with respect to the normal variable and the fast variable satisfy an initial-boundary value problem for an Poisson–Prandtl coupled system. This result shows that the zero-viscosity limit of the solution to the linearized compressible Navier–Stokes equations with highly oscillatory forces satisfies the linearized Euler equations away from the boundary, and the oscillation is propagated in a way of linear geometric optics in free space. The boundary layer is of the Prandtl type as usual, but the novelties are that the oscillation is propagated in the layer, and there is an interaction between the boundary layer and highly oscillatory waves near the boundary; for details see Theorem 2.1.

When the wavelength of the oscillatory force term is shorter than the square root of the viscosity, we observe that both the oscillation in the outflow and in the boundary layer appear only in the high order profiles, and the leading profiles are the same as the case studied in [12] with the force term without oscillation. This phenomenon will be explained in the appendix for the case that the wavelength of the oscillatory force term is the same order as the viscosity. The case that the wavelength of the oscillatory force term is longer than the square root of the viscosity is more complicated and challenging as it may destabilize the boundary layer. We shall study this problem in a forthcoming paper.

The nonlinear interaction between the boundary layer and high frequency oscillating waves for the artificial viscosity problem of a semilinear hyperbolic system was studied by Gues in [4], for which the leading profiles of solutions have three terms: the outflow satisfying the hyperbolic problem, an oscillatory wave propagated in the half space (its amplitude satisfies an initial value problem for a degenerate parabolic equation), and the boundary layer, which satisfies an initial-boundary value problem for a degenerate parabolic equation. Due to the nonlinearity of the system, problems for these three profiles are coupled together. Main differences between this paper and Gues' work [4] are that the profile of the boundary layer in the Navier–Stokes system satisfies the Prandtl system even when the force term without oscillation, and the phase function of oscillation we will study is nonlinear in general, which gives rise to the above fourth profile, describing the oscillation in the boundary layer, while the phase function of the oscillatory waves considered by Gues in [4] is linear and vanishes at the boundary, which implies that the above fourth term does not appear in that case (see Remark 5.1).

Another related work is that of Szepessy in [9], which gave a geometric optics expansion for a linearized viscous shock profile perturbed by a highly oscillatory wave in two space variables.

The remainder of this paper shall be arranged as follows. In section 2, we carry out the formal analysis to derive problems for the leading profiles of the asymptotic expansion of the solution to the linearized Navier–Stokes equations with respect to ϵ, proportional to the square root of the viscosity, and observe that one of leading profiles describes the oscillation propagated in the boundary layer, and this profile together with its vorticity with respect to the normal variable and the oscillatory variable satisfy an initial-boundary value problem for an Poisson–Prandtl coupled system. The derivation of problems for higher order profiles will be given in the appendix. The problem for the Poisson–Prandtl coupled equations is not a classical one. To our knowledge, there is no any literature devoted to this kind of problem, so we shall establish the well-posedness of this problem in section 3. In section 4, we rigorously justify the formal analysis of the zero-viscosity limit for the solution to the linearized Navier–Stokes equations. Finally, in the appendix, we shall also study the problem with the wavelength of the oscillatory force term being the same order as the viscosity.

2. Asymptotic analysis and main results. Corresponding to the problem (1.1) for the compressible Navier–Stokes equations, we consider the following linearized problem at a state $V'=(\rho',v_1',v_2')^T$ with a high frequency oscillatory force term in the half-space $\{t,x_1>0,x_2\in\mathbb{R}\}$:

$$\begin{cases} A_0(V')\partial_t V^\epsilon + A_1(V')\partial_{x_1}V^\epsilon + A_2(V')\partial_{x_2}V^\epsilon = B(\epsilon^2, D\epsilon^2)V^\epsilon + \Phi(t,x;\frac{\varphi(t,x)}{\epsilon}) \\ M^+V^\epsilon = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} V^\epsilon = 0, \qquad \text{on} \quad x_1=0 \\ V^\epsilon|_{t=0} = V_0(x), \end{cases} \tag{2.1}$$

where $V^\epsilon=(\rho^\epsilon,v_1^\epsilon,v_2^\epsilon)^T$, $\Phi(t,x;\theta)$ is periodic in $\theta\in T^1=\mathbb{R}/2\pi Z$,

$$A_0(V')=\begin{pmatrix} 1 & 0 & 0 \\ 0 & \rho' & 0 \\ 0 & 0 & \rho' \end{pmatrix},\ A_1(V')=\begin{pmatrix} v_1' & \rho' & 0 \\ c^2 & \rho' v_1' & 0 \\ 0 & 0 & \rho' v_1' \end{pmatrix},\ A_2(V')=\begin{pmatrix} v_2' & 0 & \rho' \\ 0 & \rho' v_2' & 0 \\ c^2 & 0 & \rho' v_2' \end{pmatrix}$$

with $c=\sqrt{\frac{dp(\rho')}{d\rho}}>0$ being the sound speed at V', and

$$B(\epsilon^2, D\epsilon^2)V^\epsilon = \epsilon^2(B_1\partial^2_{x_1}V^\epsilon + B_2\partial^2_{x_2}V^\epsilon + B_3\partial^2_{x_1x_2}V^\epsilon)$$

with $D\ge 0$ being a constant, and

$$B_1=\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1+D & 0 \\ 0 & 0 & 1 \end{pmatrix},\quad B_2=\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1+D \end{pmatrix},\quad B_3=\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & D \\ 0 & D & 0 \end{pmatrix},$$

where we have assumed that $\mu=\epsilon^2$ and $\mu'=D\epsilon^2$ in (1.1).

For convenience we shall assume that the background state V' is smooth and bounded. The case of finite order regularity can be handled as well, but much more bookkeeping is needed.

Suppose that

$$v_1'|_{x_1=0}=0. \tag{2.2}$$

For any fixed $(\xi_1,\xi_2)\neq(0,0)$, denote by

$$\tau_1=-(\xi_1 v_1'+\xi_2 v_2'),\quad \tau_{2,3}=-\left(\xi_1 v_1'+\xi_2 v_2' \pm c\sqrt{\xi_1^2+\xi_2^2}\right) \tag{2.3}$$

eigenvalues of the symbol $L(\tau, \xi_1, \xi_2)$ associated with the linearized Euler operator at V',

$$L(\partial_t, \partial_x) = A_0(V')\partial_t + A_1(V')\partial_{x_1} + A_2(V')\partial_{x_2} \tag{2.4}$$

which means that τ_k are roots to the following characteristic equation:

$$\det(\tau A_0(V') + \xi_1 A_1(V') + \xi_2 A_2(V')) = 0.$$

Denote by $\{\vec{r}_k\}_{k=1}^3$ and $\{\vec{l}_k\}_{k=1}^3$ the associated right and left eigenvectors, respectively,

$$\begin{cases} (\tau_k A_0(V') + \xi_1 A_1(V') + \xi_2 A_2(V'))\vec{r}_k = 0 \\ \vec{l}_k(\tau_k A_0(V') + \xi_1 A_1(V') + \xi_2 A_2(V')) = 0 \end{cases} \tag{2.5}$$

with the normalization

$$\vec{l}_j A_0 \vec{r}_k = \delta_{jk} = \begin{cases} 1, & j = k \\ 0, & j \neq k. \end{cases}$$

It follows from (2.2) that the boundary $\{x_1 = 0\}$ is uniformly characteristic with respect to the eigenvalue $\tau_1 = -(\xi_1 v_1' + \xi_2 v_2')$ associated with the linearized Euler operator (2.4).

As in the classical theory of nonlinear geometric optics [4], it is assumed that the oscillation phase $\varphi(t, x)$ in (2.1) satisfies the eikonal equation with respect to τ_1,

$$\partial_t \varphi + v_1' \partial_{x_1} \varphi + v_2' \partial_{x_2} \varphi = 0. \tag{2.6}$$

In this paper, we shall assume

$$\varphi^0(t, x_2) := \varphi(t, 0, x_2) \neq 0. \tag{2.7}$$

Obviously, the assumption $v_1'|_{x_1=0} = 0$ implies

$$\partial_t \varphi^0 + v_2'(0)\partial_{x_2} \varphi^0 = 0 \tag{2.8}$$

with $v_2'(0)$ denoting $v_2'(t, 0, x_2)$.

Moreover, we assume

$$\partial_{x_2} \varphi^0 = \partial_{x_2} \varphi|_{x_1=0} \neq 0 \tag{2.9}$$

at each point of $\{(t, x_2) \in \mathbb{R}^+ \times \mathbb{R}\}$. If $\partial_{x_2} \varphi^0 \equiv 0$, then from (2.8) we have $\partial_t \varphi^0 = 0$ as well, which implies

$$\varphi^0(t, x_2) \equiv \text{const}$$

yielding no oscillation factor in the boundary layer. The problem in the general case of φ, e.g., $\varphi(t, 0, x_2)$ degenerates in a subset of $(t, x_2) \in \mathbb{R}^+ \times \mathbb{R}$ is interesting, and shall be investigated in the future. As we shall see, the case $\varphi(t, 0, x_2) \equiv 0$ is easier to handle.

In the case of (2.6)–(2.9), we take the following ansatz for the solution of (2.1):

$$V^\epsilon(t, x) = V_{in}^\epsilon(t, x) + V_{bd}^\epsilon(t, x), \tag{2.10}$$

where

$$
\begin{cases} V_{in}^{\epsilon}(t,x) = \sum_{j\geq 0} \epsilon^j \left(a_j(t,x) + c_j\left(t,x; \frac{\varphi(t,x)}{\epsilon}\right)\right) \\ V_{bd}^{\epsilon}(t,x) = \sum_{j\geq 0} \epsilon^j \left(b_j\left(t,x_2; \frac{x_1}{\epsilon}\right) + d_j\left(t,x_2; \frac{x_1}{\epsilon}, \frac{\varphi^0(t,x_2)}{\epsilon}\right)\right), \end{cases} \tag{2.11}
$$

where $c_j(t,x;\theta)$ and $d_j(t,x_2;z,\theta)$ are $2\pi-$periodic in θ with mean value vanishing, and $b_j(t,x_2;z)$ and $d_j(t,x_2;z,\theta)$ are rapidly decreasing in z when $z \to +\infty$.

In what follows, we shall always denote by $C_p^k(T_\theta^1)$ the set of kth order smooth functions which are $2\pi-$periodic in $\theta \in T^1$, $S(\mathbb{R}_z^+)$ the set of smooth functions rapidly decreasing in z when $z \to +\infty$, and $a_j^{(k)}$ $(k = 1,2,3)$ the kth component of a_j etc.

Taking the formal expansion as

$$
\begin{aligned} &(A_0(V')\partial_t + A_1(V')\partial_{x_1} + A_2(V')\partial_{x_2})V_{in}^{\epsilon} - B(\epsilon^2, D\epsilon^2)V_{in}^{\epsilon} - \Phi\left(t,x;\tfrac{\varphi(t,x)}{\epsilon}\right) \\ &= \textstyle\sum_{j\geq -1} \epsilon^j \mathcal{F}_j, \end{aligned} \tag{2.12}
$$

in ϵ, we have

$$
\begin{cases} \mathcal{F}_{-1} = \sum_{k=0}^{2} \varphi_{x_k} A_k(V')\partial_\theta c_0, \\ \mathcal{F}_0 = L(\partial_t,\partial_x)(a_0 + c_0) - (\varphi_{x_1}^2 B_1 + \varphi_{x_2}^2 B_2 + \varphi_{x_1}\varphi_{x_2}B_3)\partial_\theta^2 c_0 - \Phi(t,x;\theta) \\ \qquad + \sum_{k=0}^{2} \varphi_{x_k} A_k(V')\partial_\theta c_1, \\ \cdots\cdots \\ \mathcal{F}_j = L(\partial_t,\partial_x)(a_j + c_j) - (\varphi_{x_1}^2 B_1 + \varphi_{x_2}^2 B_2 + \varphi_{x_1}\varphi_{x_2}B_3)\partial_\theta^2 c_j \\ \qquad + \sum_{k=0}^{2} \varphi_{x_k} A_k(V')\partial_\theta c_{j+1} + f_j \end{cases} \tag{2.13}
$$

for each $j \geq 1$, where $\varphi_{x_k} = \partial_{x_k}\varphi$ with $x_0 = t$, and

$$
\begin{aligned} f_j = &-(\varphi_{x_1x_1}B_1 + \varphi_{x_2x_2}B_2 + \varphi_{x_1x_2}B_3)\partial_\theta c_{j-1} - (2\varphi_{x_1}B_1 + \varphi_{x_2}B_3)\partial_{\theta x_1}^2 c_{j-1} \\ &-(2\varphi_{x_2}B_2 + \varphi_{x_1}B_3)\partial_{\theta x_2}^2 c_{j-1} - (B_1\partial_{x_1}^2 + B_2\partial_{x_2}^2 + B_3\partial_{x_1x_2}^2)(a_{j-2} + c_{j-2}) \end{aligned}
$$

with $a_{-1} = c_{-1} = 0$.

Set $z = \frac{x_1}{\epsilon}$. Then

$$
(A_0(V')\partial_t + A_1(V')\partial_{x_1} + A_2(V')\partial_{x_2})V_{bd}^{\epsilon} - B(\epsilon^2, D\epsilon^2)V_{bd}^{\epsilon} = \sum_{j\geq -1} \epsilon^j \mathcal{G}_j \tag{2.14}
$$

implies

$$
\begin{cases} \mathcal{G}_{-1} = (\varphi_t^0 A_0(0) + \varphi_{x_2}^0 A_2(0))\partial_\theta d_0 + A_1(0)\partial_z(b_0 + d_0), \\ \mathcal{G}_0 = L_{bd}(\partial_t,\partial_{x_2})(b_0 + d_0) + z(\varphi_t^0 A_0'(0) + \varphi_{x_2}^0 A_2'(0))\partial_\theta d_0 - B_1\partial_z^2 b_0 \\ \qquad + zA_1'(0)\partial_z(b_0 + d_0) - (B_1\partial_z^2 + (\varphi_{x_2}^0)^2 B_2 \partial_\theta^2 + \varphi_{x_2}^0 B_3 \partial_{z\theta}^2)d_0 \\ \qquad + (\varphi_t^0 A_0(0) + \varphi_{x_2}^0 A_2(0))\partial_\theta d_1 + A_1(0)\partial_z(b_1 + d_1), \\ \cdots\cdots \\ \mathcal{G}_j = L_{bd}(\partial_t,\partial_{x_2})(b_j + d_j) + z(\varphi_t^0 A_0'(0) + \varphi_{x_2}^0 A_2'(0))\partial_\theta d_j - B_1\partial_z^2 b_j \\ \qquad + zA_1'(0)\partial_z(b_j + d_j) - (B_1\partial_z^2 + (\varphi_{x_2}^0)^2 B_2 \partial_\theta^2 + \varphi_{x_2}^0 B_3 \partial_{z\theta}^2)d_j \\ \qquad + (\varphi_t^0 A_0(0) + \varphi_{x_2}^0 A_2(0))\partial_\theta d_{j+1} + A_1(0)\partial_z(b_{j+1} + d_{j+1}) + g_j \end{cases} \tag{2.15}
$$

for any $j \geq 1$, where g_j depends smoothly on $\{b_k, d_k\}_{k \leq j-1}$ and their derivatives up to order two, $A_k(0) = A_k(V')|_{x_1=0}$, $A_k'(0) = \partial_{x_1}(A_k(V'))|_{x_1=0}$, and

$$L_{bd}(\partial_t, \partial_{x_2}) = A_0(0)\partial_t + A_2(0)\partial_{x_2}.$$

From the equations in (2.1) and the assumption that each term (b_j, d_j) in V_{bd}^ϵ is rapidly decreasing in z when $z \to +\infty$, it is natural to set

$$\mathcal{F}_j \equiv 0 \qquad \text{and} \qquad \mathcal{G}_j \equiv 0 \tag{2.16}$$

in (2.12) and (2.14), respectively, for all $j \geq -1$.

The next step is to derive the governing problems for various order of profiles from (2.16) and initial and boundary conditions given in (2.1).

Let $\{\vec{r}_k(\nabla\varphi), \vec{l}_k(\nabla\varphi)\}_{k=1}^3$ be the right and left eigenvectors given in (2.5) associated with $(\xi_1, \xi_2) = (\varphi_{x_1}, \varphi_{x_2})$.

It follows from $\mathcal{F}_{-1} = 0$ that

$$c_0(t, x; \theta) = v_0(t, x; \theta)\vec{r}_1(\nabla\varphi) \tag{2.17}$$

with $v_0(t, x; \theta)$ being a scalar unknown.

Acting upon the mean value operator

$$\mathbf{m}_\theta(u) = \frac{1}{2\pi}\int_0^{2\pi} u(\theta)d\theta$$

on the equation $\mathcal{F}_0 = 0$, we deduce

$$L(\partial_t, \partial_x)a_0 = \mathbf{m}_\theta(\Phi) \tag{2.18}$$

and the difference between (2.18) and $\mathcal{F}_0 = 0$ gives

$$\begin{aligned} &L(\partial_t, \partial_x)c_0 - (\varphi_{x_1}^2 B_1 + \varphi_{x_2}^2 B_2 + \varphi_{x_1}\varphi_{x_2}B_3)\partial_\theta^2 c_0 - \Phi + \mathbf{m}_\theta(\Phi) \\ &\quad = -\sum_{k=0}^{2} \varphi_{x_k} A_k(V')\partial_\theta c_1. \end{aligned} \tag{2.19}$$

Multiplying $\vec{l}_1(\nabla\varphi)$ from the left of (2.19), and using (2.17), one discovers that $v_0(t, x; \theta)$ satisfies the following problem:

$$\begin{cases} [(\vec{l}_1 A_0 \vec{r}_1)\partial_t + (\vec{l}_1 A_1 \vec{r}_1)\partial_{x_1} + (\vec{l}_1 A_2 \vec{r}_1)\partial_{x_2}]v_0 + \vec{l}_1(A_0\partial_t\vec{r}_1 + A_1\partial_{x_1}\vec{r}_1 + A_2\partial_{x_2}\vec{r}_1)v_0 \\ \qquad -\vec{l}_1(\varphi_{x_1}^2 B_1 + \varphi_{x_2}^2 B_2 + \varphi_{x_1}\varphi_{x_2}B_3)\vec{r}_1\partial_\theta^2 v_0 = \vec{l}_1(\Phi - \mathbf{m}_\theta(\Phi)) \\ v_0|_{t=0} = 0. \end{cases} \tag{2.20}$$

Note that the vector field

$$(\vec{l}_1 A_0 \vec{r}_1)\partial_t + (\vec{l}_1 A_1 \vec{r}_1)\partial_{x_1} + (\vec{l}_1 A_2 \vec{r}_1)\partial_{x_2}$$

is tangential to the boundary $\{x_1 = 0\}$, and

$$\vec{l}_1(\varphi_{x_1}^2 B_1 + \varphi_{x_2}^2 B_2 + \varphi_{x_1}\varphi_{x_2}B_3)\vec{r}_1 = \frac{1}{\rho'}(\varphi_{x_1}^2 + \varphi_{x_2}^2) > 0.$$

The problem (2.20) is the one for a linear degenerate parabolic equation, which can be easily solved.

To solve a_0 from (2.18), we need to impose a boundary datum for $a_0^{(2)}$ on $\{x_1 = 0\}$.

It follows from the ansatz (2.10) and (2.11) that the $0(\epsilon^0)-$term of the boundary condition $M^+V^\epsilon|_{x_1=0} = 0$ in (2.1) gives rise to

$$a_0^{(k)}(t,x) + c_0^{(k)}(t,x;\theta) + b_0^{(k)}(t,x_2;z) + d_0^{(k)}(t,x_2;z,\theta^0) = 0 \tag{2.21}$$

on $\{x_1 = 0, z = 0, \theta = \theta^0\}$ for $k \in \{2,3\}$. Since $c_0^{(k)}$ and $d_0^{(k)}$ are $2\pi-$periodic in θ and θ^0, with mean values vanishing, respectively, the condition (2.21) is equivalent to

$$\begin{cases} a_0^{(k)}(t,x) + b_0^{(k)}(t,x_2;z) = 0 \quad \text{on } \{x_1 = z = 0\} \\ c_0^{(k)}(t,x;\theta) + d_0^{(k)}(t,x_2;z,\theta^0) = 0 \quad \text{on } \{x_1 = z = 0, \theta = \theta^0\} \end{cases} \tag{2.22}$$

for $k \in \{2,3\}$.

Thus, we should first study $b_0^{(2)}$ in order to determine the boundary value of $a_0^{(2)}$.

Acting upon the mean value operator $\mathbf{m}_\theta$ on $\mathcal{G}_{-1} = 0$ leads to

$$A_1(0)\partial_z b_0 = 0. \tag{2.23}$$

So, $\mathcal{G}_{-1} = 0$ gives rise to

$$(\varphi_t^0 A_0(0) + \varphi_{x_2}^0 A_2(0))\partial_\theta d_0 + A_1(0)\partial_z d_0 = 0. \tag{2.24}$$

From (2.23), we obtain immediately that

$$\partial_z b_0^{(1)} = \partial_z b_0^{(2)} = 0$$

which implies

$$b_0^{(1)} = b_0^{(2)} \equiv 0 \tag{2.25}$$

by using $b_0 \in S(I\!R_z^+)$.

Thus, it follows from (2.18) and (2.22) that $a_0(t,x)$ satisfies the following problem for the linearized Euler equations:

$$\begin{cases} (A_0(V')\partial_t + A_1(V')\partial_{x_1} + A_2(V')\partial_{x_2})a_0 = \mathbf{m}_\theta(\Phi), \quad t, x_1 > 0, \\ a_0^{(2)}|_{x_1=0} = 0, \\ a_0|_{t=0} = V_0(x). \end{cases} \tag{2.26}$$

To determine $b_0^{(3)}(t,x_2;z)$, by acting upon the mean value operator $\mathbf{m}_\theta$ on $\mathcal{G}_0 = 0$ we deduce that

$$L_{bd}(\partial_t, \partial_{x_2})b_0 + zA_1'(0)\partial_z b_0 + A_1(0)\partial_z b_1 = B_1\partial_z^2 b_0, \tag{2.27}$$

and the difference between (2.27) and $\mathcal{G}_0 = 0$ gives rise to

$$\begin{aligned} &L_{bd}(\partial_t, \partial_{x_2})d_0 + z(\varphi_t^0 A_0'(0) + \varphi_{x_2}^0 A_2'(0))\partial_\theta d_0 + zA_1'(0)\partial_z d_0 \\ &\quad -(B_1\partial_z^2 + (\varphi_{x_2}^0)^2 B_2 \partial_\theta^2 + \varphi_{x_2}^0 B_3 \partial_{z\theta}^2)d_0 \\ &\quad = -(\varphi_t^0 A_0(0) + \varphi_{x_2}^0 A_2(0))\partial_\theta d_1 - A_1(0)\partial_z d_1. \end{aligned} \tag{2.28}$$

From the third component of (2.27), we conclude that $b_0^{(3)}(t,x_2;z)$ satisfies the following problem:

$$\begin{cases} (\partial_t + v_2'(0)\partial_{x_2})b_0^{(3)} + z\dfrac{\partial v_1'(0)}{\partial x_1}\partial_z b_0^{(3)} - \dfrac{1}{\rho'(0)}\partial_z^2 b_0^{(3)} = 0, \quad t,\ z>0, \\ b_0^{(3)}|_{z=0} = -a_0^{(3)}(t,0,x_2), \\ b_0^{(3)}|_{t=0} = 0, \end{cases} \tag{2.29}$$

where $a_0^{(3)}$ is given by (2.26).

The problem (2.29) is the one for a linearized Prandtl equation, which has been solved by Xin and Yanagisawa in [12].

Now, let us derive the problems for $d_0(t,x_2;z,\theta)$ from (2.24) and (2.28).

It follows from $\varphi_t^0 + v_2'(0)\varphi_{x_2}^0 = 0$ that

$$(\varphi_t^0 A_0(0) + \varphi_{x_2}^0 A_2(0))\partial_\theta d + A_1(0)\partial_z d = \begin{pmatrix} \rho'(0)(\varphi_{x_2}^0 \partial_\theta d^{(3)} + \partial_z d^{(2)}) \\ c^2(0)\partial_z d^{(1)} \\ c^2(0)\varphi_{x_2}^0 \partial_\theta d^{(1)} \end{pmatrix}.$$

Thus, (2.24) yields

$$\varphi_{x_2}^0 \partial_\theta d_0^{(3)} + \partial_z d_0^{(2)} = 0 \tag{2.30}$$

and

$$\varphi_{x_2}^0 \partial_\theta d_0^{(1)} = 0, \quad \partial_z d_0^{(1)} = 0$$

which implies

$$d_0^{(1)} \equiv 0. \tag{2.31}$$

To solve $(d_0^{(2)}, d_0^{(3)})$, we define the operator $\mathbb{E}$ by

$$\mathbb{E}\begin{pmatrix} d^{(1)} \\ d^{(2)} \\ d^{(3)} \end{pmatrix} = \begin{pmatrix} \mathbf{m}_\theta d^{(1)} \\ \varphi_{x_2}^0 \partial_\theta d^{(2)} - \partial_z d^{(3)} \end{pmatrix}$$

for any $d = (d^{(1)}, d^{(2)}, d^{(3)})^T \in C^1(\mathbb{R}_z^+ \times T_\theta^1)$. It is easy to know that for any $d(t,x_2;z,\theta) \in C_p^1(T_\theta^1) \cap S(\mathbb{R}_z^+)$ with $\mathbf{m}_\theta(d) = 0$, we have

$$\mathbb{E}((\varphi_t^0 A_0(0) + \varphi_{x_2}^0 A_2(0))\partial_\theta d + A_1(0)\partial_z d) = 0. \tag{2.32}$$

Acting upon the operator $\mathbb{E}$ on (2.28) and using (2.32), one gets

$$\mathbb{E}(\text{left-hand side of (2.28)}) = 0. \tag{2.33}$$

Denote by A and B the second and the third components of the left-hand side of (2.28), respectively. Then, one obtains from (2.30) and (2.31) that

$$\begin{cases} A = \rho'(0)((\partial_t + v_2'(0)\partial_{x_2})d_0^{(2)} + z\dfrac{\partial v_1'(0)}{\partial x_1}\partial_z d_0^{(2)} + z\dfrac{\partial v_2'(0)}{\partial x_1}\varphi_{x_2}^0\partial_\theta d_0^{(2)}) - (\partial_z^2 + (\varphi_{x_2}^0)^2\partial_\theta^2)d_0^{(2)}, \\ B = \rho'(0)((\partial_t + v_2'(0)\partial_{x_2})d_0^{(3)} + z\dfrac{\partial v_1'(0)}{\partial x_1}\partial_z d_0^{(3)} + z\dfrac{\partial v_2'(0)}{\partial x_1}\varphi_{x_2}^0\partial_\theta d_0^{(3)}) - (\partial_z^2 + (\varphi_{x_2}^0)^2\partial_\theta^2)d_0^{(3)}. \end{cases}$$

Now, (2.33) implies that

$$\varphi^0_{x_2}\partial_\theta A - \partial_z B = 0,$$

which can be explicitly written as

$$(2.34)\quad (\partial_t + v_2'(0)\partial_{x_2})\omega_0 + z\left(\frac{\partial v_1'(0)}{\partial x_1}\partial_z + \varphi^0_{x_2}\frac{\partial v_2'(0)}{\partial x_1}\partial_\theta\right)\omega_0 - \frac{1}{\rho'(0)}(\partial_z^2 + (\varphi^0_{x_2})^2\partial_\theta^2)\omega_0 - \left(\frac{\partial v_1'(0)}{\partial x_1}\partial_z + \varphi^0_{x_2}\frac{\partial v_2'(0)}{\partial x_1}\partial_\theta\right)d_0^{(3)} = 0,$$

where $\omega_0(t,x_2;z,\theta) = \varphi^0_{x_2}\partial_\theta d_0^{(2)} - \partial_z d_0^{(3)}$.

Combining (2.30) with (2.34) and using (2.22), one obtains that $d_0^{(3)}$ and ω_0 satisfy the following problem:

$$(2.35)\quad \begin{cases} (\partial_z^2 + (\varphi^0_{x_2})^2\partial_\theta^2)d_0^{(3)} = -\partial_z\omega_0 \\ (\partial_t + v_2'(0)\partial_{x_2})\omega_0 + z\left(\dfrac{\partial v_1'(0)}{\partial x_1}\partial_z + \varphi^0_{x_2}\dfrac{\partial v_2'(0)}{\partial x_1}\partial_\theta\right)\omega_0 \\ \qquad -\dfrac{1}{\rho'(0)}(\partial_z^2 + (\varphi^0_{x_2})^2\partial_\theta^2)\omega_0 - \left(\dfrac{\partial v_1'(0)}{\partial x_1}\partial_z + \varphi^0_{x_2}\dfrac{\partial v_2'(0)}{\partial x_1}\partial_\theta\right)d_0^{(3)} = 0 \\ d_0^{(3)}|_{z=0} = -c_0^{(3)}(t,0,x_2;\theta) \\ (\omega_0 + \partial_z d_0^{(3)})|_{z=0} = -\varphi^0_{x_2}(\partial_\theta c_0^{(2)})(t,0,x_2;\theta) \\ (d_0^{(3)},\omega_0) \in S(\mathbb{R}_z^+) \\ \omega_0|_{t=0} = 0 \end{cases}$$

and $d_0^{(2)}(t,x_2;z,\theta)$ satisfies

$$(2.36)\quad \begin{cases} (\partial_z^2 + (\varphi^0_{x_2})^2\partial_\theta^2)d_0^{(2)} = \varphi^0_{x_2}\partial_\theta\omega_0 \\ d_0^{(2)}|_{z=0} = -c_0^{(2)}(t,0,x_2;\theta) \\ d_0^{(2)} \in S(\mathbb{R}_z^+), \end{cases}$$

where $(c_0^{(2)}, c_0^{(3)})$ are given by (2.17) and (2.20).

The problems for high order terms in the expansion of $V_{in}^\epsilon(t,x) + V_{bd}^\epsilon(t,x)$ can be formulated in a similar way. For completeness, we shall derive these problems in the appendix.

In section 3, we study the well-posedness of the problem (2.35) in detail, and in section 4 we justify rigorously the above formal analysis to conclude the following main result of this paper.

THEOREM 2.1. *Suppose that compatibility conditions for problems* (2.1), (2.29), (2.35), *and* (5.11), (5.13), (5.22) *given later are satisfied. Then, the solution* $V^\epsilon = (\rho^\epsilon, v_1^\epsilon, v_2^\epsilon)$ *of* (2.1) *admits the following asymptotics:*

$$(2.37)\quad V^\epsilon(t,x) = a_0(t,x) + c_0\left(t,x;\frac{\varphi(t,x)}{\epsilon}\right) + b_0\left(t,x_2;\frac{x_1}{\epsilon}\right) + d_0\left(t,x_2;\frac{x_1}{\epsilon},\frac{\varphi^0(t,x_2)}{\epsilon}\right) + O(\epsilon)$$

in $L^\infty([0,T], L^2(\mathbb{R}^2_+))$ for any $T > 0$, where $a_0(t,x)$ satisfies the problem for the linearized Euler equations (2.26)*, $c_0 = v_0(t,x; \frac{\varphi(t,x)}{\epsilon})\vec{r}_1(\nabla\varphi)$ with v_0 satisfying the degenerate parabolic equation* (2.20)*, $(b_0^{(1)}, b_0^{(2)}) = 0$ and $b_0^{(3)}(t,x_2;z)$ satisfies the linearized Prandtl equation* (2.29)*, $d_0^{(1)} = 0$, and $(d_0^{(2)}, d_0^{(3)})(t,x_2;z,\theta)$ together with its vorticity with respect to $(z,\theta)-$variables satisfy the Poisson–Prandtl coupled system* (2.35) *and the Poisson equation* (2.36)*, respectively.*

Remark 2.2. When the wavelength of the force term Φ is shorter than ϵ, i.e., $\Phi = \Phi(t,x; \frac{\varphi(t,x)}{\epsilon^\alpha})$ with $\alpha > 1$, we obtain $c_0 = d_0 \equiv 0$, and (a_0, b_0) are the same as the case studied in [12] with $\Phi \equiv 0$. This phenomenon will be studied for the case $\alpha = 2$ in the appendix.

3. The study of a Poisson–Prandtl coupled system. It is clear from problems (2.35), (2.36), and (5.22), (5.23) given later that in order to determine $(d_j^{(2)}, d_j^{(3)})$ for any $j \geq 0$, we need to study the following initial-boundary value problem for a Poisson–Prandtl coupled system in $\{t, z > 0, x \in \mathbb{R}, \theta \in T^1\}$:

$$\tag{3.1}
\begin{cases}
(\partial_z^2 + a^2\partial_\theta^2)u = f(t,x;z,\theta) - \partial_z w \\
(\partial_t + a_1\partial_x)w + z(a_2\partial_z + a_3\partial_\theta)w - a_4^2(\partial_z^2 + a^2\partial_\theta^2)w - (a_2\partial_z + a_3\partial_\theta)u = g(t,x;z,\theta) \\
u|_{z=0} = b_0(t,x;\theta), \quad u \in S(\mathbb{R}_z^+) \\
(w + \partial_z u)|_{z=0} = b_1(t,x;\theta), \quad (u,w) \in S(\mathbb{R}_z^+) \\
w|_{t=0} = 0
\end{cases}$$

for the unknowns (u,w), where (f,g) are rapidly decreasing in z when $z \to +\infty$, and periodic in $\theta \in T^1 = \mathbb{R}/2\pi Z$ as well as for $(b_0, b_1)(t,x;\theta)$ with mean values vanishing,

$$\mathbf{m}_\theta(f) = \mathbf{m}_\theta(g) = \mathbf{m}_\theta(b_0) = \mathbf{m}_\theta(b_1) = 0,$$

all coefficients in (3.1) are smooth functions of (t,x), with $a(t,x) \geq a_0$, $a_4(t,x) \geq a_0$ for a positive constant a_0.

For simplicity, we assume that (f, g, b_0, b_1) are smooth. To study smooth solutions to (3.1), as usual, one needs to impose compatibility conditions on data.

(1) The zero-th order compatibility condition.

From the initial data $w|_{t=0} = 0$, we have $\partial_z w|_{t=0} = 0$. Thus, from the first and third equations of (3.1), the datum $u_0(x,z,\theta) = u|_{t=0}$ should satisfy the problem

$$\tag{3.2}
\begin{cases}
(\partial_z^2 + a_0^2\partial_\theta^2)u_0 = f(0,x;z,\theta) \\
u_0|_{z=0} = b_0(0,x,\theta), \quad u_0 \in S(\mathbb{R}_z^+),
\end{cases}$$

where $a_0(x) = a(0,x)$.

By using Remark 3.2 given later, the problem (3.2) has a unique solution $u_0(x,z,\theta)$, and can be explicitly given by $f(0,x;z,\theta)$ and $b_0(0,x,\theta)$

From the fourth equation in (3.1), we conclude the zero-th order compatibility condition for (3.1) as follows:

$$\tag{3.3} b_1(0,x;\theta) = \partial_z u_0|_{z=0}.$$

(2) The kth order compatibility condition for any fixed integer $k \geq 1$.

In the discussion of compatibility conditions of (3.1) up to order $k-1$, we first suppose that one has the data $u_l(x,z,\theta) = \partial_t^l u|_{t=0}$ and $w_l(x,z,\theta) = \partial_t^l w|_{t=0}$ for any integer $l \leq k-1$ in terms of (f, g, b_0, b_1). From the second equation in (3.1), we immediately obtain the datum $w_k(x,z,\theta) = \partial_t^k w|_{t=0}$ in terms of $\{u_l, w_l\}_{l \leq k-1}$. By differentiating the first equation of (3.1) k-times with respect to t, and applying Remark 3.2 (given later) to solve the problem

$$\begin{cases} (\partial_z^2 + a_0^2 \partial_\theta^2) u_k = F_k(x,z,\theta) \\ u_k|_{z=0} = (\partial_t^k b_0)(0,x,\theta), \quad u_k \in S(\mathbb{R}_z^+) \end{cases}$$

with $F_k(x,z,\theta) = (\partial_t^k f - \partial_t^k (a^2 \partial_\theta^2 u) + a^2 \partial_\theta^2 \partial_t^k u - \partial_z \partial_t^k w)|_{t=0}$ being given in terms of $\{u_l\}_{l \leq k-1}$ and $\{w_l\}_{l \leq k}$, we can determine the data $u_k(x,z,\theta) = \partial_z^k u|_{t=0}$.

In this way, we get formulae of (u_k, w_k) in terms of (f, g, b_0, b_1). From the boundary condition of (3.1), it follows that the kth order compatibility condition should be

$$(w_k + \partial_z u_k)|_{z=0} = (\partial_t^k b_1)(0, x; \theta) \tag{3.4}$$

which can be explicitly formulated in terms of (f, g, b_0, b_1).

In the remainder of this section, for simplicity, we assume that any order compatibility condition is satisfied for the problem (3.1).

The goal of this section is to study the solvability of the problem (3.1) in the class that u and w are rapidly decreasing when $z \to +\infty$ and periodic in $\theta \in T^1$ with $\mathbf{m}_\theta(u, w) = 0$, which constitutes the main part of the rigorous justification for the formal analysis given in section 2. To this end, first, we derive a functional representation $u = u(w)$ of u in terms of w from the first and the third equations of (3.1), second, by substituting the relation $u = u(w)$ into the second and the fourth equations of (3.1), we solve the unknown $w = w(t, x; z, \theta)$.

3.1. Derivation of the representation = ( ). To deduce the relation $u = u(w)$, we first consider the following boundary value problem:

$$\begin{cases} (\partial_z^2 + a^2 \partial_\theta^2) u = F(t, x; z, \theta) \\ u|_{z=0} = b_0(t, x; \theta), \quad u \in S(\mathbb{R}_z^+), \end{cases} \tag{3.5}$$

where F is rapidly decreasing when $z \to +\infty$, and (b_0, F) are periodic in $\theta \in T^1$ with mean values vanishing.

Obviously, to solve the problem (3.5) is equivalent to studying the following problem:

$$\begin{cases} (\partial_z^2 + a^2 \partial_\theta^2) u = F(t, x; z, \theta) \\ u|_{z=0} = b_0(t, x; \theta), \quad u_z|_{z=0} = u_0(t, x; \theta), \end{cases} \tag{3.6}$$

where u_0, periodic in $\theta \in T^1$ with $\mathbf{m}_\theta(u_0) = 0$, will be determined by $(b_0(t, x; \theta), F(t, x; z, \theta))$ such that the problem (3.6) admits a unique solution $u(t, x; z, \theta) \in C_p^2(T_\theta^1) \cap S(\mathbb{R}_z^+)$ with $\mathbf{m}_\theta(u) = 0$.

Denote by

$$\begin{cases} F(t, x; z, \theta) = \displaystyle\sum_{k \neq 0} F^{(k)}(t, x; z) e^{ik\theta} \\ b_0(t, x; \theta) = \displaystyle\sum_{k \neq 0} b_0^{(k)}(t, x) e^{ik\theta} \end{cases} \tag{3.7}$$

the Fourier expansions of (F, b_0) with respect to $\theta \in T^1$.

LEMMA 3.1. *The necessary and sufficient condition for the solution $u(t,x;z,\theta)$ of (3.6) to be rapidly decreasing when $z \to +\infty$ is*

$$u_0(t,x;\theta) = -\sum_{k=1}^{\infty}\left(kab_0^{(k)} + \int_0^\infty e^{-ka\xi}F^{(k)}(t,x;\xi)d\xi\right)e^{ik\theta} + \sum_{k=-1}^{-\infty}\left(kab_0^{(k)} - \int_0^\infty e^{ka\xi}F^{(k)}(t,x;\xi)d\xi\right)e^{ik\theta}. \tag{3.8}$$

Proof. (1) First, we solve the following problem:

$$\begin{cases} (\partial_z^2 + a^2\partial_\theta^2)w = F(t,x;z,\theta) \\ w|_{z=0} = 0, \quad w_z|_{z=0} = w_0(t,x;\theta). \end{cases} \tag{3.9}$$

We will find $w_0(t,x;\theta)$, periodic in $\theta \in T^1$ with $\mathbf{m}_\theta(w_0) = 0$, such that the solution $w(t,x;z,\theta)$ to (3.9) is rapidly decreasing when $z \to +\infty$.

If we set

$$\begin{cases} w(t,x;z,\theta) = \sum_{k\neq 0} w^{(k)}(t,x;z)e^{ik\theta} \\ w_0(t,x;\theta) = \sum_{k\neq 0} w_0^{(k)}(t,x)e^{ik\theta}, \end{cases} \tag{3.10}$$

then the problem (3.9) is equivalent to the following one for $w^{(k)}(t,x;z)$:

$$\begin{cases} (\partial_z^2 - k^2a^2)w^{(k)} = F^{(k)}(t,x;z) \\ w^{(k)}|_{z=0} = 0, \quad w_z^{(k)}|_{z=0} = w_0^{(k)}(t,x;\theta) \end{cases} \tag{3.11}$$

for any $k \neq 0$.

Obviously, the solution to (3.11) can be expressed as

$$w^{(k)}(t,x;z) = \frac{1}{2ka}\left(w_0^{(k)}(t,x) + \int_0^z e^{-ka\xi}F^{(k)}(t,x;\xi)d\xi\right)e^{kaz} - \frac{1}{2ka}\left(w_0^{(k)}(t,x) + \int_0^z e^{ka\xi}F^{(k)}(t,x;\xi)d\xi\right)e^{-kaz}. \tag{3.12}$$

When $k > 0$, the necessary condition for $w^{(k)} \in S(\mathbb{R}_z^+)$ is

$$\lim_{z\to+\infty}\left(w_0^{(k)}(t,x) + \int_0^z e^{-ka\xi}F^{(k)}(t,x;\xi)d\xi\right) = 0$$

which implies

$$w_0^{(k)}(t,x) = -\int_0^\infty e^{-ka\xi}F^{(k)}(t,x;\xi)d\xi. \tag{3.13}$$

Substituting (3.13) into (3.12) it follows

$$w^{(k)}(t,x;z) = -\frac{1}{2ka}\int_z^\infty e^{ka(z-\xi)}F^{(k)}(t,x;\xi)d\xi + \frac{1}{2ka}\int_0^\infty e^{-ka(z+\xi)}F^{(k)}(t,x;\xi)d\xi - \frac{1}{2ka}\int_0^z e^{ka(\xi-z)}F^{(k)}(t,x;\xi)d\xi. \tag{3.14}$$

Since $F^{(k)} \in S(\mathbb{R}_z^+)$, we deduce

$$\int_0^\infty e^{-ka(z+\xi)} F^{(k)}(t,x;\xi)d\xi \in S(\mathbb{R}_z^+)$$

and

$$\int_z^\infty e^{ka(z-\xi)} F^{(k)}(t,x;\xi)d\xi \in S(\mathbb{R}_z^+).$$

On the other hand, we have

$$\left| z^l \int_0^z e^{ka(\xi-z)} F^{(k)}(t,x;\xi)d\xi \right| \le \sum_{0\le j\le l} \binom{l}{j} \left| \int_0^z (z-\xi)^{l-j}\xi^j e^{-ka(z-\xi)} F^{(k)}(t,x;\xi)d\xi \right|$$

which is bounded for all $l \ge 0$ by using $F^{(k)} \in S(\mathbb{R}_z^+)$. Thus, we also have

$$\int_0^z e^{ka(\xi-z)} F^{(k)}(t,x;\xi)d\xi \in S(\mathbb{R}_z^+).$$

Therefore, the function $w^{(k)}(t,x;z)$ given in (3.14) is rapidly decreasing when $z \to +\infty$.

Similarly, we deduce that when $k < 0$, the necessary and sufficient condition for $w^{(k)}$ given in (3.1) belonging to $S(\mathbb{R}_z^+)$ is

$$w_0^{(k)}(t,x) = -\int_0^\infty e^{ka\xi} F^{(k)}(t,x;\xi)d\xi \tag{3.15}$$

and in this case, the solution to (3.11) can be expressed as

$$\begin{aligned} w^{(k)}(t,x;z) = &\frac{1}{2ka}\int_z^\infty e^{ka(\xi-z)} F^{(k)}(t,x;\xi)d\xi + \frac{1}{2ka}\int_0^z e^{ka(z-\xi)} F^{(k)}(t,x;\xi)d\xi \\ &- \frac{1}{2ka}\int_0^\infty e^{ka(\xi+z)} F^{(k)}(t,x;\xi)d\xi. \end{aligned} \tag{3.16}$$

Combining (3.13), (3.14), and (3.15) with (3.16) shows that

$$\begin{aligned} w(t,x;z,\theta) = &\sum_{k=1}^{\infty} \frac{1}{2ka}\left[\int_0^\infty e^{-ka(z+\xi)} F^{(k)}(t,x;\xi)d\xi - \int_0^z e^{-ka(z-\xi)} F^{(k)}(t,x;\xi)d\xi \right. \\ &\left. - \int_z^\infty e^{ka(z-\xi)} F^{(k)}(t,x;\xi)d\xi\right] e^{ik\theta} + \sum_{k=-1}^{-\infty} \frac{1}{2ka}\left[\int_0^z e^{ka(z-\xi)} F^{(k)}(t,x;\xi)d\xi \right. \\ &\left. - \int_0^\infty e^{ka(z+\xi)} F^{(k)}(t,x;\xi)d\xi + \int_z^\infty e^{-ka(z-\xi)} F^{(k)}(t,x;\xi)d\xi\right] e^{ik\theta} \\ &\in S(\mathbb{R}_z^+) \end{aligned} \tag{3.17}$$

is the unique solution to

$$\begin{cases} (\partial_z^2 + a^2\partial_\theta^2)w = F(t,x;z,\theta) \\ w|_{z=0} = 0 \\ w_z|_{z=0} = -\sum_{k=1}^{\infty}\int_0^\infty e^{k(i\theta-a\xi)} F^{(k)}(t,x;\xi)d\xi - \sum_{k=-1}^{-\infty}\int_0^\infty e^{k(i\theta+a\xi)} F^{(k)}(t,x;\xi)d\xi. \end{cases} \tag{3.18}$$

(2) Let $v = u - w$ with u being the solution to (3.6). Then v solves the following problem:

$$
\begin{cases}
(\partial_z^2 + a^2\partial_\theta^2)v = 0 \\
v|_{z=0} = b_0(t,x;\theta) \\
v_z|_{z=0} = u_0(t,x;\theta) + \sum_{k=1}^{\infty}\int_0^\infty e^{k(i\theta - a\xi)}F^{(k)}(t,x;\xi)d\xi + \sum_{k=-1}^{-\infty}\int_0^\infty e^{k(i\theta + a\xi)}F^{(k)}(t,x;\xi)d\xi.
\end{cases}
\tag{3.19}
$$

Denote by

$$
\begin{cases}
v(t,x;z,\theta) = \sum_{k\neq 0} v^{(k)}(t,x;z)e^{ik\theta}, \\
u_0(t,x;\theta) = \sum_{k\neq 0} u_0^{(k)}(t,x)e^{ik\theta}
\end{cases}
$$

the Fourier expansions of (v, u_0). Then, from (3.19), $v^{(k)}$ satisfies

$$
\begin{cases}
(\partial_z^2 - k^2a^2)v^{(k)} = 0, \\
v^{(k)}|_{z=0} = b_0^{(k)}(t,x), \\
v_z^{(k)}|_{z=0} = \begin{cases} u_0^{(k)}(t,x) + \int_0^\infty e^{-ka\xi}F^{(k)}(t,x;\xi)d\xi, & k \geq 1, \\ u_0^{(k)}(t,x) + \int_0^\infty e^{ka\xi}F^{(k)}(t,x;\xi)d\xi, & k \leq -1. \end{cases}
\end{cases}
\tag{3.20}
$$

It follows that

$$
\begin{aligned}
v^{(k)}(t,x;z) = & \left[\frac{1}{2}b_0^{(k)} + \frac{1}{2ka}\left(u_0^{(k)} + \int_0^{+\infty} e^{-ka\xi}F^{(k)}(t,x;\xi)d\xi\right)\right]e^{kaz} \\
& + \left[\frac{1}{2}b_0^{(k)} - \frac{1}{2ka}\left(u_0^{(k)} + \int_0^{+\infty} e^{-ka\xi}F^{(k)}(t,x;\xi)d\xi\right)\right]e^{-kaz}
\end{aligned}
\tag{3.21}
$$

when $k > 0$, and

$$
\begin{aligned}
v^{(k)}(t,x;z) = & \left[\frac{1}{2}b_0^{(k)} + \frac{1}{2ka}\left(u_0^{(k)} + \int_0^{+\infty} e^{ka\xi}F^{(k)}(t,x;\xi)d\xi\right)\right]e^{kaz} \\
& + \left[\frac{1}{2}b_0^{(k)} - \frac{1}{2ka}\left(u_0^{(k)} + \int_0^{+\infty} e^{ka\xi}F^{(k)}(t,x;\xi)d\xi\right)\right]e^{-kaz}
\end{aligned}
\tag{3.22}
$$

when $k < 0$.

From (3.21) and (3.22), we conclude that one should have the condition (3.8) to guarantee $v^{(k)} \in S(\mathbb{R}_z^+)$, and in this case we have

$$
v(t,x;z,\theta) = \sum_{k=1}^{\infty} b_0^{(k)}(t,x)e^{-k(az - i\theta)} + \sum_{k=-1}^{-\infty} b_0^{(k)}(t,x)e^{k(az + i\theta)}. \qquad \square
\tag{3.23}
$$

Remark 3.2. From (3.17) and (3.23), it immediately follows that the unique solution to the problem (3.5) is given by

(3.24)

$$\begin{aligned} u(t,x;z,\theta) = &\sum_{k=1}^{\infty} b_0^{(k)}(t,x)e^{-k(az-i\theta)} + \sum_{k=-1}^{-\infty} b_0^{(k)}(t,x)e^{k(az+i\theta)} \\ &+\sum_{k=1}^{\infty}\frac{1}{2ka}\bigg[\int_0^{\infty} e^{-ka(z+\xi)}F^{(k)}(t,x;\xi)d\xi - \int_0^{z} e^{-ka(z-\xi)}F^{(k)}(t,x;\xi)d\xi \\ &- \int_z^{\infty} e^{ka(z-\xi)}F^{(k)}(t,x;\xi)d\xi\bigg]e^{ik\theta} + \sum_{k=-1}^{-\infty}\frac{1}{2ka}\bigg[\int_0^{z} e^{ka(z-\xi)}F^{(k)}(t,x;\xi)d\xi \\ &- \int_0^{\infty} e^{ka(z+\xi)}F^{(k)}(t,x;\xi)d\xi + \int_z^{\infty} e^{-ka(z-\xi)}F^{(k)}(t,x;\xi)d\xi\bigg]e^{ik\theta}. \end{aligned}$$

For the problem (3.1), let the Fourier expansion of w be

$$w(t,x;z,\theta) = \sum_{k\neq 0} w^{(k)}(t,x;z)e^{ik\theta}.$$

Using (3.24), we conclude with the following proposition.

PROPOSITION 3.3. *For the problem* (3.1), *the solution u has the following representation in terms of w:*

$$\begin{aligned} u(t,x;z,\theta) = &\sum_{k=1}^{\infty} b_0^{(k)}(t,x)e^{k(i\theta-az)} + \sum_{k=-1}^{-\infty} b_0^{(k)}(t,x)e^{k(i\theta+az)} \\ &+\sum_{k=1}^{\infty}\frac{1}{2}\bigg\{\int_0^{\infty} e^{-ka(z+\xi)}\left(\frac{f^{(k)}(t,x;\xi)}{ka} - w^{(k)}(t,x;\xi)\right)d\xi \\ &- \int_0^{z} e^{-ka(z-\xi)}\left(\frac{f^{(k)}(t,x;\xi)}{ka} + w^{(k)}(t,x;\xi)\right)d\xi \\ &- \int_z^{\infty} e^{ka(z-\xi)}\left(\frac{f^{(k)}(t,x;\xi)}{ka} - w^{(k)}(t,x;\xi)\right)d\xi\bigg\}e^{ik\theta} \\ &- \sum_{k=-1}^{-\infty}\frac{1}{2}\bigg\{\int_0^{\infty} e^{ka(z+\xi)}\left(\frac{f^{(k)}(t,x;\xi)}{ka} + w^{(k)}(t,x;\xi)\right)d\xi \\ &- \int_0^{z} e^{ka(z-\xi)}\left(\frac{f^{(k)}(t,x;\xi)}{ka} + w^{(k)}(t,x;\xi)\right)d\xi \\ &- \int_z^{\infty} e^{-ka(z-\xi)}\left(\frac{f^{(k)}(t,x;\xi)}{ka} + w^{(k)}(t,x;\xi)\right)d\xi\bigg\}e^{ik\theta} \end{aligned} \tag{3.25}$$

and

$$\begin{aligned} \partial_z u|_{z=0} = &\sum_{k=-1}^{-\infty} kab_0^{(k)}(t,x)e^{ki\theta} - \sum_{k=1}^{\infty} kab_0^{(k)}(t,x)e^{ki\theta} \\ &- \sum_{k=1}^{\infty}\int_0^{\infty} e^{k(i\theta-a\xi)}(f^{(k)}(t,x;\xi) - \partial_\xi w^{(k)}(t,x;\xi))d\xi \\ &- \sum_{k=-1}^{-\infty}\int_0^{\infty} e^{k(i\theta+a\xi)}(f^{(k)}(t,x;\xi) - \partial_\xi w^{(k)}(t,x;\xi))d\xi. \end{aligned} \tag{3.26}$$

3.2. The problem of a linear integro-Prandtl equation. It follows from Proposition 3.3 that to solve the problem (3.1), it suffices to use (3.25) and (3.26) to study the following problem for w:

$$\begin{cases} (\partial_t + a_1\partial_x)w + z(a_2\partial_z + a_3\partial_\theta)w - a_4^2(\partial_z^2 + a^2\partial_\theta^2)w - (a_2\partial_z + a_3\partial_\theta)u = g, \\ w|_{z=0} = b_1(t,x;\theta) - \partial_z u(t,x;0,\theta), \quad w \in S(\mathbb{R}_z^+), \\ w|_{t=0} = 0. \end{cases} \tag{3.27}$$

Obviously, in (3.27) both the equation and boundary condition have integral terms coming from those of u. First, let us transform (3.27) into a problem with the boundary condition being a standard Dirichlet form.

The compatibility conditions for the problem (3.27) follow immediately from those for the problem (3.1) given at the beginning of this section.

Denote by

$$\begin{cases} w(t,x;z,\theta) = \sum\limits_{k\neq 0} w^{(k)}(t,x;z)e^{ik\theta}, \\ g(t,x;z,\theta) = \sum\limits_{k\neq 0} g^{(k)}(t,x;z)e^{ik\theta}, \\ b_1(t,x;\theta) = \sum\limits_{k\neq 0} b_1^{(k)}(t,x)e^{ik\theta} \end{cases} \tag{3.28}$$

the Fourier expansions with respect to $\theta \in T^1$.

It follows from (3.27) that $w^{(k)}(t,x;z)$ satisfies the following problem:

$$\begin{cases} (\partial_t + a_1\partial_x)w^{(k)} + z(a_2\partial_z + ika_3)w^{(k)} - a_4^2(\partial_z^2 - k^2a^2)w^{(k)} + a_2 w^{(k)} \\ \qquad + \dfrac{k}{2}(ia_3 - aa_2)\Big[\displaystyle\int_0^{+\infty} e^{-ka(z+\xi)}\Big(w^{(k)}(t,x;\xi) - \frac{f^{(k)}(t,x;\xi)}{ka}\Big)d\xi \\ \qquad + \displaystyle\int_0^z e^{-ka(z-\xi)}\Big(w^{(k)}(t,x;\xi) + \frac{f^{(k)}(t,x;\xi)}{ka}\Big)d\xi\Big] \\ \qquad - \dfrac{k}{2}(aa_2 + ia_3)\displaystyle\int_z^{+\infty} e^{ka(z-\xi)}\Big(w^{(k)}(t,x;\xi) - \frac{f^{(k)}(t,x;\xi)}{ka}\Big)d\xi \\ \qquad\quad = g^{(k)}(t,x;z) - k(aa_2 + ia_3)b_0^{(k)}e^{-kaz} \\ w^{(k)}|_{z=0} = b_1^{(k)} + kab_0^{(k)} + \displaystyle\int_0^{+\infty} e^{-ka\xi}(f^{(k)}(t,x;\xi) - \partial_\xi w^{(k)}(t,x;\xi))d\xi \\ w^{(k)}|_{t=0} = 0, \quad w^{(k)} \in S(\mathbb{R}_z^+) \end{cases} \tag{3.29}$$

for any $k \geq 1$, and

$$
\begin{cases}
(\partial_t + a_1\partial_x)w^{(k)} + z(a_2\partial_z + ika_3)w^{(k)} - a_4^2(\partial_z^2 - k^2a^2)w^{(k)} + a_2 w^{(k)} \\
\quad + \frac{k}{2}(aa_2 + ia_3)\Big[\int_0^{+\infty} e^{ka(z+\xi)}\Big(w^{(k)}(t,x;\xi) + \frac{f^{(k)}(t,x;\xi)}{ka}\Big)d\xi \\
\quad + \int_0^z e^{ka(z-\xi)}\Big(w^{(k)}(t,x;\xi) - \frac{f^{(k)}(t,x;\xi)}{ka}\Big)d\xi\Big] \\
\quad + \frac{k}{2}(aa_2 - ia_3)\int_z^{+\infty} e^{-ka(z-\xi)}\Big(w^{(k)}(t,x;\xi) + \frac{f^{(k)}(t,x;\xi)}{ka}\Big)d\xi \\
\qquad\qquad = g^{(k)}(t,x;z) + k(aa_2 + ia_3)b_0^{(k)}e^{kaz} \\
w^{(k)}|_{z=0} = b_1^{(k)} - kab_0^{(k)} + \int_0^{+\infty} e^{ka\xi}(f^{(k)}(t,x;\xi) - \partial_\xi w^{(k)}(t,x;\xi))d\xi \\
w^{(k)}|_{t=0} = 0, \quad w^{(k)} \in S(\mathbb{R}_z^+)
\end{cases}
\tag{3.30}
$$

for any $k \leq -1$.

The boundary conditions of $w^{(k)}(t,x;z)$ at $\{z = 0\}$ given in (3.29) and (3.30) can be expressed as:

$$
\begin{cases}
\int_0^{+\infty} e^{-ka\xi}(f^{(k)} - kaw^{(k)})(t,x;\xi)d\xi + b_1^{(k)}(t,x) + kab_0^{(k)}(t,x) = 0, \quad k \geq 1, \\
\int_0^{+\infty} e^{ka\xi}(f^{(k)} + kaw^{(k)})(t,x;\xi)d\xi + b_1^{(k)}(t,x) - kab_0^{(k)}(t,x) = 0, \quad k \leq -1.
\end{cases}
\tag{3.31}
$$

In terms of the transformation

$$
Y^{(k)}(t,x;z) = \begin{cases}
\int_z^{+\infty} e^{ka(z-\xi)}w^{(k)}(t,x;\xi)d\xi, \quad k \geq 1 \\
\int_z^{+\infty} e^{ka(\xi-z)}w^{(k)}(t,x;\xi)d\xi, \quad k \leq -1,
\end{cases}
\tag{3.32}
$$

problems (3.29), (3.30), and (3.31) can be reformulated as

$$
\begin{cases}
(\partial_t + a_1\partial_x)Y^{(k)} - a_4^2(\partial_z^2 - k^2a^2)Y^{(k)} + z(a_2\partial_z + ika_3)Y^{(k)} \\
\qquad + ka_5\int_z^{+\infty} e^{ka(z-\xi)}Y^{(k)}(t,x;\xi)d\xi + ka_6\int_0^z e^{ka(\xi-z)}Y^{(k)}(t,x;\xi)d\xi \\
\quad = G^{(k)}(t,x;z) \\
Y^{(k)}|_{z=0} = W_0^{(k)}(t,x), \quad Y^{(k)} \in S(\mathbb{R}_z^+) \\
Y^{(k)}|_{t=0} = 0
\end{cases}
\tag{3.33}
$$

for any $k \geq 1$, and

$$(3.34)\quad \begin{cases} (\partial_t + a_1\partial_x)Y^{(k)} - a_4^2(\partial_z^2 - k^2a^2)Y^{(k)} + z(a_2\partial_z + ika_3)Y^{(k)} \\ \qquad + ka_5\displaystyle\int_z^{+\infty} e^{ka(\xi-z)}Y^{(k)}(t,x;\xi)d\xi + ka_6\int_0^z e^{ka(z-\xi)}Y^{(k)}(t,x;\xi)d\xi \\ \quad = G^{(k)}(t,x;z) \\ Y^{(k)}|_{z=0} = W_0^{(k)}(t,x), \quad Y^{(k)} \in S(\mathbb{R}_z^+) \\ Y^{(k)}|_{t=0} = 0 \end{cases}$$

for any $k \leq -1$, where

$$\begin{cases} G^{(k)} = \displaystyle\int_z^{+\infty} e^{ka(z-\xi)}g^{(k)}(t,x;\xi)d\xi + \frac{aa_2 - ia_3}{4ka^2}\Big(\int_0^{+\infty} e^{-ka(z+\xi)}f^{(k)}(t,x;\xi)d\xi \\ \qquad + \displaystyle\int_z^{+\infty} e^{ka(z-\xi)}f^{(k)}(t,x;\xi)d\xi + \int_0^z e^{ka(\xi-z)}f^{(k)}(t,x;\xi)d\xi\Big) \\ \qquad - \dfrac{aa_2 + ia_3}{2a}\displaystyle\int_z^{\infty}(\xi - z)e^{ka(z-\xi)}f^{(k)}(t,x;\xi)d\xi + \frac{aa_2 - ia_3}{2ka^2}e^{-kaz}b_1^{(k)}(t,x) \\ W_0^{(k)}(t,x) = b_0^{(k)} + \dfrac{1}{ka}b_1^{(k)} + \dfrac{1}{ka}\displaystyle\int_0^{+\infty} e^{-ka\xi}f^{(k)}(t,x;\xi)d\xi \end{cases}$$

for any $k \geq 1$,

$$\begin{cases} G^{(k)} = \displaystyle\int_z^{+\infty} e^{ka(\xi-z)}g^{(k)}(t,x;\xi)d\xi - \frac{aa_2 + ia_3}{4ka^2}\Big(\int_0^{+\infty} e^{ka(z+\xi)}f^{(k)}(t,x;\xi)d\xi \\ \qquad + \displaystyle\int_z^{\infty} e^{ka(\xi-z)}f^{(k)}(t,x;\xi)d\xi + \int_0^z e^{ka(z-\xi)}f^{(k)}(t,x;\xi)d\xi\Big) \\ \qquad - \dfrac{aa_2 - ia_3}{2a}\displaystyle\int_z^{\infty}(\xi - z)e^{ka(\xi-z)}f^{(k)}(t,x;\xi)d\xi - \frac{aa_2 + ia_3}{2ka^2}e^{kaz}b_1^{(k)}(t,x) \\ W_0^{(k)}(t,x) = b_0^{(k)} - \dfrac{1}{ka}b_1^{(k)} - \dfrac{1}{ka}\displaystyle\int_0^{+\infty} e^{ka\xi}f^{(k)}(t,x;\xi)d\xi \end{cases}$$

for any $k \leq -1$, and

$$\begin{cases} a_5 = a_t + a_1a_x + \dfrac{aa_2 + ia_3}{2}, \quad a_6 = -\dfrac{1}{2}(aa_2 + ia_3), \quad k \geq 1, \\ a_5 = -\Big(a_t + a_1a_x + \dfrac{aa_2 - ia_3}{2}\Big), \quad a_6 = \dfrac{1}{2}(aa_2 + ia_3), \quad k \leq -1. \end{cases}$$

The compatibility conditions for problems (3.33) and (3.34) can be easily formulated in a classical way. For example, the zero-th order compatibility condition for (3.33) is

$$W_0^{(k)}(0,x) = 0$$

and the first order one is

$$G^{(k)}(0,x;0) = (\partial_t W_0^{(k)})(0,x).$$

It is not difficult to verify that compatibility conditions for problems (3.33) and (3.34) are implied directly by those for the problem (3.1).

The problem (3.33) shall be solved in the following steps, and (3.34) can be studied similarly.

3.2.1. Step 1: Homogenization of the initial data. Let $\chi(z) \in C_0^\infty(\mathbb{R})$ be an arbitrary smooth function with compact support and $\chi(0) = 1$. Then, the function

$$Y_0^{(k)}(t, x; z) = \chi(z) W_0^{(k)}(t, x)$$

satisfies the initial and boundary conditions given in (3.33) due to the compatibility conditions.

Using the transformation $\tilde{Y}^{(k)} = Y^{(k)} - Y_0^{(k)}$, if necessary, it suffices to study the problem (3.33) for the special case $Y^{(k)}|_{z=0} \equiv 0$, which will be assumed in what follows.

3.2.2. Step 2: Construction of approximation solutions. We construct an approximate solution sequence $\{Y_n^{(k)}\}_{n\geq 1}$ of (3.33) by solving the following problem for each $n \geq 1$:

(3.35)
$$\begin{cases} (\partial_t + a_1\partial_x)Y_n^{(k)} - a_4^2(\partial_z^2 - k^2a^2)Y_n^{(k)} + z(a_2\partial_z + ika_3)Y_n^{(k)} - \dfrac{1}{n}\partial_x^2 Y_n^{(k)} \\ \quad +ka_5 \displaystyle\int_z^{+\infty} e^{ka(z-\xi)} Y_{n-1}^{(k)}(t, x; \xi) d\xi + ka_6 \int_0^z e^{ka(\xi - z)} Y_{n-1}^{(k)}(t, x; \xi) d\xi = G^{(k)}(t, x; z) \\ Y_n^{(k)}|_{z=0} = 0, \quad Y_n^{(k)} \in S(\mathbb{R}_z^+) \\ Y_n^{(k)}|_{t=0} = 0, \end{cases}$$

where $Y_0^{(k)}(t, x, z) \equiv 0$.

It remains to study properties of the sequence $\{Y_n^{(k)}\}_{n\geq 1}$. Most of this part will follow the idea of Xin and Yanagisawa in section 4 of [12] for studying the linearized Prandtl equation.

In what follows, for any $j \in \mathbb{N}$, we shall denote by C_j a constant depending only upon the bounds of derivatives of coefficients appeared in (3.35) up to order j.

3.2.3. Step 3: The boundedness of { ( )} ≥1 in ²− norm. Denote by $\langle z \rangle = (1 + z^2)^{\frac{1}{2}}$, and $\Omega = \mathbb{R}_+^2 = \{(x, z) \in \mathbb{R}^2 |\ z > 0\}$. For any fixed integer $l \in \mathbb{N}$, multiplying (3.35) by $< z >^{2l}\ \overline{Y}_n^{(k)}$, and integrating the resulting equation over Ω, one gets

$$\begin{aligned} &\frac{d}{dt}\int_\Omega \langle z\rangle^{2l} |Y_n^{(k)}|^2 dxdz - \int_\Omega \partial_x a_1 \langle z\rangle^{2l} |Y_n^{(k)}|^2 dxdz \\ &\quad + 2k^2 \int_\Omega a^2 a_4^2 \langle z\rangle^{2l} |Y_n^{(k)}|^2 dxdz + 4l\mathcal{R}\int_\Omega a_4^2 z \langle z\rangle^{2(l-1)} Y_n^{(k)} \partial_z \overline{Y}_n^{(k)} dxdz \\ &\quad + 2\int_\Omega a_4^2 \langle z\rangle^{2l} |\partial_z Y_n^{(k)}|^2 dxdz + \frac{2}{n}\int_\Omega \langle z\rangle^{2l} |\partial_x Y_n^{(k)}|^2 dxdz + 2kA_0 \\ &\quad = 2\mathcal{R}\int_\Omega \langle z\rangle^{2l} G^{(k)} \overline{Y}_n^{(k)} dxdz, \end{aligned} \tag{3.36}$$

where $\mathcal{R}(\cdot)$ denotes the real part of the related functions, and

$$\begin{aligned} A_0 = \mathcal{R}\int_\Omega \langle z\rangle^{2l} \overline{Y}_n^{(k)} \Big(& a_5 \int_z^{+\infty} e^{ka(z-\xi)} Y_{n-1}^{(k)}(t, x; \xi) d\xi \\ & + a_6 \int_0^z e^{ka(\xi - z)} Y_{n-1}^{(k)}(t, x; \xi) d\xi \Big) dxdz. \end{aligned} \tag{3.37}$$

By a simple computation, one deduces

$$
\begin{aligned}
&\int_\Omega \langle z\rangle^{2l} |\overline{Y}_n^{(k)} \int_0^z e^{ka(\xi-z)} Y_{n-1}^{(k)}(t,x;\xi) d\xi| dxdz \\
&\quad \le \frac{1}{2}\int_\Omega \int_0^z \langle z\rangle^{2l} e^{ka(\xi-z)} (|Y_n^{(k)}(t,x;z)|^2 + |Y_{n-1}^{(k)}(t,x;\xi)|^2) d\xi dxdz \\
&\quad \le \frac{c(l,a_0)}{k}\int_\Omega \langle z\rangle^{2l} (|Y_n^{(k)}|^2 + |Y_{n-1}^{(k)}|^2) dxdz,
\end{aligned}
\tag{3.38}
$$

where $c(l,a_0)$ is a constant depending only upon $l\in\mathbb{N}$ and a_0 satisfying $0<a_0\le a(t,x)$.

Similarly, we have

$$
\begin{aligned}
&\int_\Omega \langle z\rangle^{2l} |\overline{Y}_n^{(k)} \int_z^{+\infty} e^{ka(z-\xi)} Y_{n-1}^{(k)}(t,x;\xi) d\xi| dxdz \\
&\quad \le \frac{c(l,a_0)}{k}\int_\Omega \langle z\rangle^{2l} (|Y_n^{(k)}|^2 + |Y_{n-1}^{(k)}|^2) dxdz.
\end{aligned}
\tag{3.39}
$$

Substituting (3.38) and (3.39) into (3.37) shows that

$$
|A_0| \le \frac{C_0}{k}\int_\Omega \langle z\rangle^{2l} (|Y_n^{(k)}|^2 + |Y_{n-1}^{(k)}|^2) dxdz. \tag{3.40}
$$

Combining (3.40) and (3.36), we get

$$
\begin{aligned}
&\frac{d}{dt}\int_\Omega \langle z\rangle^{2l} |Y_n^{(k)}|^2 dxdz + \int_\Omega \langle z\rangle^{2l} (|\partial_z Y_n^{(k)}|^2 + k^2|Y_n^{(k)}|^2) dxdz \\
&\quad \le C_0 \int_\Omega \langle z\rangle^{2l} (|Y_n^{(k)}|^2 + |Y_{n-1}^{(k)}|^2) dxdz + \int_\Omega \langle z\rangle^{2l} |G^{(k)}|^2 dxdz.
\end{aligned}
\tag{3.41}
$$

To study (3.41), first we have to note the following lemma.

LEMMA 3.4. *Given nonnegative functions $f\in C^0[0,\infty)$, $b_n\in C^0[0,\infty)$, $a_n\in C^1[0,\infty)$ satisfying $a_n(0)\le a$ for a constant a for any $n\in\mathbb{N}$, if we have*

$$
a_n'(t) + b_n(t) \le C_0(a_n(t) + a_{n-1}(t)) + f(t) \quad \forall n\ge 1
$$

for a constant $C_0\ge 0$ independent of n, then the estimate

$$
a_n(t) + \int_0^t e^{C_0(t-s)} b_n(s) ds \le a e^{2C_0 t} + \int_0^t e^{2C_0(t-s)} f(s) ds
$$

holds for any $n\in\mathbb{N}$.

This Gronwall type estimate can be obtained by induction on n.

By using Lemma 3.4 in (3.42), we immediately refer to the following lemma.

LEMMA 3.5. *Denote by $\langle z\rangle = (1+z^2)^{\frac{1}{2}}$ and $\Omega = \mathbb{R}_+^2 = \{(x,z)\in\mathbb{R}^2 |\ z>0\}$. For any fixed integer $l\in\mathbb{N}$, there is a constant C_0 depending only upon l and a_0 satisfying $0<a_0\le a(t,x)$, such that the following estimate*

$$
\begin{aligned}
&\max_{0\le t\le T}\int_\Omega \langle z\rangle^{2l} |Y_n^{(k)}|^2 dxdz + \int_0^T\int_\Omega \langle z\rangle^{2l} (|\partial_z Y_n^{(k)}|^2 + k^2|Y_n^{(k)}|^2) dxdzdt \\
&\quad \le \int_0^T\int_\Omega e^{2C_0(T-t)} \langle z\rangle^{2l} |G^{(k)}|^2 dxdzdt
\end{aligned}
\tag{3.42}
$$

holds for any $T\ge 0$ and $n\in\mathbb{N}$.

3.2.4. Step 4: Estimates of spatial tangential derivatives ( ). For any $\alpha \in \mathbb{N}$, set $Y^{(k)}_{n,\alpha} = \partial_x^\alpha Y^{(k)}_n$. From (3.35), we know that $Y^{(k)}_{n,\alpha}$ satisfies the following problem:

(3.43)
$$\begin{cases} (\partial_t + a_1\partial_x)Y^{(k)}_{n,\alpha} - a_4^2(\partial_z^2 - k^2a^2)Y^{(k)}_{n,\alpha} + z(a_2\partial_z + ika_3)Y^{(k)}_{n,\alpha} - \dfrac{1}{n}\partial_x^2 Y^{(k)}_{n,\alpha} \\ \qquad +ka_5\displaystyle\int_z^{+\infty} e^{ka(z-\xi)}Y^{(k)}_{n-1,\alpha}(\cdot,\xi)d\xi + ka_6\int_0^z e^{ka(\xi-z)}Y^{(k)}_{n-1,\alpha}(\cdot,\xi)d\xi + R_\alpha = \partial_x^\alpha G^{(k)} \\ Y^{(k)}_{n,\alpha}|_{z=0} = 0, \quad Y^{(k)}_{n,\alpha} \in S(\mathbb{R}^+_z) \\ Y^{(k)}_{n,\alpha}|_{t=0} = 0, \end{cases}$$

where

$$\begin{aligned} R_\alpha &= [\partial_x^\alpha, a_1\partial_x - a_4^2(\partial_z^2 - k^2a^2) + z(a_2\partial_z + ika_3)]Y^{(k)}_n \\ &\quad + k\sum_{0<j\le\alpha}\binom{\alpha}{j}\Big(\partial_x^j a_5\int_z^{+\infty}\partial_x^{\alpha-j}(e^{ka(z-\xi)}Y^{(k)}_{n-1}(\cdot,\xi))d\xi \\ &\quad + \partial_x^j a_6\int_0^z \partial_x^{\alpha-j}(e^{ka(\xi-z)}Y^{(k)}_{n-1}(\cdot,\xi))d\xi\Big) \\ &\quad + ka_5\int_z^{+\infty}[\partial_x^\alpha, e^{ka(z-\xi)}]Y^{(k)}_{n-1}(t,x;\xi)d\xi + ka_6\int_0^z[\partial_x^\alpha, e^{ka(\xi-z)}]Y^{(k)}_{n-1}(t,x;\xi)d\xi. \end{aligned}$$

Similar to (3.41), by multiplying the equation in (3.43) by $\langle z\rangle^{2l}\overline{Y}^{(k)}_{n,\alpha}$ for any fixed $l \in \mathbb{N}$, and integrating the resulting equation over Ω, we obtain

$$\begin{aligned} &\frac{d}{dt}\int_\Omega \langle z\rangle^{2l}|Y^{(k)}_{n,\alpha}|^2dxdz + \int_\Omega\langle z\rangle^{2l}(|\partial_z Y^{(k)}_{n,\alpha}|^2 + k^2|Y^{(k)}_{n,\alpha}|^2)dxdz \\ &\qquad \le C_0\int_\Omega\langle z\rangle^{2l}(|Y^{(k)}_{n,\alpha}|^2 + |Y^{(k)}_{n-1,\alpha}|^2)dxdz + \int_\Omega < z >^{2l} |\partial_x^\alpha G^{(k)}|^2dxdz \\ &\qquad\quad -2\mathcal{R}\int_\Omega\langle z\rangle^{2l}R_\alpha\overline{Y}^{(k)}_{n,\alpha}dxdz. \end{aligned} \tag{3.44}$$

On the other hand, we have

$$\begin{aligned} &\Big|\int_\Omega\langle z\rangle^{2l}R_\alpha\overline{Y}^{(k)}_{n,\alpha}dxdz\Big| \\ &\quad \le C_0\int_\Omega\langle z\rangle^{2l}|Y^{(k)}_{n,\alpha}|^2dxdz + \epsilon\int_\Omega\langle z\rangle^{2l}(|\partial_z Y^{(k)}_{n,\alpha}|^2 + k^2|Y^{(k)}_{n,\alpha}|^2)dxdz \\ &\qquad + \sum_{0<j\le\alpha}\frac{C_j}{\epsilon}\int_\Omega\langle z\rangle^{2l}(|\partial_z Y^{(k)}_{n,\alpha-j}|^2 + k^2|Y^{(k)}_{n,\alpha-j}|^2 + \langle z\rangle^2|Y^{(k)}_{n,\alpha-j}|^2)dxdz \\ &\qquad + \sum_{0<j\le\alpha}C_j\int_\Omega\langle z\rangle^{2l}(\langle z\rangle^2|\partial_z Y^{(k)}_{n,\alpha-j}|^2 + |Y^{(k)}_{n-1,\alpha-j}|^2)dxdz \end{aligned} \tag{3.45}$$

for any $\epsilon > 0$.

Substituting (3.45) into (3.44), and letting ϵ be small, we obtain

$$\begin{aligned}
&\frac{d}{dt}\int_\Omega \langle z\rangle^{2l}\,|Y_{n,\alpha}^{(k)}|^2 dxdz + \int_\Omega \langle z\rangle^{2l}(|\partial_z Y_{n,\alpha}^{(k)}|^2 + k^2|Y_{n,\alpha}^{(k)}|^2)dxdz \\
&\qquad \le C_0\int_\Omega \langle z\rangle^{2l}(|Y_{n,\alpha}^{(k)}|^2 + |Y_{n-1,\alpha}^{(k)}|^2)dxdz + \int_\Omega \langle z\rangle^{2l}|\partial_x^\alpha G^{(k)}|^2 dxdz \\
&\qquad\quad + \sum_{0<j\le\alpha} C_j\Big(\int_\Omega \langle z\rangle^{2(l+1)}(|\partial_z Y_{n,\alpha-j}^{(k)}|^2 + |Y_{n,\alpha-j}^{(k)}|^2)dxdz \\
&\qquad\quad + \int_\Omega \langle z\rangle^{2l}(|Y_{n-1,\alpha-j}^{(k)}|^2 + k^2|Y_{n,\alpha-j}^{(k)}|^2)dxdz\Big).
\end{aligned}$$

By using Lemma 3.4 and induction on $\alpha\in\mathbb{N}$, Lemma 3.6 follows.

LEMMA 3.6. *For the problem* (3.35), *let* $Y_{n,\alpha}^{(k)} = \partial_x^\alpha Y_n^{(k)}$; *the following estimate*

$$\begin{aligned}
&\max_{0\le t\le T}\int_\Omega \langle z\rangle^{2l}|Y_{n,\alpha}^{(k)}|^2 dxdz + \int_0^T\int_\Omega \langle z\rangle^{2l}(|\partial_z Y_{n,\alpha}^{(k)}|^2 + k^2|Y_{n,\alpha}^{(k)}|^2)dxdzdt \\
&\qquad \le C(T)\sum_{j=0}^{\alpha}\int_0^T\int_\Omega \langle z\rangle^{2(l+\alpha-j)}|\partial_x^j G^{(k)}|^2 dxdzdt
\end{aligned} \tag{3.46}$$

holds for any $\alpha\in\mathbb{N}$.

3.2.5. Step 5: Estimates of derivatives , ( ). For any fixed integer $j\ge 0$, set $V_{n,j}^{(k)} = \partial_t^j Y_n^{(k)}$.

From (3.35) we know that $V_{n,j}^{(k)}$ satisfies the following problem:

(3.47)
$$\begin{cases}
(\partial_t + a_1\partial_x)V_{n,j}^{(k)} - a_4^2(\partial_z^2 - k^2a^2)V_{n,j}^{(k)} + z(a_2\partial_z + ika_3)V_{n,j}^{(k)} - \frac{1}{n}\partial_x^2 V_{n,j}^{(k)} \\
\quad + ka_5\displaystyle\int_z^{+\infty} e^{ka(z-\xi)}V_{n-1,j}^{(k)}(\cdot,\xi)d\xi + ka_6\int_0^z e^{ka(\xi-z)}V_{n-1,j}^{(k)}(\cdot,\xi)d\xi + Q_j = \partial_t^j G^{(k)} \\
V_{n,j}^{(k)}|_{z=0} = 0, \quad V_{n,j}^{(k)}\in S(\mathbb{R}_z^+) \\
V_{n,j}^{(k)}|_{t=0} = V_{n,j,0}^{(k)}(x,z),
\end{cases}$$

where

$$\begin{aligned}
Q_j &= [\partial_t^j, a_1\partial_x - a_4^2(\partial_z^2 - k^2a^2) + z(a_2\partial_z + ika_3)]Y_n^{(k)} \\
&\quad + k\sum_{0<m\le j}\binom{j}{m}\Big(\partial_t^m a_5\int_z^{+\infty}\partial_t^{j-m}(e^{ka(z-\xi)}Y_{n-1}^{(k)}(t,x;\xi))d\xi \\
&\quad + \partial_t^m a_6\int_0^z \partial_t^{j-m}(e^{ka(\xi-z)}Y_{n-1}^{(k)}(t,x;\xi))d\xi\Big) \\
&\quad + ka_5\int_z^{+\infty}[\partial_t^j, e^{ka(z-\xi)}]Y_{n-1}^{(k)}(t,x;\xi)d\xi + ka_6\int_0^z[\partial_t^j, e^{ka(\xi-z)}]Y_{n-1}^{(k)}(t,x;\xi)d\xi
\end{aligned}$$

and

(3.48)
$$\begin{aligned}
V_{n,j,0}^{(k)} &= \partial_t^{j-1}G^{(k)}|_{t=0} - \Big(a_1\partial_x - a_4^2(\partial_z^2 - k^2a^2) + z(a_2\partial_z + ika_3) - \frac{1}{n}\partial_x^2\Big)V_{n,j-1,0}^{(k)} \\
&\quad - ka_5\int_z^{+\infty} e^{ka(z-\xi)}V_{n-1,j-1,0}^{(k)}d\xi - ka_6\int_0^z e^{ka(\xi-z)}V_{n-1,j-1,0}^{(k)}d\xi - Q_{j-1}|_{t=0}
\end{aligned}$$

is defined by induction on j with $V^{(k)}_{n,1,0} = G^{(k)}(0, x, z)$.

Multiplying (3.47) by $\langle z\rangle^{2l}\overline{V}^{(k)}_{n,j}$, and integrating the resulting equation over Ω, one gets

$$\begin{aligned}
&\frac{d}{dt}\int_\Omega \langle z\rangle^{2l}\,|V^{(k)}_{n,j}|^2 dxdz + \int_\Omega \langle z\rangle^{2l}(|\partial_z V^{(k)}_{n,j}|^2 + k^2|V^{(k)}_{n,j}|^2)dxdz\\
&\qquad\le C_0\int_\Omega \langle z\rangle^{2l}(|V^{(k)}_{n,j}|^2 + |V^{(k)}_{n-1,j}|^2)dxdz + \int_\Omega \langle z\rangle^{2l}|\partial_t^j G^{(k)}|^2 dxdz\\
&\qquad\quad -2\mathcal{R}\int_\Omega \langle z\rangle^{2l} Q_j \overline{V}^{(k)}_{n,j} dxdz.
\end{aligned}\tag{3.49}$$

On the other hand, we have

$$\begin{aligned}
&\left|\int_\Omega \langle z\rangle^{2l}\,Q_j\overline{V}^{(k)}_{n,j}dxdz\right|\\
&\qquad\le C_0\int_\Omega \langle z\rangle^{2l}|V^{(k)}_{n,j}|^2 dxdz + \epsilon\int_\Omega \langle z\rangle^{2l}(|\partial_z V^{(k)}_{n,j}|^2 + k^2|V^{(k)}_{n,j}|^2)dxdz\\
&\qquad\quad + \sum_{0<m\le j}\frac{C_m}{\epsilon}\int_\Omega \langle z\rangle^{2l}(|\partial_z V^{(k)}_{n,j-m}|^2 + k^2|V^{(k)}_{n,j-m}|^2\\
&\qquad\quad + \langle z\rangle^2|V^{(k)}_{n,j-m}|^2 + |V^{(k)}_{n-1,j-m}|^2)dxdz\\
&\qquad\quad + \sum_{0<m\le j} C_m\int_\Omega \langle z\rangle^{2l}(\langle z\rangle^2|\partial_z V^{(k)}_{n,j-m}|^2 + |\partial_x V^{(k)}_{n,j-m}|^2)dxdz
\end{aligned}\tag{3.50}$$

for any $\epsilon > 0$.

Substituting (3.50) into (3.49), and letting ϵ be small, we obtain

$$\begin{aligned}
&\frac{d}{dt}\int_\Omega \langle z\rangle^{2l}\,|V^{(k)}_{n,j}|^2 dxdz + \int_\Omega \langle z\rangle^{2l}(|\partial_z V^{(k)}_{n,j}|^2 + k^2|V^{(k)}_{n,j}|^2)dxdz\\
&\quad\le C_0\int_\Omega \langle z\rangle^{2l}(|V^{(k)}_{n,j}|^2 + |V^{(k)}_{n-1,j}|^2)dxdz + \int_\Omega \langle z\rangle^{2l}|\partial_t^j G^{(k)}|^2 dxdz\\
&\qquad + \sum_{0<m\le j} C_m\Big(\int_\Omega \langle z\rangle^{2(l+1)}(|\partial_z V^{(k)}_{n,j-m}|^2 + |V^{(k)}_{n,j-m}|^2)dxdz\\
&\qquad + \int_\Omega \langle z\rangle^{2l}(|V^{(k)}_{n-1,j-m}|^2 + k^2|V^{(k)}_{n,j-m}|^2 + |\partial_x V^{(k)}_{n,j-m}|^2)dxdz\Big).
\end{aligned}\tag{3.51}$$

Thus, to complete the estimate on $V^{(k)}_{n,j}$, we should study $\{\partial_x V^{(k)}_{n,m}\}_{m=1}^{j-1}$ first.

It follows from (3.47) that $\partial_x^p V^{(k)}_{n,j} = \partial_x^p\partial_t^j Y^{(k)}_n$ satisfies the following problem:

$$\begin{cases}
(\partial_t + a_1\partial_x - a_4^2(\partial_z^2 - k^2a^2) + z(a_2\partial_z + ika_3) - \frac{1}{n}\partial_x^2)\partial_x^p V^{(k)}_{n,j}\\
\qquad + ka_5\displaystyle\int_z^{+\infty} e^{ka(z-\xi)}\partial_x^p V^{(k)}_{n-1,j}d\xi + ka_6\int_0^z e^{ka(\xi-z)}\partial_x^p V^{(k)}_{n-1,j}d\xi + Q_{j,p} = \partial_x^p\partial_t^j G^{(k)}\\
\partial_x^p V^{(k)}_{n,j}|_{z=0} = 0, \quad \partial_x^p V^{(k)}_{n,j} \in S(\mathbb{R}_z^+)\\
\partial_x^p V^{(k)}_{n,j}|_{t=0} = \partial_x^p V^{(k)}_{n,j,0}(x,z),
\end{cases}\tag{3.52}$$

where $V^{(k)}_{n,j,0}(x,z)$ is given in (3.48), and

$$Q_{j,p} = \partial_x^p Q_j + [\partial_x^p, a_1\partial_x - a_4^2(\partial_z^2 - k^2a^2) + z(a_2\partial_z + ika_3)]V^{(k)}_{n,j}$$
$$+ k \sum_{0<m\le p} \binom{p}{m} \Big(\partial_x^m a_5 \int_z^{+\infty} \partial_x^{p-m}(e^{ka(z-\xi)}V^{(k)}_{n-1,j}(t,x;\xi))d\xi$$
$$+ \partial_x^m a_6 \int_0^z \partial_x^{p-m}(e^{ka(\xi-z)}V^{(k)}_{n-1,j}(t,x;\xi))d\xi\Big)$$
$$+ ka_5 \int_z^{+\infty} [\partial_x^p, e^{ka(z-\xi)}]V^{(k)}_{n-1,j}(t,x;\xi)d\xi + ka_6 \int_0^z [\partial_x^p, e^{ka(\xi-z)}]V^{(k)}_{n-1,j}(t,x;\xi)d\xi$$

with Q_j being given in (3.47).

Multiplying (3.52) by $< z >^{2l} \partial_x^p \overline{V}^{(k)}_{n,j}$, and integrating the resulting equation over Ω, one gets

$$\frac{d}{dt}\int_\Omega \langle z\rangle^{2l}\,|\partial_x^p V^{(k)}_{n,j}|^2 dxdz + \int_\Omega \langle z\rangle^{2l}(|\partial_z\partial_x^p V^{(k)}_{n,j}|^2 + k^2|\partial_x^p V^{(k)}_{n,j}|^2)dxdz$$
$$\le C_0 \int_\Omega \langle z\rangle^{2l}(|\partial_x^p V^{(k)}_{n,j}|^2 + |\partial_x^p V^{(k)}_{n-1,j}|^2)dxdz \tag{3.53}$$
$$+ \int_\Omega \langle z\rangle^{2l}|\partial_x^p\partial_t^j G^{(k)}|^2 dxdz - 2\mathcal{R}\int_\Omega \langle z\rangle^{2l} Q_{j,p}\partial_x^p \overline{V}^{(k)}_{n,j}dxdz.$$

A direct computation shows

$$\left|\int_\Omega \langle z\rangle^{2l} Q_{j,p}\partial_x^p\overline{V}^{(k)}_{n,j}dxdz\right|$$
$$\le C_0\int_\Omega \langle z\rangle^{2l}|\partial_x^p V^{(k)}_{n,j}|^2 dxdz + \epsilon\int_\Omega \langle z\rangle^{2l}(|\partial_z\partial_x^p V^{(k)}_{n,j}|^2 + k^2|\partial_x^p V^{(k)}_{n,j}|^2)dxdz$$
$$+ \sum_{0<m\le j, q\le p} \frac{C_{q+m}}{\epsilon}\int_\Omega \langle z\rangle^{2(l+1)}(|\partial_z\partial_x^q V^{(k)}_{n,j-m}|^2 + |\partial_x^q V^{(k)}_{n,j-m}|^2)dxdz$$
$$+ k^2 \sum_{0<m\le j, q\le p}\frac{C_{q+m}}{\epsilon}\int_\Omega \langle z\rangle^{2l}|\partial_x^q V^{(k)}_{n,j-m}|^2 dxdz$$
$$+ \sum_{q<p}\frac{C_q}{\epsilon}\int_\Omega \langle z\rangle^{2l}(|\partial_z\partial_x^q V^{(k)}_{n,j}|^2 + k^2|\partial_x^q V^{(k)}_{n,j}|^2)dxdz$$
$$+ \sum_{q<p} C_q \int_\Omega \langle z\rangle^{2l}(\langle z\rangle^2|\partial_z\partial_x^q V^{(k)}_{n,j}|^2 + |\partial_x^q V^{(k)}_{n-1,j}|^2)dxdz$$
$$+ \sum_{1\le q\le p} C_q\int_\Omega \langle z\rangle^{2l}|\partial_x^q V^{(k)}_{n,j}|^2 dxdz + \sum_{0<m\le j} C_m\int_\Omega \langle z\rangle^{2l}|\partial_x^{q+1} V^{(k)}_{n,j-m}|^2 dxdz$$
$$+ \sum_{0<m\le j, q\le p} C_{q+m}\int_\Omega \langle z\rangle^{2l}|\partial_x^q V^{(k)}_{n-1,j-m}|^2 dxdz.$$

Thus, from (3.53) one has

$$\begin{aligned}
&\frac{d}{dt}\int_\Omega \langle z\rangle^{2l}|\partial_x^p V_{n,j}^{(k)}|^2 dxdz + \int_\Omega \langle z\rangle^{2l}(|\partial_z\partial_x^p V_{n,j}^{(k)}|^2 + k^2|\partial_x^p V_{n,j}^{(k)}|^2)dxdz \\
&\quad \le C_0\int_\Omega \langle z\rangle^{2l}(|\partial_x^p V_{n,j}^{(k)}|^2 + |\partial_x^p V_{n-1,j}^{(k)}|^2)dxdz + \int_\Omega \langle z\rangle^{2l}|\partial_x^p\partial_t^j G^{(k)}|^2 dxdz \\
&\qquad + C_{j+p}\left\{\sum_{0<m\le j, q\le p}\left(\int_\Omega \langle z\rangle^{2(l+1)}(|\partial_z\partial_x^q V_{n,j-m}^{(k)}|^2 + |\partial_x^q V_{n,j-m}^{(k)}|^2)dxdz\right.\right. \\
&\qquad \left.\left. + \int_\Omega \langle z\rangle^{2l}(k^2|\partial_x^q V_{n,j-m}^{(k)}|^2 + |\partial_x^q V_{n-1,j-m}^{(k)}|^2)dxdz\right)\right\} \\
&\qquad + C_p\sum_{q<p}\int_\Omega \langle z\rangle^{2l}(\langle z\rangle^2|\partial_z\partial_x^q V_{n,j}^{(k)}|^2 + k^2|\partial_x^q V_{n,j}^{(k)}|^2 + |\partial_x^q V_{n-1,j}^{(k)}|^2)dxdz \\
&\qquad + C_j\sum_{0<m\le j}\int_\Omega \langle z\rangle^{2l}|\partial_x^{q+1}V_{n,j-m}^{(k)}|^2 dxdz.
\end{aligned} \tag{3.54}$$

By using Lemmas 3.4 and 3.5 in (3.51) for the case $j=1$, it follows that $V_{n,1}^{(k)} = \partial_t V_n^{(k)}$ satisfies

$$\begin{aligned}
&\max_{0\le t\le T}\int_\Omega \langle z\rangle^{2l}|V_{n,1}^{(k)}|^2 dxdz + \int_0^T\int_\Omega \langle z\rangle^{2l}(|\partial_z V_{n,1}^{(k)}|^2 + k^2|V_{n,1}^{(k)}|^2)dxdzdt \\
&\qquad \le C(T)\int_0^T\int_\Omega \langle z\rangle^{2l}(|\partial_x G^{(k)}|^2 + |\partial_t G^{(k)}|^2 + |G^{(k)}|^2)dxdzdt.
\end{aligned} \tag{3.55}$$

Using Lemma 3.6 and (3.55) in (3.54) for the case $j=1$ and $p=1$, and using Lemma 3.4, it follows that $\partial_x V_{n,1}^{(k)} = \partial_t\partial_x Y_n^{(k)}$ satisfies

$$\begin{aligned}
&\max_{0\le t\le T}\int_\Omega < z >^{2l} |\partial_x V_{n,1}^{(k)}|^2 dxdz + \int_0^T\int_\Omega \langle z\rangle^{2l}(|\partial_z\partial_x V_{n,1}^{(k)}|^2 + k^2|\partial_x V_{n,1}^{(k)}|^2)dxdzdt \\
&\quad \le C(T)\left(\sum_{j=0}^{2}\int_0^T\int_\Omega \langle z\rangle^{2(l+2-j)}|\partial_x^j G^{(k)}|^2 dxdzdt\right. \\
&\qquad \left. + \sum_{j=0}^{1}\int_0^T\int_\Omega \langle z\rangle^{2(l+1-j)}|\partial_x^j\partial_t G^{(k)}|^2 dxdzdt\right).
\end{aligned} \tag{3.56}$$

Employing (3.56), (3.46), and (3.55) in (3.51) for the case $j=2$, and using Lemma 3.4 we deduce that $V_{n,2}^{(k)} = \partial_t^2 Y_n^{(k)}$ satisfies

$$\begin{aligned}
&\max_{0\le t\le T}\int_\Omega \langle z\rangle^{2l}|V_{n,2}^{(k)}|^2 dxdz + \int_0^T\int_\Omega \langle z\rangle^{2l}(|\partial_z V_{n,2}^{(k)}|^2 + k^2|V_{n,2}^{(k)}|^2)dxdzdt \\
&\quad \le C(T)\left(\int_\Omega \langle z\rangle^{2l}|V_{n,2,0}^{(k)}|^2 dxdz + \sum_{|\alpha|\le 2}\int_0^T\int_\Omega \langle z\rangle^{2(l+2-|\alpha|)}|\partial_{t,x}^\alpha G^{(k)}|^2 dxdzdt\right),
\end{aligned}$$

which implies

$$\begin{aligned}
&\max_{0\le t\le T}\int_\Omega \langle z\rangle^{2l}|V_{n,2}^{(k)}|^2dxdz+\int_0^T\int_\Omega\langle z\rangle^{2l}(|\partial_zV_{n,2}^{(k)}|^2+k^2|V_{n,2}^{(k)}|^2)dxdzdt\\
&\quad\le C\Bigg(\sum_{|\alpha|\le 2}\int_0^T\int_\Omega\langle z\rangle^{2(l+2-|\alpha|)}|\partial_{t,x}^\alpha G^{(k)}|^2dxdzdt\\
&\qquad+k^4\int_0^T\int_\Omega\langle z\rangle^{2l}(|G^{(k)}|^2+|\partial_tG^{(k)}|^2)dxdzdt\\
&\qquad+\sum_{j=1}^2\sum_{i=0}^1\int_0^T\int_\Omega\langle z\rangle^{2(l+2-j)}|\partial_t^i\partial_z^jG^{(k)}|^2dxdzdt\\
&\qquad+\sum_{j=1}^2\int_0^T\int_\Omega\langle z\rangle^{2l}|\partial_t\partial_x^jG^{(k)}|^2dxdzdt\Bigg)
\end{aligned}\tag{3.57}$$

from the formula of $V_{n,2,0}^{(k)}$ given in (3.48), where the positive constant C depends only on T, C_0, and C_2.

By induction on $|\alpha|\in\mathbb{N}$, we obtain the following lemma.

LEMMA 3.7. *Let $Y_n^{(k)}$ be the solution sequence to the problem* (3.35). *Then for any $\alpha\in\mathbb{N}^2$, $\partial_{t,x}^\alpha Y_n^{(k)}$ satisfies the following estimate:*

$$\begin{aligned}
&\max_{0\le t\le T}\int_\Omega\langle z\rangle^{2l}|\partial_{t,x}^\alpha Y_n^{(k)}|^2dxdz+\int_0^T\int_\Omega\langle z\rangle^{2l}(|\partial_z\partial_{t,x}^\alpha Y_n^{(k)}|^2+k^2|\partial_{t,x}^\alpha Y_n^{(k)}|^2)dxdzdt\\
&\le C(T)\Bigg(\sum_{j=2}^{|\alpha|}\sum_{m=0}^{|\alpha|-j}\sum_{\substack{r+i+p\le j-1\\0\le q\le 2i\\0\le h\le 1}}\int_0^T\int_\Omega\langle z\rangle^{2(l+|\alpha|-j-m+2i-q)}k^{4r}|\partial_t^h\partial_z^q\partial_x^{m+2p}G^{(k)}|^2dxdzdt\\
&\qquad+\sum_{|\beta|\le|\alpha|}\int_0^T\int_\Omega\langle z\rangle^{2(l+|\alpha|-|\beta|)}|\partial_{t,x}^\beta G^{(k)}|^2dxdzdt\Bigg).
\end{aligned}\tag{3.58}$$

3.2.6. Step 6: Estimates of normal derivatives , ( ). For any $j\in\mathbb{N}$ and $\alpha=(\alpha_1,\alpha_2)\in\mathbb{N}^2$, set $W_{n,\alpha,j}^{(k)}=\partial_z^j\partial_{t,x}^\alpha Y_n^{(k)}$.

From (3.58) and the obvious identity

$$\partial_z^j\partial_t^{\alpha_1}\partial_x^{\alpha_2}Y_n^{(k)}(t,x,z)=\partial_z^j\partial_t^{\alpha_1}\partial_x^{\alpha_2}Y_n^{(k)}(0,x,z)+\int_0^t\partial_z^j\partial_t^{\alpha_1+1}\partial_x^{\alpha_2}Y_n^{(k)}(s,x,z)ds$$

we get

$$\max_{0\le t\le T}\int_\Omega \langle z\rangle^{2l}|\partial_z\partial_t^{\alpha_1}\partial_x^{\alpha_2}Y_n^{(k)}|^2dxdz \le \int_\Omega \langle z\rangle^{2l}|\partial_z\partial_t^{\alpha_1}\partial_x^{\alpha_2}Y_n^{(k)}(t=0)|^2dxdz \tag{3.59}$$
$$+C(T)\Bigg(\sum_{j=2}^{|\alpha|+1}\sum_{m=0}^{|\alpha|+1-j}\int_\Omega \langle z\rangle^{2(l+|\alpha|+1-j-m)}|\partial_x^m\partial_t^jY_n^{(k)}(t=0)|^2dxdz$$
$$+\sum_{|\beta|\le|\alpha|+1}\int_0^T\int_\Omega \langle z\rangle^{2(l+|\alpha|+1-|\beta|)}|\partial_{t,x}^\beta G^{(k)}|^2dxdzdt\Bigg).$$

From (3.52) it follows that

(3.60)
$$\int_\Omega \langle z\rangle^{2l}\ |\partial_z^2\partial_t^{\alpha_1}\partial_x^{\alpha_2}Y_n^{(k)}|^2dxdz$$
$$\le C_0\Bigg\{\int_\Omega \langle z\rangle^{2(l+1)}(|\partial_z\partial_t^{\alpha_1}\partial_x^{\alpha_2}Y_n^{(k)}|^2+k^2|\partial_t^{\alpha_1}\partial_x^{\alpha_2}Y_n^{(k)}|^2)dxdz$$
$$+\int_\Omega \langle z\rangle^{2l}(|\partial_t^{\alpha_1+1}\partial_x^{\alpha_2}Y_n^{(k)}|^2+|\partial_t^{\alpha_1}\partial_x^{\alpha_2+1}Y_n^{(k)}|^2+|\partial_t^{\alpha_1}\partial_x^{\alpha_2+2}Y_n^{(k)}|^2$$
$$+k^4|\partial_t^{\alpha_1}\partial_x^{\alpha_2}Y_n^{(k)}|^2+|\partial_t^{\alpha_1}\partial_x^{\alpha_2}Y_{n-1}^{(k)}|^2+|\partial_t^{\alpha_1}\partial_x^{\alpha_2}G^{(k)}|^2+Q_{\alpha_1,\alpha_2})dxdz\Bigg\}$$

and

(3.61)
$$\int_\Omega \langle z\rangle^{2l}|Q_{\alpha_1,\alpha_2}|^2dxdz$$
$$\le C\Bigg\{\sum_{j\le\alpha_1-1,m\le\alpha_2}\int_\Omega \langle z\rangle^{2l}(|\partial_z^2\partial_t^j\partial_x^mY_n^{(k)}|^2+k^4|\partial_t^j\partial_x^mY_n^{(k)}|^2+|\partial_t^j\partial_x^mY_{n-1}^{(k)}|^2)dxdz$$
$$+\sum_{j\le\alpha_1-1,m\le\alpha_2}\int_\Omega \langle z\rangle^{2(l+1)}(|\partial_z\partial_t^j\partial_x^mY_n^{(k)}|^2+k^2|\partial_t^j\partial_x^mY_n^{(k)}|^2)dxdz$$
$$+\int_\Omega \langle z\rangle^{2l}(|\partial_t^{\alpha_1}\partial_x^{\alpha_2}Y_n^{(k)}|^2+\sum_{j\le\alpha_1-1}|\partial_t^j\partial_x^{\alpha_2+1}Y_n^{(k)}|^2)dxdz$$
$$+\sum_{m\le\alpha_2-1}\int_\Omega \langle z\rangle^{2l}(|\partial_z^2\partial_t^{\alpha_1}\partial_x^mY_n^{(k)}|^2+k^4|\partial_t^{\alpha_1}\partial_x^mY_n^{(k)}|^2+|\partial_t^{\alpha_1}\partial_x^mY_{n-1}^{(k)}|^2)dxdz$$
$$+\sum_{m\le\alpha_2-1}\int_\Omega \langle z\rangle^{2(l+1)}(|\partial_z\partial_t^{\alpha_1}\partial_x^mY_n^{(k)}|^2+k^2|\partial_t^{\alpha_1}\partial_x^mY_n^{(k)}|^2)dxdz\Bigg\}.$$

Substituting (3.61) into (3.60) yields

$$
\begin{aligned}
&\int_\Omega \langle z\rangle^{2l}\ |\partial_z^2\partial_t^{\alpha_1}\partial_x^{\alpha_2}Y_n^{(k)}|^2dxdz \\
&\qquad \le C\Bigg\{\sum_{\beta\le\alpha}\int_\Omega \langle z\rangle^{2(l+1)}(|\partial_z\partial_{t,x}^\beta Y_n^{(k)}|^2+k^2|\partial_{t,x}^\beta Y_n^{(k)}|^2)dxdz \\
&\qquad\quad +\int_\Omega\langle z\rangle^{2l}\Bigg(\sum_{\beta\le\alpha}(k^4|\partial_{t,x}^\beta Y_n^{(k)}|^2+|\partial_{t,x}^\beta Y_{n-1}^{(k)}|^2)+\sum_{\beta<\alpha}|\partial_z^2\partial_{t,x}^\beta Y_n^{(k)}|^2 \\
&\qquad\quad +\sum_{j\le\alpha_1}|\partial_t^j\partial_x^{\alpha_2+1}Y_n^{(k)}|^2+|\partial_t^{\alpha_1}\partial_x^{\alpha_2+2}Y_n^{(k)}|^2 \\
&\qquad\quad +|\partial_t^{\alpha_1+1}\partial_x^{\alpha_2}Y_n^{(k)}|^2+|\partial_t^{\alpha_1}\partial_x^{\alpha_2}G^{(k)}|^2\Bigg)dxdz\Bigg\},
\end{aligned}
\tag{3.62}
$$

where the notations $\beta\le\alpha$ and $\beta<\alpha$ for $\alpha,\beta\in\mathbb{N}^2$ mean that $\beta_1\le\alpha_1,\beta_2\le\alpha_2$, and $\beta_1\le\alpha_1,\beta_2\le\alpha_2$, $\beta_1+\beta_2<\alpha_1+\alpha_2$, respectively.

By using (3.59) and (3.58) in (3.62), we get

$$
\begin{aligned}
&\max_{0\le t\le T}\int_\Omega\langle z\rangle^{2l}|\partial_z^2\partial_{t,x}^\alpha Y_n^{(k)}|^2dxdz \\
&\qquad \le C\Bigg\{k^4\sum_{\beta\le\alpha}\Bigg(\int_\Omega\langle z\rangle^{2(l+|\alpha|-|\beta|)}|\partial_{t,x}^\beta Y_n^{(k)}(t=0)|^2dxdz \\
&\qquad\qquad +\int_0^T\int_\Omega\langle z\rangle^{2(l+|\alpha|-|\beta|)}|\partial_{t,x}^\beta G^{(k)}|^2dxdzdt\Bigg) \\
&\qquad\quad +\sum_{|\beta|\le|\alpha|+2}\Bigg(\int_\Omega\langle z\rangle^{2(l+|\alpha|+2-|\beta|)}|\partial_{t,x}^\beta Y_n^{(k)}(t=0)|^2dxdz \\
&\qquad\quad +\int_0^T\int_\Omega\langle z\rangle^{2(l+|\alpha|+2-|\beta|)}|\partial_{t,x}^\beta G^{(k)}|^2dxdzdt\Bigg) \\
&\qquad\quad +\sum_{\beta\le\alpha}\int_\Omega\langle z\rangle^{2(l+1)}|\partial_z\partial_{t,x}^\beta Y_n^{(k)}(t=0)|^2dxdz \\
&\qquad\quad +\sum_{\beta<\alpha}\int_0^T\int_\Omega\langle z\rangle^{2l}|\partial_z^2\partial_{t,x}^\beta Y_n^{(k)}|^2dxdzdt\Bigg\}
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\max_{0\le t\le T}\int_\Omega \langle z\rangle^{2l}|\partial_z^2\partial_{t,x}^\alpha Y_n^{(k)}|^2dxdz\\
&\quad\le C(T)\Bigg\{k^4\sum_{\beta\le\alpha}\bigg(\int_\Omega\langle z\rangle^{2(l+|\alpha|-|\beta|)}|\partial_{t,x}^\beta Y_n^{(k)}(t=0)|^2dxdz\\
&\qquad+\int_0^T\int_\Omega\langle z\rangle^{2(l+|\alpha|-|\beta|)}|\partial_{t,x}^\beta G^{(k)}|^2dxdzdt\bigg)\\
&\qquad+\sum_{|\beta|\le|\alpha|+2}\bigg(\int_\Omega\langle z\rangle^{2(l+|\alpha|+2-|\beta|)}|\partial_{t,x}^\beta Y_n^{(k)}(t=0)|^2dxdz\\
&\qquad+\int_0^T\int_\Omega\langle z\rangle^{2(l+|\alpha|+2-|\beta|)}|\partial_{t,x}^\beta G^{(k)}|^2dxdzdt\bigg)\\
&\qquad+\sum_{\beta\le\alpha}\int_\Omega\langle z\rangle^{2(l+1)}|\partial_z\partial_{t,x}^\beta Y_n^{(k)}(t=0)|^2dxdz\Bigg\}.
\end{aligned}\tag{3.63}
$$

Differentiating (3.52) with respect to z and by induction on $j\in\mathbb{N}$, one can obtain the following lemma.

LEMMA 3.8. *The solution $Y_n^{(k)}$ of* (3.35) *satisfies the following estimate:*

$$
\begin{aligned}
&\max_{0\le t\le T}\int_\Omega \langle z\rangle^{2l}|\partial_z^j\partial_{t,x}^\alpha Y_n^{(k)}|^2dxdz\\
&\le C(T)\Bigg\{\sum_{m=0}^{[j/2]}k^{4m}\sum_{|\beta|\le|\alpha|+j-2m}\bigg(\int_\Omega\langle z\rangle^{2(l+|\alpha|+j-2m-|\beta|)}|\partial_{t,x}^\beta Y_n^{(k)}(t=0)|^2dxdz\\
&\quad+\int_0^T\int_\Omega\langle z\rangle^{2(l+|\alpha|+j-2m-|\beta|)}|\partial_{t,x}^\beta G^{(k)}|^2dxdzdt\bigg)\\
&\quad+\sum_{|\beta|\le|\alpha|+j-1-2m}\int_\Omega <z>^{2(l+|\alpha|+j-1-2m-|\beta|)}|\partial_z\partial_{t,x}^\beta Y_n^{(k)}(t=0)|^2dxdz\\
&\quad+\max_{0\le t\le T}\sum_{m=1}^{j-2}\int_\Omega\langle z\rangle^{2(l+j-2-m)}|\partial_z^m\partial_{t,x}^\alpha G^{(k)}|^2dxdzdt\Bigg\}.
\end{aligned}\tag{3.64}
$$

Remark 3.9. From the problem (3.35), it is easy to estimate the first and third terms on the right-hand side of (3.64) by the source term $G^{(k)}$. For example, one can obtain

(3.65)
$$
\begin{aligned}
&\int_\Omega \langle z\rangle^{2(l+|\alpha|+j-2m-|\beta|)}|\partial_{t,x}^\beta Y_n^{(k)}(t=0)|^2 dxdz\\
&\leq C(T)\left(\sum_{\substack{r+i+p\leq\beta_t-1\\ 0\leq q\leq 2i}}\int_\Omega\langle z\rangle^{2(l+|\alpha|+j-2m-|\beta|+2i-q)}k^{4r}|\partial_z^q\partial_x^{\beta_x+2p}G^{(k)}|_{t=0}|^2dxdz\right.\\
&+\sum_{r+i+p\leq\beta_t-1}\int_\Omega\langle z\rangle^{2(l+|\alpha|+j-2m-|\beta|)}k^{4r}|\partial_t^i\partial_x^{\beta_x+2p}G^{(k)}|_{t=0}|^2dxdz\\
&\left.+\sum_{\substack{i+p\leq\beta_t-2\\ 0\leq q\leq 2i}}\int_\Omega\langle z\rangle^{2(l+|\alpha|-j-2m-|\beta|+2i-q)}|\partial_t\partial_z^q\partial_x^{\beta_x+2p}G^{(k)}|_{t=0}|^2dxdz\right)
\end{aligned}
$$

with $\beta=(\beta_t,\beta_x)$.

In summary, we conclude with the following proposition.

PROPOSITION 3.10. *The approximate solution sequence* $\{Y_n^{(k)}\}_{n\geq1}$ *constructed by* (3.35) *is bounded in* $W^{k,\infty}([0,T],H^s(\Omega))$ *for any fixed* $k,s\in\mathbb{N}$*; moreover,* $\{Y_n^{(k)}\}_{n\geq1}$ *satisfies the estimates given in Lemmas* 3.6*,* 3.7*, and* 3.8.

3.2.7. Step 7: Convergence of { ( )} ≥1. As usual, based on the high order norm boundedness estimate (3.64) of $\{Y_n^{(k)}\}_{n\geq1}$, it suffices to consider the convergence of $\{Y_n^{(k)}\}_{n\geq1}$ in the L^2-norm.

Let $W_n^{(k)}=Y_{n+1}^{(k)}-Y_n^{(k)}$. It follows from (3.35) that $W_n^{(k)}$ solves the following problem:

(3.66)
$$
\begin{cases}
(\partial_t+a_1\partial_x)W_n^{(k)}-a_4^2(\partial_z^2-k^2a^2)W_n^{(k)}+z(a_2\partial_z+ika_3)W_n^{(k)}-\dfrac{1}{n+1}\partial_x^2W_n^{(k)}\\
\quad+ka_5\displaystyle\int_z^{+\infty}e^{ka(z-\xi)}W_{n-1}^{(k)}(\cdot,\xi)d\xi+ka_6\int_0^z e^{ka(\xi-z)}W_{n-1}^{(k)}(\cdot,\xi)d\xi=-\dfrac{1}{n(n+1)}\partial_x^2Y_n^{(k)}\\
W_n^{(k)}|_{z=0}=0,\quad W_n^{(k)}\in S(\mathbb{R}_z^+)\\
W_n^{(k)}|_{t=0}=0.
\end{cases}
$$

In a way similar to (3.42), we deduce that for all $n\geq1$,

$$
\begin{aligned}
&\frac{d}{dt}\int_\Omega\langle z\rangle^{2l}|W_n^{(k)}|^2dxdz+\int_\Omega\langle z\rangle^{2l}(|\partial_zW_n^{(k)}|^2+k^2|W_n^{(k)}|^2)dxdz\\
&\qquad\leq C_0\int_\Omega\langle z\rangle^{2l}(|W_n^{(k)}|^2+|W_{n-1}^{(k)}|^2)dxdz+\frac{C_0}{n(n+1)}
\end{aligned}
\tag{3.67}
$$

by using the boundedness of $\{Y_n^{(k)}\}_{n\geq1}$. Applying Lemma 3.4 in (3.67) we achieve the following proposition.

PROPOSITION 3.11. *For any fixed $T > 0$ and $l \in \mathbb{N}$, it holds that*

$$\max_{0 \le t \le T} \int_\Omega \langle z \rangle^{2l} |Y_{n+1}^{(k)} - Y_n^{(k)}|^2 dxdz \longrightarrow 0 \tag{3.68}$$

when n goes to infinite.

By bringing together Propositions 3.10 with 3.11, we deduce the following theorem.

THEOREM 3.12. *Consider problems* (3.33) *and* (3.34)*. Let $T > 0$ be fixed. Suppose that coefficients a and $\{a_j\}_{j=1}^6$ belong to $H^\infty([0,T] \times \mathbb{R}_x)$ with $a(t,x) \ge a_0$ and $a_4(t,x) \ge a_0$ for a constant $a_0 > 0$, $W_0^{(k)} \in H^\infty([0,T] \times \mathbb{R}_x)$ and $G^{(k)} \in H^\infty([0,T] \times \mathbb{R}_x \times \mathbb{R}_z^+)$ satisfying all compatibility conditions, with $z^l G^{(k)} \in H^\infty([0,T] \times \mathbb{R}_x \times \mathbb{R}_z^+)$ for all $l \in \mathbb{N}$. Then there exist unique solutions $\{Y^{(k)}\}_{k \in Z \setminus \{0\}}$ to problems* (3.33) *and* (3.34) *satisfying $< z >^l Y^{(k)} \in H^\infty([0,T] \times \mathbb{R}_x \times \mathbb{R}_z^+)$ for all $l \in \mathbb{N}$.*

Combining Theorem 3.12 with the transformation (3.32) and Proposition 3.3, we obtain the following existence and uniqueness result of solutions (u, w) to the Poisson–Prandtl coupled problem (3.1).

THEOREM 3.13. *Let $T > 0$ be fixed and all coefficients a and $\{a_j\}_{j=1}^4$ be the same as given in Theorem* 3.12*. Assume that $(b_0, b_1) \in C_p^\infty(T^1, H^\infty([0,T] \times \mathbb{R}_x))$ and $(f, g) \in C_p^\infty(T^1, H^\infty([0,T] \times \mathbb{R}_x \times \mathbb{R}_z^+))$, smooth periodic in $\theta \in T^1$ with valued in $H^\infty([0,T] \times \mathbb{R}_x)$ and $H^\infty([0,T] \times \mathbb{R}_x \times \mathbb{R}_z^+)$, respectively, satisfying*

$$\mathbf{m}_\theta(f) = \mathbf{m}_\theta(g) = \mathbf{m}_\theta(b_0) = \mathbf{m}_\theta(b_1) = 0$$

and $(z^l f, z^l g) \in C_p^\infty(T^1, H^\infty([0,T] \times \mathbb{R}_x \times \mathbb{R}_z^+))$ for any $l \in \mathbb{N}$, and, moreover, all compatibility conditions for (3.1) *hold. Then, there exist unique solutions $(u, w) \in C_p^\infty(T^1, H^\infty([0,T] \times \mathbb{R}_x \times \mathbb{R}_z^+))$ to the problem* (3.1) *satisfying $\mathbf{m}_\theta(u) = \mathbf{m}_\theta(w) = 0$ and $(z^l u, z^l w) \in C_p^\infty(T^1, H^\infty([0,T] \times \mathbb{R}_x \times \mathbb{R}_z^+))$ for all $l \in \mathbb{N}$.*

4. Rigorous justification of the zero-viscosity limit. In this section, we shall rigorously justify the formal analysis given in section 2.

First, let us suppose that

(H1) all compatibility conditions for the problem (2.1) are satisfied,

(H2) all compatibility conditions for problems (2.29), (2.35), and (5.11), (5.13), and (5.22) given later are satisfied,

which will be studied in detail at the end of this section. It is easy to see that compatibility conditions for the problem (2.26) follow from those for (2.1) by setting $\epsilon = 0$.

Let V^ϵ be the solution to the problem (2.1). From section 3, we know that one can uniquely determine $(d_0^{(2)}, d_0^{(3)})$ and $\{(d_{j+1}^{(2)}, d_{j+1}^{(3)})\}_{j \ge 0}$ from problems (2.35) and (2.36) and (5.22) and (5.23), respectively. Thus, from sections 2 and 5.1 we obtain each order smooth profile $\{(a_j, c_j, b_j, d_j)\}_{j \ge 0}$ in the formal expansion of the solution

$$V^\epsilon(t,x) \sim \sum_{j \ge 0} \epsilon^j (a_j(t,x) + c_j\left(t, x; \frac{\varphi(t,x)}{\epsilon}\right) + b_j\left(t, x_2; \frac{x_1}{\epsilon}\right) + d_j\left(t, x_2; \frac{x_1}{\epsilon}, \frac{\varphi^0(t,x_2)}{\epsilon}\right). \tag{4.1}$$

Denote by

$$V_J^\epsilon(t,x) = \sum_{j=0}^J \epsilon^j \left(a_j(t,x) + c_j\left(t, x; \frac{\varphi(t,x)}{\epsilon}\right) + b_j\left(t, x_2; \frac{x_1}{\epsilon}\right) + d_j\left(t, x_2; \frac{x_1}{\epsilon}, \frac{\varphi^0(t,x_2)}{\epsilon}\right)\right)$$

the Jth order approximate solution for any fixed $J \in \mathbb{N}$.

From the discussion in sections 2 and 5.1, it is easy to see that $W_J^\epsilon = V^\epsilon - V_J^\epsilon$ satisfies the following problem:

$$
\begin{cases}
A_0(V')\partial_t W_J^\epsilon + A_1(V')\partial_{x_1} W_J^\epsilon + A_2(V')\partial_{x_2} W_J^\epsilon = B(\epsilon^2, D\epsilon^2) W_J^\epsilon + R_J^\epsilon, \\
M^+ W_J^\epsilon = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} W_J^\epsilon = 0, \qquad \text{on} \quad x_1 = 0, \\
W_J^\epsilon|_{t=0} = 0,
\end{cases} \tag{4.2}
$$

where the remainder $R_J^\epsilon(t,x)$ satisfies

$$
\|R_J^\epsilon\|_{L^\infty([0,T],L^2(\mathbb{R}_+^2))} \le C\epsilon^{J-1} \tag{4.3}
$$

for any $T > 0$ and a constant $C > 0$.

By using the classical theory of the linearized Navier–Stokes equations in the problem (4.2), we immediately conclude

$$
\|V^\epsilon - V_J^\epsilon\|_{L^\infty([0,T],L^2(\mathbb{R}_+^2))} \le C_1\epsilon^{J-1} \tag{4.4}
$$

which implies

$$
\|V^\epsilon - V_J^\epsilon\|_{L^\infty([0,T],L^2(\mathbb{R}_+^2))} \le C_2\epsilon^{J+1} \tag{4.5}
$$

for any $J \in \mathbb{N}$ with the constant C_2 depending only upon T and J.

Therefore, we obtain the following theorem.

THEOREM 4.1. *Under the assumptions* (H1) *and* (H2), *the solution* $V^\epsilon = (\rho^\epsilon, v_1^\epsilon, v_2^\epsilon)$ *of* (2.1) *has the following asymptotics:*

$$
\begin{aligned}
V^\epsilon(t,x) = \sum_{j=0}^{J} \epsilon^j \Bigg(& a_j(t,x) + c_j\left(t, x; \frac{\varphi(t,x)}{\epsilon}\right) + b_j\left(t, x_2; \frac{x_1}{\epsilon}\right) \\
& + d_j\left(t, x_2; \frac{x_1}{\epsilon}, \frac{\varphi^0(t,x_2)}{\epsilon}\right)\Bigg) + O(\epsilon^{J+1})
\end{aligned} \tag{4.6}
$$

in $L^\infty([0,T], L^2(\mathbb{R}_+^2))$ *for any fixed* $J \in \mathbb{N}$, *where* $a_j(t,x)$ *satisfy problems* (2.26) *and* (5.11) *for the linearized Euler equations;* $c_j(t,x;\frac{\varphi(t,x)}{\epsilon})$ *are determined from* (2.17) *and* (2.20), *and* (5.3)–(5.5); $(b_j^{(1)}, b_j^{(2)})$ *are given in* (2.25) *and* (5.10); *and* $b_j^{(3)}(t,x_2;z)$ *satisfy problems* (2.29) *and* (5.13) *for the linearized Prandtl equation;* $d_j^{(1)}$ *are given in* (2.31) *and* (5.17); *and* $(d_j^{(2)}, d_j^{(3)})(t,x_2;z,\theta)$ *together with their vorticity with respect to* $(z,\theta)-$*variables satisfy problems* (2.35) *and* (2.36), *and* (5.22) and (5.23) *for the Poisson–Prandtl coupled system.*

Remark 4.2. (1) From (4.6), one immediately concludes Theorem 2.1.

(2) The asymptotic relation (4.6) holds in high order Sobolev spaces with weighted norms due to the high frequency of oscillations in $\{c_j, d_j\}_{j\ge 0}$ and the multiple scales in boundary layers $\{b_j, d_j\}_{j\ge 0}$, e.g., in $L^\infty([0,T], H_\epsilon^s(\mathbb{R}^2))$ with the norm of $H_\epsilon^s(\mathbb{R}^2)$ being defined as

$$
\|u\|_{s,\epsilon} = \left(\sum_{|\alpha|\le s} \epsilon^{2|\alpha|} \|\partial_x^\alpha u\|_{L^2(I\!R_+^2)}^2 \right)^{\frac{1}{2}}.
$$

Finally, for completeness, let us investigate the assumptions (H1) and (H2).

(I) The compatibility condition for the problem of linearized Navier–Stokes equations (2.1) can be formulated in the classical way as follows.

(I1) The zero-th order compatibility condition is

$$V_0^{(2)} = V_0^{(3)} = 0 \qquad \text{on} \quad \{x_1 = 0\}. \tag{4.7}$$

(I2) The jth order compatibility condition ($j \geq 1$).

Set $\Phi^\epsilon(t,x) = \Phi(t,x; \frac{\varphi(t,x)}{\epsilon})$. For any fixed $j \in \mathbb{N}$ with $j \geq 1$, it follows from the equations in (2.1) that

$$\begin{aligned}\partial_t^j V^\epsilon = (A_0(V'))^{-1}\{&B(\epsilon^2, D\epsilon^2)\partial_t^{j-1}V^\epsilon + \partial_t^{j-1}\Phi^\epsilon \\ &-[\partial_t^{j-1}, A_0(V')]\partial_t V^\epsilon - \partial_t^{j-1}(A_1(V')\partial_{x_1}V^\epsilon + A_2(V')\partial_{x_2}V^\epsilon)\}\end{aligned}$$

by induction on j. By using the initial data $V^\epsilon|_{t=0} = V_0(x)$, we know that $V_j^\epsilon(x) = \partial_t^j V^\epsilon|_{t=0}$ is a linear function of $\{\partial_x^\alpha V_0\}_{|\alpha|\leq 2j}$ and $\{\partial_t^k \partial_x^\alpha \Phi^\epsilon(t=0)\}_{k\leq j-1, |\alpha|\leq 2(j-1-k)}$. Then, the jth order compatibility condition for the problem (2.1) is

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} V_j^\epsilon = 0 \qquad \text{on} \quad \{x_1 = 0\}. \tag{4.8}$$

Next, we study the assumption (H1).

($\mathbb{I}$) The compatibility condition for the problem of linearized Prandtl equation (2.29).

($\mathbb{I}$1) The zero-th order compatibility condition is

$$a_0^{(3)} = 0 \qquad \text{on} \quad \{t = x_1 = 0\}, \tag{4.9}$$

which is a simple consequence of the zero-th order compatibility condition (4.7) by noting $a_0|_{t=0} = V_0(x)$ in (2.26).

($\mathbb{I}$2) The jth order compatibility condition ($j \geq 1$).

It follows from the equation and the initial data in (2.29) that

$$\partial_t^j b_0^{(3)}|_{t=0} = 0.$$

So, the jth order compatibility condition for the problem (2.29) is

$$\partial_t^j a_0^{(3)} = 0 \qquad \text{on} \quad \{t = x_1 = 0\}, \tag{4.10}$$

where $a_0^{(3)}(t,x)$ is determined by the problem (2.26).

The compatibility conditions for the problem (5.13) can be obtained in the same ways as those for the problem (2.29) given above.

Both of problems (2.35) and (5.22) are the special cases of the problem (3.1), so their compatibility conditions can be stated in the same way as those for the problem (3.1) given in section 3.

Finally, we note that in general compatibility conditions for problems of profiles, $\{a_j, c_j, b_j, d_j\}_{j\geq 0}$ could not be implied by those for the original linearized Navier–Stokes equations (2.1). The simplest case to guarantee all compatibility conditions given as above valid is that

$$\begin{cases} \partial_t^k \partial_x^\alpha \Phi(t,x;\theta) = 0, & \text{on } \{t = x_1 = 0\} \\ \partial_x^\alpha V_0(x) = 0, & \text{on } \{x_1 = 0\} \end{cases}$$

hold for any $k \in \mathbb{N}$ and $\alpha \in \mathbb{N}^2$.

5. Appendix.

5.1. Problems of high order profiles. In section 2, we derived problems for the leading profiles in the expansion of the solution $V^\epsilon(t,x) = V^\epsilon_{in}(t,x) + V^\epsilon_{bd}(t,x)$ to the problem (2.1). In this subsection, let us briefly derive problems for high order terms in the expansion.

By induction, suppose that $\{a_k(t,x), c_k(t,x;\theta), b_k(t,x_2;z), d_k(t,x_2;z,\theta^0)\}_{k\le j}$ are known already, we want to determine profiles $\{a_{j+1}(t,x), c_{j+1}(t,x;\theta), b_{j+1}(t,x_2;z), d_{j+1}(t,x_2;z,\theta^0)\}$ in the expansion (2.11).

It follows from (2.13) and the fact $\mathbf{m}_\theta(\mathcal{F}_j)=0$ that

$$L(\partial_t,\partial_x)a_j = (B_1\partial^2_{x_1} + B_2\partial^2_{x_2} + B_3\partial^2_{x_1x_2})a_{j-2} \tag{5.1}$$

and the difference between $\mathcal{F}_j = 0$ and (5.1) gives rise to

$$\sum_{k=0}^{2} \varphi_{x_k} A_k(V')\partial_\theta c_{j+1} = \tilde f_j, \tag{5.2}$$

where

$$\begin{aligned}\tilde f_j = {} & (\varphi^2_{x_1}B_1 + \varphi^2_{x_2}B_2 + \varphi_{x_1}\varphi_{x_2}B_3)\partial^2_\theta c_j - L(\partial_t,\partial_x)c_j \\ & + (\varphi_{x_1x_1}B_1 + \varphi_{x_2x_2}B_2 + \varphi_{x_1x_2}B_3)\partial_\theta c_{j-1} \\ & + (2\varphi_{x_1}B_1 + \varphi_{x_2}B_3)\partial^2_{\theta x_1}c_{j-1} + (2\varphi_{x_2}B_2 + \varphi_{x_1}B_3)\partial^2_{\theta x_2}c_{j-1} \\ & + (B_1\partial^2_{x_1} + B_2\partial^2_{x_2} + B_3\partial^2_{x_1x_2})c_{j-2}\end{aligned}$$

satisfies $\mathbf{m}_\theta(\tilde f_j) = 0$.

If we set

$$c_{j+1}(t,x;\theta) = \sum_{k=1}^{3} v^{(k)}_{j+1}(t,x;\theta)\vec r_k(\nabla\varphi), \tag{5.3}$$

then from (5.2) it follows

$$(\varphi_t - \tau_k(\nabla\varphi))\partial_\theta v^{(k)}_{j+1} = (\vec l_k(\nabla\varphi)\cdot \tilde f_j)(t,x;\theta), \quad k=2,3, \tag{5.4}$$

where $\tau_k(\nabla\varphi)$ are defined in (2.3). Due to the assumption (2.6), we obtain that $(v^{(2)}_{j+1}, v^{(3)}_{j+1})$ can be uniquely determined by (5.4) with $\mathbf{m}_\theta(v^{(2)}_{j+1}, v^{(3)}_{j+1}) = 0$.

To solve $v^{(1)}_{j+1}$, acting $\vec l_1(\nabla\varphi)$ from the left on the same equation as (5.2) with j being replaced by $j+1$, and using (5.3), one gets that $v^{(1)}_{j+1}$ satisfies the following problem:

$$\begin{cases} [(\vec l_1 A_0\vec r_1)\partial_t + (\vec l_1 A_1 \vec r_1)\partial_{x_1} + (\vec l_1 A_2\vec r_1)\partial_{x_2}]v^{(1)}_{j+1} + \vec l_1(A_0\partial_t\vec r_1 + A_1\partial_{x_1}\vec r_1 + A_2\partial_{x_2}\vec r_1)v^{(1)}_{j+1} \\ \quad -\vec l_1(\varphi^2_{x_1}B_1 + \varphi^2_{x_2}B_2 + \varphi_{x_1}\varphi_{x_2}B_3)\vec r_1\partial^2_\theta v^{(1)}_{j+1} = h_{j+1} \\ v^{(1)}_{j+1}|_{t=0} = 0, \end{cases} \tag{5.5}$$

which is similar to the problem (2.20), where

$$\begin{aligned} h_{j+1} = \vec l_1[& (\varphi_{x_1x_1}B_1 + \varphi_{x_2x_2}B_2 + \varphi_{x_1x_2}B_3)\partial_\theta c_j + (2\varphi_{x_1}B_1 + \varphi_{x_2}B_3)\partial^2_{\theta x_1}c_j \\ & + (2\varphi_{x_2}B_2 + \varphi_{x_1}B_3)\partial^2_{\theta x_2}c_j + (B_1\partial^2_{x_1} + B_2\partial^2_{x_2} + B_3\partial^2_{x_1x_2})c_{j-1} \\ & - (L(\partial_t,\partial_x) - (\varphi^2_{x_1}B_1 + \varphi^2_{x_2}B_2 + \varphi_{x_1}\varphi_{x_2}B_3)\partial^2_\theta)(v^{(2)}_{j+1}\vec r_2 + v^{(3)}_{j+1}\vec r_3)]. \end{aligned}$$

As in (2.21) and (2.22), the $0(\epsilon^{j+1})-$term of the boundary condition $M^+V^\epsilon|_{x_1=0} = 0$ in (2.1) gives

$$\begin{cases} a_{j+1}^{(k)}(t,x) + b_{j+1}^{(k)}(t,x_2;z) = 0 & \text{on } \{x_1 = z = 0\} \\ c_{j+1}^{(k)}(t,x;\theta) + d_{j+1}^{(k)}(t,x_2;z,\theta^0) = 0 & \text{on } \{x_1 = z = 0, \theta = \theta^0\} \end{cases} \tag{5.6}$$

for $k \in \{2,3\}$.

Thus, to solve a_{j+1} from the same equation as (5.1) with j being replaced by $j+1$, one should study $b_{j+1}^{(2)}$ first in order to determine the boundary value of $a_{j+1}^{(2)}$ on $\{x_1 = 0\}$.

Acting upon the averaging operator $\mathbf{m}_\theta$ on $\mathcal{G}_j = 0$ from (2.15), and using the assumption $\mathbf{m}_\theta(d_k) = 0$ for any $k \geq 0$, we get

$$A_1(0)\partial_z b_{j+1} = \tilde{g}_j(t,x_2;z), \tag{5.7}$$

and the difference between $\mathcal{G}_j = 0$ and (5.7) gives rise to

$$(\varphi_t^0 A_0(0) + \varphi_{x_2}^0 A_2(0))\partial_\theta d_{j+1} + A_1(0)\partial_z d_{j+1} = g_j^\star(t,x_2;z,\theta), \tag{5.8}$$

where

$$\begin{cases} \tilde{g}_j(t,x_2;z) = B_1\partial_z^2 b_j - (A_0(0)\partial_t + A_2(0)\partial_{x_2})b_j - zA_1'(0)\partial_z b_j - \mathbf{m}_\theta(g_j), \\ g_j^\star(t,x_2;z,\theta) = (B_1\partial_z^2 + (\varphi_{x_2}^0)^2 B_2\partial_\theta^2 + \varphi_{x_2}^0 B_3\partial_{z\theta}^2)d_j - (A_0(0)\partial_t + A_2(0)\partial_{x_2})d_j \\ \qquad -z(\varphi_t^0 A_0'(0) + \varphi_{x_2}^0 A_2'(0))\partial_\theta d_j - zA_1'(0)\partial_z d_j - g_j + \mathbf{m}_\theta(g_j). \end{cases}$$

From (5.7), we deduce immediately that $(b_{j+1}^{(1)}, b_{j+1}^{(2)})$ solves the following problem:

$$\begin{cases} \begin{pmatrix} 0 & \rho'(0) \\ c^2(0) & 0 \end{pmatrix} \begin{pmatrix} \partial_z b_{j+1}^{(1)} \\ \partial_z b_{j+1}^{(2)} \end{pmatrix} = \begin{pmatrix} \tilde{g}_j^{(1)} \\ \tilde{g}_j^{(2)} \end{pmatrix} \\ (b_{j+1}^{(1)}, b_{j+1}^{(2)}) \in S(\mathbb{R}_z^+), \end{cases} \tag{5.9}$$

which implies

$$\begin{cases} b_{j+1}^{(1)}(t,x_2;z) = -c^{-2}(0)\int_z^{+\infty} \tilde{g}_j^{(2)}(t,x_2;\xi)d\xi \\ b_{j+1}^{(2)}(t,x_2;z) = -(\rho'(0))^{-1}\int_z^{+\infty} \tilde{g}_j^{(1)}(t,x_2;\xi)d\xi. \end{cases} \tag{5.10}$$

Therefore, from the same equation as (5.1) with j being replaced by $j+1$, we know that $a_{j+1}(t,x)$ solves the following problem:

$$\begin{cases} L(\partial_t,\partial_x)a_{j+1} = (B_1\partial_{x_1}^2 + B_2\partial_{x_2}^2 + B_3\partial_{x_1x_2}^2)a_{j-1}, \\ a_{j+1}^{(2)}|_{x_1=0} = (\rho'(0))^{-1}\int_0^{+\infty} \tilde{g}_j^{(1)}(t,x_2;\xi)d\xi, \\ a_{j+1}|_{t=0} = 0. \end{cases} \tag{5.11}$$

To determine $b_{j+1}^{(3)}(t,x_2;z)$, we act upon the averaging operator $\mathbf{m}_\theta$ on $\mathcal{G}_{j+1} = 0$ with $\mathcal{G}_{j+1}$ being given as in (2.15), and obtain

$$L_{bd}(\partial_t,\partial_{x_2})b_{j+1} + zA_1'(0)\partial_z b_{j+1} - B_1\partial_z^2 b_{j+1} + A_1(0)\partial_z b_{j+2} + \mathbf{m}_\theta(g_{j+1}) = 0. \tag{5.12}$$

The third component of (5.12) shows that $b_{j+1}^{(3)}$ solves the following initial-boundary value problem for the linearized Prandtl equation:

(5.13)
$$\begin{cases} (\partial_t + v_2'(0)\partial_{x_2})b_{j+1}^{(3)} + z\dfrac{\partial v_1'(0)}{\partial x_1}\partial_z b_{j+1}^{(3)} - \dfrac{1}{\rho'(0)}\partial_z^2 b_{j+1}^{(3)} = -\dfrac{c^2(0)}{\rho'(0)}\partial_{x_2} b_{j+1}^{(1)} - \mathbf{m}_\theta(g_{j+1}^{(3)}) \\ b_{j+1}^{(3)}|_{z=0} = -a_{j+1}^{(3)}(t,0,x_2), \quad b_{j+1}^{(3)} \in S(\mathbb{R}_z^+) \\ b_{j+1}^{(3)}|_{t=0} = 0, \end{cases}$$

where $a_{j+1}^{(3)}$ is the third component of a_{j+1} given in (5.11), and $b_{j+1}^{(1)}$ is given already in (5.10).

However, it remains to be determined that $d_{j+1}(t,x_2;z,\theta)$. From (5.8), we get

$$\begin{cases} \varphi_{x_2}^0 \partial_\theta d_{j+1}^{(1)} = \dfrac{1}{c^2(0)} g_j^{\star(3)}, \quad \partial_z d_{j+1}^{(1)} = \dfrac{1}{c^2(0)} g_j^{\star(2)} \\ d_{j+1}^{(1)} \in S(\mathbb{R}_z^+) \end{cases} \tag{5.14}$$

and

$$\partial_z d_{j+1}^{(2)} + \varphi_{x_2}^0 \partial_\theta d_{j+1}^{(3)} = \frac{1}{\rho'(0)} g_j^{\star(1)}. \tag{5.15}$$

By using the fact (2.32) in (5.8), we know

$$\mathbb{E}(g_j^\star) = 0,$$

which implies especially

$$\partial_z g_j^{\star(3)} - \varphi_{x_2}^0 \partial_\theta g_j^{\star(2)} = 0. \tag{5.16}$$

Obviously, (5.16) is the compatibility condition for solving $d_{j+1}^{(1)}$ from (5.14), and

$$d_{j+1}^{(1)} = -c^{-2}(0)\int_z^{+\infty} g_j^{\star(2)}(t,x_2;\xi,\theta)d\xi. \tag{5.17}$$

Acting upon the operator $\mathbb{E}$ on the same equations as in (5.8) with j being replaced by $j+1$, it follows that

$$\mathbb{E}(L_{bd}(\partial_t,\partial_{x_2})d_{j+1} + z(\varphi_t^0 A_0'(0) + \varphi_{x_2}^0 A_2'(0))\partial_\theta d_{j+1} + zA_1'(0)\partial_z d_{j+1}$$

$$-(B_1\partial_z^2 + (\varphi_{x_2}^0)^2 B_2 \partial_\theta^2 + \varphi_{x_2}^0 B_3 \partial_{z\theta}^2)d_{j+1} + g_{j+1} - \mathbf{m}_\theta(g_{j+1})) = 0. \tag{5.18}$$

Denote by $\tilde{A}$ and $\tilde{B}$ the second and third components of the above term on which $\mathbb{E}$

acts. Due to (5.15), they can be expressed as

$$\begin{cases} \tilde{A} = \rho'(0)\left((\partial_t + v_2'(0)\partial_{x_2})d_{j+1}^{(2)} + z\left(\dfrac{\partial v_1'(0)}{\partial x_1}\partial_z + \varphi_{x_2}^0\dfrac{\partial v_2'(0)}{\partial x_1}\partial_\theta\right)d_{j+1}^{(2)}\right) \\ \qquad - (\partial_z^2 + (\varphi_{x_2}^0)^2\partial_\theta^2)d_{j+1}^{(2)} + z\dfrac{\partial c^2(0)}{\partial x_1}\partial_z d_{j+1}^{(1)} + g_{j+1}^{(2)} - \mathbf{m}_\theta(g_{j+1}^{(2)}) - \dfrac{D}{\rho'(0)}\partial_z g_j^{\star(1)} \\ \tilde{B} = \rho'(0)\left((\partial_t + v_2'(0)\partial_{x_2})d_{j+1}^{(3)} + z\left(\dfrac{\partial v_1'(0)}{\partial x_1}\partial_z + \varphi_{x_2}^0\dfrac{\partial v_2'(0)}{\partial x_1}\partial_\theta\right)d_{j+1}^{(3)}\right) \\ \qquad - (\partial_z^2 + (\varphi_{x_2}^0)^2\partial_\theta^2)d_{j+1}^{(3)} + c^2(0)\partial_{x_2}d_{j+1}^{(1)} + z\dfrac{\partial c^2(0)}{\partial x_1}\varphi_{x_2}^0\partial_\theta d_{j+1}^{(1)} + g_{j+1}^{(3)} \\ \qquad - \mathbf{m}_\theta(g_{j+1}^{(3)}) - \frac{D\varphi_{x_2}^0}{\rho'(0)}\partial_\theta g_j^{\star(1)}. \end{cases}$$

We deduce from (5.18) that

$$\omega_{j+1}(t,x_2;z,\theta) = \varphi_{x_2}^0\partial_\theta d_{j+1}^{(2)} - \partial_z d_{j+1}^{(3)} \tag{5.19}$$

satisfies

(5.20)

$$(\partial_t + v_2'(0)\partial_{x_2})\omega_{j+1} + z\left(\frac{\partial v_1'(0)}{\partial x_1}\partial_z + \varphi_{x_2}^0\frac{\partial v_2'(0)}{\partial x_1}\partial_\theta\right)\omega_{j+1} - \frac{1}{\rho'(0)}(\partial_z^2 + (\varphi_{x_2}^0)^2\partial_\theta^2)\omega_{j+1}$$
$$- \left(\frac{\partial v_1'(0)}{\partial x_1}\partial_z + \varphi_{x_2}^0\frac{\partial v_2'(0)}{\partial x_1}\partial_\theta\right)d_{j+1}^{(3)} = R_{j+1},$$

where

(5.21)

$$R_{j+1} = \frac{1}{\rho'(0)}\left[\partial_z g_{j+1}^{(3)} - \mathbf{m}_\theta(\partial_z g_{j+1}^{(3)}) - \varphi_{x_2}^0\partial_\theta g_{j+1}^{(2)} + c^2(0)\partial_{zx_2}^2 d_{j+1}^{(1)} + \frac{\partial c^2(0)}{\partial x_1}\varphi_{x_2}^0\partial_\theta d_{j+1}^{(1)}\right]$$

with $d_{j+1}^{(1)}$ being given in (5.17).

Combining (5.15), (5.19), (5.20), and (5.6) leads to $(d_{j+1}^{(2)}, d_{j+1}^{(3)}, \omega_{j+1})$ satisfying the following problems:

(5.22)

$$\begin{cases} (\partial_z^2 + (\varphi_{x_2}^0)^2\partial_\theta^2)d_{j+1}^{(3)} = \dfrac{\varphi_{x_2}^0}{\rho'(0)}\partial_\theta g_j^{\star(1)} - \partial_z\omega_{j+1} \\ (\partial_t + v_2'(0)\partial_{x_2})\omega_{j+1} + z\left(\dfrac{\partial v_1'(0)}{\partial x_1}\partial_z + \varphi_{x_2}^0\dfrac{\partial v_2'(0)}{\partial x_1}\partial_\theta\right)\omega_{j+1} - \dfrac{1}{\rho'(0)}(\partial_z^2 + (\varphi_{x_2}^0)^2\partial_\theta^2)\omega_{j+1} \\ \quad -\left(\dfrac{\partial v_1'(0)}{\partial x_1}\partial_z + \varphi_{x_2}^0\dfrac{\partial v_2'(0)}{\partial x_1}\partial_\theta\right)d_{j+1}^{(3)} = R_{j+1} \\ d_{j+1}^{(3)}|_{z=0} = -c_{j+1}^{(3)}(t,0,x_2;\theta) \\ (\omega_{j+1} + \partial_z d_{j+1}^{(3)})|_{z=0} = -\varphi_{x_2}^0(\partial_\theta c_{j+1}^{(2)})(t,0,x_2;\theta) \\ (d_{j+1}^{(3)}, \omega_{j+1}) \in S(\mathbb{R}_z^+) \\ \omega_{j+1}|_{t=0} = 0 \end{cases}$$

and

$$\begin{cases} (\partial_z^2 + (\varphi^0_{x_2})^2\partial_\theta^2)d^{(2)}_{j+1} = \varphi^0_{x_2}\partial_\theta\omega_{j+1} + \dfrac{1}{\rho'(0)}\partial_z g_j^{\star(1)}, \\ d^{(2)}_{j+1}|_{z=0} = -c^{(2)}_{j+1}(t,0,x_2;\theta), \\ d^{(2)}_{j+1} \in S(\mathbb{R}^+_z), \end{cases} \tag{5.23}$$

which are similar to problems (2.35) and (2.36).

Remark 5.1. When $\varphi|_{x_1=0} = \varphi^0(t,x_2) \equiv 0$, the terms d_j disappear, similar to Gues [4], the boundary conditions (2.21) and (5.6) become

$$(a_j^{(k)}(t,x) + c_j^{(k)}(t,x;\theta) + b_j^{(k)}(t,x_2;z))|_{x_1=z=\theta=0} = 0$$

for any $j \geq 0, k = 2,3$. In this case, we obtain that $a_0(t,x)$ and $c_0(t,x;\theta)$ satisfy the same problems as (2.26) and (2.17), but for $j \geq 0$, $b_j(t,x_2;z)$ satisfies problems (2.29) and (5.13) with different boundary conditions.

5.2. The case of shorter wavelength. In this subsection, we are going to study the problem (2.1) when the wavelength of the oscillatory force term is shorter than the square root of the viscosity, i.e., $\Phi = \Phi(t,x;\frac{\varphi(t,x)}{\epsilon^\alpha})$ with $\alpha > 1$. It is observed in this case that the leading profiles of the solution V^ϵ are the same as the case without oscillations [12], and the oscillation shall appear only at the high order profiles. To explain the idea, we shall investigate the case that $\alpha = 2$.

Take the following ansatz for the solution of (2.1) with $\Phi = \Phi(t,x;\frac{\varphi(t,x)}{\epsilon^2})$ and $\varphi(t,x)$ being the same as in (2.6):

$$V^\epsilon(t,x) = V^\epsilon_{in}(t,x) + V^\epsilon_{bd}(t,x), \tag{5.24}$$

where the outflow V^ϵ_{in} admits the expansion

$$V^\epsilon_{in}(t,x) = \sum_{j\geq 0}\epsilon^j\left(a_j(t,x) + c_j\left(t,x;\frac{\varphi(t,x)}{\epsilon^2}\right)\right) \tag{5.25}$$

and the flow near the boundary $V^\epsilon_{bd}(t,x)$ will be developed later, where $c_j(t,x;\theta)$ are 2π-periodic in θ with mean value vanishing.

Plugging the expansion (5.25) into

$$L(\partial_t,\partial_x)V^\epsilon_{in} - B(\epsilon^2, D\epsilon^2)V^\epsilon_{in} - \Phi\left(t,x;\frac{\varphi(t,x)}{\epsilon^2}\right) = 0, \tag{5.26}$$

it follows that the vanishing of the $O(\epsilon^{-2})$, $O(\epsilon^{-1})$, and $O(\epsilon^0)-$terms on the left side of (5.26) give rise to

$$\sum_{k=0}^{2}\varphi_{x_k}A_k(V')\partial_\theta c_0 = (\varphi^2_{x_1}B_1 + \varphi^2_{x_2}B_2 + \varphi_{x_1}\varphi_{x_2}B_3)\partial_\theta^2 c_0, \tag{5.27}$$

$$\sum_{k=0}^{2}\varphi_{x_k}A_k(V')\partial_\theta c_1 = (\varphi^2_{x_1}B_1 + \varphi^2_{x_2}B_2 + \varphi_{x_1}\varphi_{x_2}B_3)\partial_\theta^2 c_1, \tag{5.28}$$

and

$$\begin{aligned} L(\partial_t,\partial_x)(a_0+c_0) + \sum_{k=0}^{2}\varphi_{x_k}A_k(V')\partial_\theta c_2 - (\varphi^2_{x_1}B_1 + \varphi^2_{x_2}B_2 + \varphi_{x_1}\varphi_{x_2}B_3)\partial_\theta^2 c_2 \\ - \Phi(t,x,\theta) = 0, \end{aligned} \tag{5.29}$$

respectively.

Using the assumption (2.6) in (5.27), it follows that

$$\begin{cases} \varphi_{x_1}\partial_\theta c_0^{(2)} + \varphi_{x_2}\partial_\theta c_0^{(3)} = 0, \\ c^2\varphi_{x_1}\partial_\theta c_0^{(1)} = ((1+D)\varphi_{x_1}^2 + \varphi_{x_2}^2)\partial_\theta^2 c_0^{(2)} + D\varphi_{x_1}\varphi_{x_2}\partial_\theta^2 c_0^{(3)}, \\ c^2\varphi_{x_2}\partial_\theta c_0^{(1)} = (\varphi_{x_1}^2 + (1+D)\varphi_{x_2}^2)\partial_\theta^2 c_0^{(3)} + D\varphi_{x_1}\varphi_{x_2}\partial_\theta^2 c_0^{(2)}, \end{cases} \tag{5.30}$$

which implies

$$\varphi_{x_2}\partial_\theta^2 c_0^{(2)} - \varphi_{x_1}\partial_\theta^2 c_0^{(3)} = 0. \tag{5.31}$$

Combining (5.31) and the first equation in (5.30) leads to

$$c_0^{(2)} = c_0^{(3)} \equiv 0 \tag{5.32}$$

by using $\mathbf{m}_\theta(c_0) = 0$.

By substituting (5.32) into the last two equations in (5.30), it follows that

$$c_0^{(1)} \equiv 0. \tag{5.33}$$

Similarly, from (5.28) it follows that

$$c_1(t, x, \theta) \equiv 0. \tag{5.34}$$

From (5.29) we obtain

$$L(\partial_t, \partial_x)a_0 = \mathbf{m}_\theta(\Phi) \tag{5.35}$$

and

$$\sum_{k=0}^{2} \varphi_{x_k} A_k(V')\partial_\theta c_2 - (\varphi_{x_1}^2 B_1 + \varphi_{x_2}^2 B_2 + \varphi_{x_1}\varphi_{x_2} B_3)\partial_\theta^2 c_2 = \Phi - \mathbf{m}_\theta(\Phi). \tag{5.36}$$

which implies that $c_2 = (c_2^{(1)}, c_2^{(2)}, c_2^{(3)})^T$ satisfies

$$\begin{cases} \rho'(\varphi_{x_1}\partial_\theta c_2^{(2)} + \varphi_{x_2}\partial_\theta c_2^{(3)}) = \Phi^{(1)} - \mathbf{m}_\theta(\Phi^{(1)}), \\ c^2\varphi_{x_1}\partial_\theta c_2^{(1)} = ((1+D)\varphi_{x_1}^2 + \varphi_{x_2}^2)\partial_\theta^2 c_2^{(2)} + D\varphi_{x_1}\varphi_{x_2}\partial_\theta^2 c_2^{(3)} + \Phi^{(2)} - \mathbf{m}_\theta(\Phi^{(2)}), \\ c^2\varphi_{x_2}\partial_\theta c_2^{(1)} = (\varphi_{x_1}^2 + (1+D)\varphi_{x_2}^2)\partial_\theta^2 c_2^{(3)} + D\varphi_{x_1}\varphi_{x_2}\partial_\theta^2 c_2^{(2)} + \Phi^{(3)} - \mathbf{m}_\theta(\Phi^{(3)}). \end{cases} \tag{5.37}$$

From (5.37), it follows that

$$\begin{cases} \rho'(\varphi_{x_1}\partial_\theta^2 c_2^{(2)} + \varphi_{x_2}\partial_\theta^2 c_2^{(3)}) = \partial_\theta\Phi^{(1)}, \\ (\varphi_{x_1}^2 + \varphi_{x_2}^2)(\varphi_{x_2}\partial_\theta^2 c_2^{(2)} - \varphi_{x_1}\partial_\theta^2 c_2^{(3)}) = \varphi_{x_1}(\Phi^{(3)} - \mathbf{m}_\theta(\Phi^{(3)})), -\varphi_{x_2}(\Phi^{(2)} - \mathbf{m}_\theta(\Phi^{(2)})) \end{cases} \tag{5.38}$$

from which $c_2^{(2)}(t, x; \theta)$ and $c_2^{(3)}(t, x; \theta)$ are determined uniquely. Consequently, $c_2^{(1)}$ can be uniquely determined from the last two equations in (5.36) as well.

To determine $a_0(t,x)$ from the linearized Euler equations (5.35), one needs to have the boundary data on $\{x_1=0\}$, which shall be deduced from the boundary layer profiles given later. Suppose that this is known already, then we have determined the leading profiles of the flow away from the boundary $\{x_1=0\}$. In this way, one can obtain profiles of $V_{in}^\epsilon(t,x)$ up to any fixed order, which satisfy problems similar to those of a_0 and $c_2(t,x,\theta)$. Therefore, the outflow V_{in}^ϵ has the formal expansion

$$V_{in}^\epsilon(t,x)=\sum_{j\geq 0}\epsilon^j a_j(t,x)+\sum_{j\geq 2}\epsilon^j c_j\left(t,x;\frac{\varphi(t,x)}{\epsilon^2}\right). \tag{5.39}$$

The formula (5.39) inspires us to take the following ansatz for the boundary layer part V_{bd}^ϵ to (2.1):

$$V_{bd}^\epsilon(t,x)=\sum_{j\geq 0}\epsilon^j b_j\left(t,x_2;\frac{x_1}{\epsilon}\right)+\sum_{j\geq 2}\epsilon^j d_j\left(t,x_2;\frac{x_1}{\epsilon^2},\frac{\varphi^0(t,x_2)}{\epsilon^2}\right), \tag{5.40}$$

where $d_j(t,x_2;\eta,\theta)$ are $2\pi-$periodic in θ with mean value vanishing, and $b_j(t,x_2;z)$ and $d_j(t,x_2;\eta,\theta)$ are rapidly decreasing in $z=\frac{x_1}{\epsilon}$ and $\eta=\frac{x_1}{\epsilon^2}$, respectively, when $z,\eta\to+\infty$.

Plugging the ansatz (5.40) into

$$L(\partial_t,\partial_x)V_{bd}^\epsilon-B(\epsilon^2,D\epsilon^2)V_{bd}^\epsilon=0$$

and grouping each power of ϵ, it follows from the vanishing of $O(\epsilon^{-1})$ and $O(\epsilon^0)$-terms that

$$A_1(0)\partial_z b_0=0 \tag{5.41}$$

and

$$\begin{cases}(A_0(0)\partial_t+A_2(0)\partial_{x_2}+zA_1'(0)\partial_z)b_0=B_1\partial_z^2 b_0,\\ (\varphi_t^0A_0(0)+\varphi_{x_2}^0A_2(0))\partial_\theta d_2+A_1(0)\partial_\eta d_2=B_1\partial_\eta^2 d_2+(\varphi_{x_2}^0)^2B_2\partial_\theta^2 d_2+\varphi_{x_2}^0B_3\partial_{\eta\theta}^2 d_2.\end{cases} \tag{5.42}$$

From (5.41), we deduce

$$\partial_z b_0^{(1)}=\partial_z b_0^{(2)}=0,$$

which implies

$$b_0^{(1)}=b_0^{(2)}\equiv 0 \tag{5.43}$$

by using $b_0\in\mathcal{S}(\mathbb{R}_z^+)$.

As in (2.22), from the boundary condition in (2.1) we get

$$\begin{cases}a_0^{(k)}(t,x)+b_0^{(k)}(t,x_2;z)=0, & \text{on } \{x_1=z=0\},\\ c_2^{(k)}(t,x;\theta)+d_2^{(k)}(t,x_2;\eta,\theta^0)=0, & \text{on } \{x_1=\eta=0,\theta=\theta^0\}\end{cases} \tag{5.44}$$

for $k\in\{2,3\}$.

Therefore, from (5.35), (5.43), and (5.44) we obtain that $a_0(t,x)$ satisfies the same initial boundary value problem of the linearized Euler equations (2.26) as for the case studied in section 2. From the first equation in (5.42), it also implies that the boundary layer profile $b_0^{(3)}(t,x_2,z)$ satisfies the same problem of the linearized Prandtl equation (2.29) derived in section 2.

From the second equation in (5.42), we deduce that $d_2=(d_2^{(1)},d_2^{(2)},d_2^{(3)})^T$ satisfies

$$\begin{cases} \partial_\eta d_2^{(2)}+\varphi^0_{x_2}\partial_\theta d_2^{(3)}=0,\\ c^2(0)\partial_\eta d_2^{(1)}=(1+D)\partial_\eta^2 d_2^{(2)}+(\varphi^0_{x_2})^2\partial_\theta^2 d_2^{(2)}+D\varphi^0_{x_2}\partial^2_{\eta\theta}d_2^{(3)},\\ c^2(0)\varphi^0_{x_2}\partial_\theta d_2^{(1)}=\partial_\eta^2 d_2^{(3)}+(1+D)(\varphi^0_{x_2})^2\partial_\theta^2 d_2^{(3)}+D\varphi^0_{x_2}\partial^2_{\eta\theta}d_2^{(2)}. \end{cases} \tag{5.45}$$

From (5.45), it is easy to see that $\omega=\varphi^0_{x_2}\partial_\theta d_2^{(2)}-\partial_\eta d_2^{(3)}$ and $d_2^{(3)}$ satisfy the following problem:

$$\begin{cases} [\partial_\eta^2+(\varphi^0_{x_2})^2\partial_\theta^2]\omega=0,\\ [\partial_\eta^2+(\varphi^0_{x_2})^2\partial_\theta^2]d_2^{(3)}=-\partial_\eta\omega,\\ (\omega+\partial_\eta d_2^{(3)})|_{\eta=0}=-\varphi^0_{x_2}\partial_\theta c_2^{(2)}|_{x_1=0},\\ d_2^{(3)}|_{\eta=0}=-c_2^{(3)}|_{x_1=0}, \end{cases} \tag{5.46}$$

where $c_2^{(2)}$ and $c_2^{(3)}$ are given in (5.38).

In an argument even simpler than in section 3, one concludes that the problem (5.46) admits unique smooth solutions $(d_2^{(3)},\omega)\in\mathcal{C}_p^\infty(T^1,H^\infty([0,T]\times\mathbb{R}_{x_2}\times\mathbb{R}_\eta^+))$ rapidly decreasing in η when $\eta\to+\infty$.

Obviously, from the definition of ω and the first equation in (5.45) we know that $d_2^{(2)}(t,x_2;\eta,\theta)$ satisfies the following problem:

$$\begin{cases} [\partial_\eta^2+(\varphi^0_{x_2})^2\partial_\theta^2]d_2^{(2)}=\varphi^0_{x_2}\partial_\theta\omega,\\ d_2^{(2)}|_{\eta=0}=-c_2^{(2)}(t,0,x_2,\theta), \end{cases} \tag{5.47}$$

which is the same as (2.36). Thus, from (5.47) one can uniquely determine $d_2^{(2)}\in\mathcal{C}_p^\infty(T^1,\,H^\infty([0,T]\times\mathbb{R}_{x_2}\times\mathbb{R}_\eta^+))$ rapidly decreasing in η when $\eta\to+\infty$.

Substituting $(d_2^{(2)},d_2^{(3)})$ into the right-hand side of the third equation in (5.45) we can determine $d_2^{(1)}(t,x_2,\eta,\theta)$ by using $\mathbf{m}_\theta(d_2^{(1)})=0$.

Up to now, we have solved all leading profiles of the solution $V^\epsilon(t,x)$ to (2.1) with $\Phi=\Phi(t,x,\frac{\varphi(t,x)}{\epsilon^2})$ both near and away from the boundary. Similar to the discussion given at the last subsection, one can obtain all other high order profiles for this problem.

In a way similar to section 4, we can obtain the following theorem.

THEOREM 5.2. *Under the assumption that all compatibility conditions are satisfied for the problem* (2.1) *with* $\Phi=\Phi(t,x,\frac{\varphi(t,x)}{\epsilon^2})$, *the solution* V^ϵ *of* (2.1) *has the following asymptotics:*

$$V^\epsilon(t,x)=a_0(t,x)+b_0\left(t,x_2;\frac{x_1}{\epsilon}\right)+O(\epsilon)$$

in $L^\infty([0,T]\times\mathbb{R}^2_+)$ *for any* $T>0$, *where* $a_0(t,x)$ *satisfies the problem for the linearized Euler equations* (2.26), $(b_0^{(1)},b_0^{(2)})=0$, *and* $b_0^{(3)}(t,x_2;z)$ *satisfies the linearized Prandtl equation* (2.29).

Acknowledgement. The authors would like to thank the referees for their suggestion for improving the presentation of this work.

REFERENCES

[1] W. E, *Boundary layer theory and the zero-viscosity limit of the Navier-Stokes equation*, Acta Math. Sin. (Engl. Ser.), 16 (2000), pp. 207–218.

[2] E. Grenier, *On the stability of boundary layers of incompressible Euler equations*, J. Differential Equations, 164 (2000), pp. 180–222.

[3] E. Grenier, *On the nonlinear instability of Euler and Prandtl equations*, Comm. Pure Appl. Math., 53 (2000), pp. 1067–1091.

[4] O. Gues, *Viscous boundary layers and high frequency oscillations*, in Singularities and Oscillations, J. Rauch and M. Taylor eds., IMA Vol. Math. Appl., Springer-Verlag 1997, pp. 61–77.

[5] O. A. Oleinik and V. N. Samokhin, *Mathematical Models in Boundary Layers Theory*, Applied Mathematics and Mathematical Computation, 15, Chapman & Hall/CRC, Boca Raton, FL, 1999.

[6] L. Prandtl, *Über Flüssigkeitsbewegungen bei sehr kleiner Reibung*, in Verh. Int. Math. Kongr., Heidelberg, Germany, 1904, Teubner, Germany, 1905, pp. 484–494.

[7] M. Sammartino and R. E. Caflisch, *Zero viscosity limit for analytic solutions of the Navier-Stokes equations on a half-space,* I. *Existence for Euler and Prandtl equations*, Comm. Math. Phys., 192 (1998), pp. 433–461.

[8] M. Sammartino and R. E. Caflisch, *Zero viscosity limit for analytic solutions of the Navier-Stokes equations on a half-space,* II. *Construction of the Navier-Stokes solution*, Comm. Math. Phys., 192 (1998), pp. 463–491.

[9] A. Szepessy, *High frequency asymptotics for* 2d *viscous shocks*, Indiana Univ. Math. J., 49 (2000), pp. 1623–1671.

[10] R. Temam and X. M. Wang, *Boundary layers associated with incompressible Navier-Stokes equations: the noncharacteristic boundary case*, J. Differential Equations, 179 (2002), pp. 647–686.

[11] Z. P. Xin, *Viscous boundary layers and their stability.* I*,* J. Partial Differential Equations, 11 (1998), pp. 97–124.

[12] Z. P. Xin and T. Yanagisawa, *Zero-viscosity limit of the linearized Navier-Stokes equations for a compressible viscous fluid in the half-plane*, Comm. Pure Appl. Math., 52 (1999), pp. 479–541.

[13] Z. P. Xin and L. Q. Zhang, *On the global existence of solutions to the Prandtl's system*, Adv. Math., 181 (2004), pp. 88–133.

[14] Z. P. Xin, L. Q. Zhang, and J. N. Zhao, *Global well-posedness for the two dimensional Prandtl's boundary layer equations*, preprint.

SIAM J. MATH. ANAL.
Vol. 37, No. 4, pp. 1299–1332

PERIODICITY AND UNIQUENESS OF GLOBAL MINIMIZERS OF AN ENERGY FUNCTIONAL CONTAINING A LONG-RANGE INTERACTION*

XINFU CHEN† AND YOSHIHITO OSHITA‡

Abstract. We consider, on an interval of arbitrary length, global minimizers of a class of energy functionals containing a small parameter ε and a long-range interaction. Such functionals arise from models for phase separation in diblock copolymers and from stationary solutions of FitzHugh–Nagumo systems. We show that every global minimizer is periodic with a period of order $\varepsilon^{1/3}$. Also, we identify the number of global minimizers and provide asymptotic expansions for the periods and global minimizers.

Key words. singular perturbation, elliptic systems, transition layer

AMS subject classifications. 34E15, 35J50

DOI. 10.1137/S0036141004441155

1. Introduction.

We consider global minimizers of the energy functional

$$\mathbf{E}(u,v,\varepsilon,\ell) := \int_0^\ell \left\{ \tfrac{1}{2}\varepsilon^2 u_x^2 + F(u) + \tfrac{1}{2}v_x^2 + \tfrac{1}{2}\gamma v^2 \right\} dx \tag{1.1}$$

under the constraint

$$-v_{xx} + \gamma v = u - m \quad \text{in} \quad (0,\ell), \qquad v_x(0) = v_x(\ell) = 0. \tag{1.2}$$

Here $\gamma \geqslant 0$ and $m \in (0,1)$ are fixed constants, ε is a small positive parameter, and $\ell > 0$ is arbitrary. The function F is a smooth double-equal-well potential; more precisely,

$$F \in C^3(\mathbb{R}), \quad F(0) = F(1) = 0 < F(s) \quad \forall s \in \mathbb{R} \setminus \{0,1\}, \quad F''(0)F''(1) \neq 0. \tag{1.3}$$

When $\gamma = 0$, the solvability of v in (1.2) requires the average $\bar{u}$ of u be equal to m. In this case, the functional (1.1) was first introduced by Ohta and Kawasaki [20], and later by Bahiana and Oono [3], as a free energy modeling a microphase separation in diblock copolymers. In this model, u and $1-u$ represent the concentrations of two different repulsive monomers that constitute the diblock copolymers. The nonlocal term v represents the long-range interaction of copolymer chains. In numerical simulations, phase separations with fine structures were observed to be in agreement with those from laboratory experiments. In [18], Nishiura and Ohnishi derived and studied a gradient flow of the functional defined on a finite interval. For a detailed physical background on diblock copolymers, see Hamley [10]; other related mathematical treatments can be found in [6, 9, 11, 19, 22, 23, 24].

*Received by the editors February 13, 2004; accepted for publication (in revised form) April 25, 2005; published electronically January 6, 2006.
http://www.siam.org/journals/sima/37-4/44115.html
†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (xinfu@pitt.edu).
This author's research was partially supported by National Science Foundation grants DMS–9971043 and 0203991.
‡Laboratory of Nonlinear Studies and Computations, Research Institute for Electronic Science, Hokkaido University, Kita-ku, Sapporo 060-0812, Japan (oshita@nsc.es.hokudai.ac.jp).
This author's research was supported by the Japan Society for the Promotion of Science.

The Euler–Lagrange equation for a critical point of the functional is

$$\varepsilon^2 u_{xx} = F'(u) + v \quad \text{in} \quad (0, \ell), \quad u_x(0) = u_x(\ell) = 0. \tag{1.4}$$

Any solution to the system (1.2), (1.4) can be regarded as a stationary pattern of the FitzHugh–Nagumo equations; see [7, 8, 12, 16, 17] and the references therein. Recently, one of the authors used the functional (1.1) for $\gamma > 0$ obtaining mesoscopic patterns of the FitzHugh–Nagumo dynamics in higher space dimensions [21].

The functional (1.1) admits many local minimizers [16, 19, 22]; this paper focuses on global minimizers. Recently, Ren and Wei [24] carried out an impressive mathematical analysis on global minimizers of (1.1) with $\ell = 1$ and $\gamma = 0$. Their results are in line with a classical result of Müller [15] on energy functionals of the type

$$\int_0^1 \left\{ \varepsilon^2 w_{xx}^2 + (w_x^2 - 1)^2 + w^2 \right\} dx.$$

Under periodic or homogeneous Dirichlet boundary conditions, Müller proved that global minimizers have periods of order $\varepsilon^{1/3}$ and are in general unique; see also [1]. The above energy can be regarded as a special case of (1.1), after taking $\gamma = 0$, $\ell = 1$, $m = 1/2$, and $F(s) = 2s^2(1-s)^2$, and setting $w = -2v_x$. In this paper we provide a unified framework to analyze the stationary diblock copolymer equation ($\gamma = 0$) and the stationary FitzHugh–Nagumo system ($\gamma > 0$). Besides using many ingenious ideas from Müller [15], Alberti and Müller [1], Nishiura [16], Ohnishi et al. [19], and Ren and Wei [22, 23, 24], we develop new techniques to

(i) treat the case $\gamma > 0$,
(ii) remove technical assumptions on F in [1, 15, 24] (here we only assume (1.3)),
(iii) simplify some of their proofs, and
(iv) provide a (seemingly) more complete result by introducing ℓ as a parameter.

The energy functional (1.1) is designed to have two length scales (e.g., [6, 16, 20, 19, 24]). To see this, let's denote the average of u and v over $(0, \ell)$ by

$$\bar{u} = \frac{1}{\ell} \int_0^\ell u(x) dx, \qquad \bar{v} = \frac{1}{\ell} \int_0^\ell v(x) dx.$$

An integration of the equation for v gives $\gamma \bar{v} = \bar{u} - m$. As $\|v\|_{L^2}^2 = \|v - \bar{v}\|_{L^2}^2 + \|\bar{v}\|_{L^2}^2$, the functional can also be written as

$$\mathbf{E}(u, v, \varepsilon, \ell) = \int_0^\ell \left\{ \tfrac{1}{2}\varepsilon^2 u_x^2 + F(u) + \tfrac{1}{2\gamma}(\bar{u} - m)^2 \right\} dx + \int_0^\ell \left\{ \tfrac{1}{2} v_x^2 + \tfrac{\gamma}{2}(v - \bar{v})^2 \right\} dx. \tag{1.5}$$

We call the first integral the *interfacial energy* and the second integral the *interaction energy*.

For the interfacial energy to be small, u has to stay close to either 0 or 1, whereas its average stays close to m. The transition layer (interface) where u changes from 0 to 1 has to be of ε scale; see Carr, Gurtin, and Slemrod [4], who studied minimizers of the interfacial energy functional in the class $\{u \mid \bar{u} = m\}$. Indeed, near each interface located at z, $u(x) \sim Q(\pm(x - z)/\varepsilon)$, where Q is the profile of the transition being the unique solution to

$$\begin{cases} \ddot{Q}(\xi) = f(Q(\xi)) \quad \forall \xi \in \mathbb{R}, \\ Q(-\infty) = 0, \quad Q(\infty) = 1, \quad \int_{\mathbb{R}} \xi \dot{Q}(\xi)\, d\xi = 0. \end{cases}$$

Here $f(u) \equiv F'(u)$ and the last condition is a normalization to fix the translation invariance. Each layer (interface) contains an interfacial energy about $\sigma\varepsilon$, where $\sigma = \int_0^1 \sqrt{2F(s)}ds$; the fewer the interfaces there are, the smaller the total interfacial energy is. The thickness of each interfacial region is of order ε.

On the other hand, the interaction energy is proportional to the cubic power of the length of phase regions (the sets where $u \sim 0$ or 1): when the distance between two interfaces is l and the equation for v is approximated by $v_{xx} \approx m$ or $m-1$, the interaction energy can be calculated to be about $\frac{1}{6}m^2(1-m)^2l^3$. Hence, the energy density

$$\frac{\text{interfacial energy} + \text{interaction energy}}{\text{length}} \approx \frac{1}{l}\left(\sigma\varepsilon + \frac{m^2(1-m)^2}{6}l^3\right).$$

Consequently, the energy density becomes optimal (the smallest) if $l \sim L_0\varepsilon^{1/3}$, where

$$L_0 := \left(\frac{3\sigma}{m^2(1-m)^2}\right)^{1/3}, \qquad \sigma := \int_0^1 \sqrt{2F(s)}ds. \tag{1.6}$$

Thus, at least formally, energies are minimized under an *interaction scale* of size $\varepsilon^{1/3}$ and an *interfacial scale* of size ε. An arbitrary sample of the material studied would in general exhibit a high oscillating pattern of frequency about $\varepsilon^{-1/3}$ per unit length.

To state our analytical result, we introduce the following:

$$\mathbf{K}_N(\ell) = \{(u,v) \mid u \in H^1(0,\ell), v'' - \gamma v = m - u, v'(0) = v'(\ell) = 0\},$$
$$E(\varepsilon,\ell) = \inf_{(u,v)\in\mathbf{K}_N(\ell)} \mathbf{E}(u,v,\varepsilon,\ell), \qquad \rho(\varepsilon,\ell) = \frac{1}{\ell}E(\varepsilon,\ell).$$

Associated with each $(u,v) \in \mathbf{K}_N(\ell)$, there are two natural companions of the same energy density:

(i) *Reflection* about $x = \ell/2$: $(\tilde{u}(x), \tilde{v}(x)) := (u(\ell - x), v(\ell - x))$.

(ii) *Even-periodic extension*: extend (u,v) first evenly to $[-\ell, 0)$ and then periodically to $\mathbb{R}$.

The reflection $(\tilde{u},\tilde{v})$ can be viewed as an ℓ-spatial translation of the even-periodic extension $(\hat{u},\hat{v})$: for every $x \in (0,\ell)$, $\hat{u}(x+\ell) = \hat{u}(x+\ell-2\ell) = \hat{u}(x-\ell) = u(\ell-x) = \tilde{u}(x)$.

Using the even-periodic extensions as test functions, one immediately derives that

$$\rho(\varepsilon, k\ell) \le \rho(\varepsilon,\ell) \quad \text{ or } \quad \rho(\varepsilon,\ell) \le \rho(\varepsilon,\ell/k) \qquad \forall\, \ell > 0,\ k = 2,3,\dots.$$

Hence, for the uniqueness of global minimizers, it would be better to modulo out the periodicity. For this purpose, we introduce

$$\mathbf{K}_N^+(\ell) = \{(u,v) \in \mathbf{K}_N(\ell) \mid v' \ge 0 \text{ in } (0,\ell)\},$$
$$E^+(\varepsilon,\ell) = \inf_{(u,v)\in\mathbf{K}_N^+(\ell)} \mathbf{E}(u,v,\varepsilon,\ell), \qquad \rho^+(\varepsilon,\ell) = \frac{1}{\ell}E^+(\varepsilon,\ell).$$

In what follows, we always assume that F satisfies (1.3), $m \in (0,1)$ is a fixed constant, and $\gamma = \gamma(\varepsilon) \in [0,\gamma_0]$, where γ_0 is a fixed constant. The following technical result, stated first for easy reference, can be regarded as a lemma.

THEOREM 1. *There exists $\varepsilon_0 > 0$ such that when $\varepsilon \in (0,\varepsilon_0]$, the following holds.*

1. *With* σ *and* L_0 *defined as in* (1.6),

$$\text{(1.7)}\qquad \frac{L_0}{\sigma\varepsilon^{2/3}}\,\rho^+(\varepsilon,\ell) \leqslant \sqrt[3]{5} \qquad \text{if} \quad \frac{1}{\sqrt[3]{3}} \leqslant \frac{\ell}{L_0\varepsilon^{1/3}} \leqslant \frac{2}{\sqrt[3]{3}},$$

$$\text{(1.8)}\qquad \frac{L_0}{\sigma\varepsilon^{2/3}}\,\rho^+(\varepsilon,\ell) \geqslant \sqrt[3]{6} \qquad \text{if} \quad \frac{\ell}{L_0\varepsilon^{1/3}} \geqslant \sqrt[3]{4} \quad \text{or} \quad \frac{\ell}{L_0\varepsilon^{1/3}} \leqslant \frac{1}{4}.$$

2. *For each* $\ell \in [\frac{1}{16}L_0\varepsilon^{1/3}, \sqrt[3]{5}L_0\varepsilon^{1/3}]$, *there is a unique minimizer of* $\mathbf{E}(\cdot,\cdot,\varepsilon,\ell)$ *in* $\mathbf{K}_N^+(\ell)$.
3. *As a function of* ℓ, $E^+(\varepsilon,\ell)$ *is smooth in* $[\frac{1}{16}L_0\varepsilon^{1/3}, \sqrt[3]{5}L_0\varepsilon^{1/3}]$ *and, denoting* $E^+_\ell = \frac{d}{d\ell}E^+$,

$$\begin{aligned} E^+(\varepsilon,\ell) &= \sigma\varepsilon + \tfrac{1}{6}m^2(1-m)^2\ell^3 + O(1)\ell^4,\\ E^+_\ell(\varepsilon,\ell) &= \tfrac{1}{2}m^2(1-m)^2\ell^2 + O(1)\ell^3,\\ E^+_{\ell\ell}(\varepsilon,\ell) &= m^2(1-m)^2\ell + O(1)\ell^2,\\ \ell^2\rho^+_\ell(\varepsilon,\ell) &= \tfrac{1}{3}m^2(1-m)^2\ell^3 - \sigma\varepsilon + O(1)\ell^4,\\ \ell^3\rho^+_{\ell\ell}(\varepsilon,\ell) &= \tfrac{1}{3}m^2(1-m)^2\ell^3 + 2\sigma\varepsilon + O(1)\ell^4. \end{aligned}$$

In particular, both $E^+(\varepsilon,\cdot)$ *and* $\rho^+(\varepsilon,\cdot)$ *are strictly convex in*

$$\left[\frac{1}{16}L_0\varepsilon^{1/3}, \sqrt[3]{5}L_0\varepsilon^{1/3}\right].$$

4. *There is a unique positive constant* $\ell^\varepsilon = L_0\varepsilon^{1/3}[1+O(1)\varepsilon^{1/3}]$ *such that*

$$\rho^+(\varepsilon,\ell) > \rho^+(\varepsilon,\ell^\varepsilon) \qquad \forall \ell \in (0,\ell^\varepsilon)\cup(\ell^\varepsilon,\infty).$$

In addition, for each positive integer n, *there is a unique solution* $\ell_n = \ell_n(\varepsilon)$ *to*

$$\text{(1.9)}\qquad \rho^+\left(\varepsilon, \frac{\ell_n}{n+1}\right) = \rho^+\left(\varepsilon, \frac{\ell_n}{n}\right), \qquad \frac{\ell_n}{n+1} < \ell^\varepsilon < \frac{\ell_n}{n}.$$

Furthermore, uniformly in $n \geq 1$, *the solution has the expansion*

$$\text{(1.10)}\qquad \frac{\ell_n(\varepsilon)}{\ell^\varepsilon} = \left(n+\tfrac{1}{2}\right)\left[\frac{n^2+n}{n^2+n+1/4}\right]^{2/3} + O(1)\varepsilon^{1/3}.$$

In this paper, every $O(1)$ refers to a quantity that is bounded uniformly in ε and in ℓ.

The following is our main result, where $\ell_0(\varepsilon) := \ell_1(\varepsilon)/2$.

Theorem 2 (periodicity and uniqueness of global minimizers). *For every sufficiently small positive* ε *the following holds:*

1. *For every integer* $n \geq 1$ *and* $\ell \in (\ell_{n-1}(\varepsilon), \ell_n(\varepsilon))$, (1.1) *under the constraint* (1.2) *has a global minimizer. It must have a half-period* ℓ/n *and is unique up to a half-period translation.*

2. *If* $\ell = \ell_n$ $(n \geq 1)$, (1.1) *under the constraint* (1.2) *has exactly (modulo a half-period translation) two minimizers; one has a half-period* ℓ/n *and the other a half-period* $\ell/(n+1)$.

In both cases 1 *and* 2, *a global minimizer in a half-period l is the unique global minimizer of $\mathbf{E}(\cdot,\cdot,\varepsilon,l)$ in $\mathbf{K}_N^+(l)$ or its reflection about the quarter-period $x = l/2$.*

Here letting $\hat{u}$ be the even-periodic extension of u, we call $\hat{u}(\cdot - l)$ the half-period translation of u if $\hat{u}$ has the period $2l$.

Remark 1.1. (1) Suppose $\{(u^\ell, v^\ell)\}_{\ell\geq\ell_0(\varepsilon)}$ is a family of global minimizers normalized so that $v^\ell_{xx}(0) \geq 0$. Then their even-periodic extensions satisfy

$$\lim_{\ell\to\infty}(u^\ell, v^\ell) = (u^{\ell^\varepsilon}, v^{\ell^\varepsilon}) = (u^{k\ell^\varepsilon}, v^{k\ell^\varepsilon}) \qquad \forall x \in (-\infty,\infty), k = 2,3,\ldots,$$

$$\lim_{\ell\to\infty}\rho(\varepsilon,\ell) = \rho(\varepsilon,\ell^\varepsilon) = \rho(\varepsilon,k\ell^\varepsilon) \qquad \forall\, k = 2,3,\ldots.$$

These limits follow from the continuous dependence of global minimizers in $\mathbf{K}_N^+(\ell)$ for ℓ near ℓ^ε.

(2) The maximum period of all minimizers (for $\ell \geq \ell^\varepsilon$) is

$$2\ell_1 = \frac{4L_0}{\sqrt[3]{3}}\,\varepsilon^{1/3} + O(1)\varepsilon^{2/3}.$$

Namely, every global minimizer in any length of interval has a fine structure of length scale $\varepsilon^{1/3}$.

(3) Theorem 1 part 3 is indeed the announced Proposition 4.7 in [19], stating that the minimum energy and the corresponding energy density are convex functions of ℓ in a neighborhood of $L_0\varepsilon^{1/3}$.

The paper is organized as follows. In section 2, we use Theorem 1 to prove our main result, Theorem 2. The proof of part 4 of Theorem 1 is also included. In section 3 and section 4, we establish energy upper and lower bounds, respectively; that is, we prove Theorem 1 part 1. The rest of the paper is devoted to the proof of Theorem 1 parts 2 and 3. First we provide certain estimates on energy minimizers in section 5; in particular, we provide rigorously up to $O(\varepsilon\ell)$ order expansions. Then in section 6 we estimate the principal eigenvalue of the linearized operator around any minimizer. We remark that Nishiura [16] has already proven the positivity of the principal eigenvalue. Here we identify its precise value. The estimate shows that minimizers in any $o(\ell^2)$ neighborhood are unique. Since the expansion provided in section 5 for minimizers is unique, we conclude that minimizers, for ℓ close to $L_0\varepsilon^{1/3}$, are unique. The uniqueness of minimizers allows us to differentiate the minimizer with respect to ℓ and to prove the convexity of minimum energy with respect to ℓ, which will be done in the last section.

2. Proof of the main result. In this section, we use Theorem 1 to prove our main result, Theorem 2.

2.1. The idea. The key to the proof is to show that all zeros of v_x of a global minimizer (u,v) are equally spaced, with a distance approximately equal to $L_0\varepsilon^{1/3}$. Once this is done, the rest of the proof follows from a straightforward calculation using the estimates of ρ^+ and the uniqueness of global minimizers in $\mathbf{K}_N^+$ stated in Theorem 1.

To show that the zeros of v_x are equally spaced, we compare the energy $E(\varepsilon,\ell)$ with $H(\varepsilon,P) := \sum_{i=1}^n E^+(\varepsilon,l_i)$, where $P = \{l_1,\ldots,l_n\}$ is an arbitrary partition of ℓ, defined as follows.

Definition 1. *A* partition *P of a positive number ℓ is a collection of finitely many positive numbers $l_1,\ldots,l_n$ satisfying $\sum_{i=1}^n l_i = \ell$. A partition $P = \{l_1,\ldots,l_n\}$ is called an* equal partition *if all $l_1,\ldots,l_n$ are identical, namely, equal to ℓ/n.*

Using a step-by-step modification of partition lengths, we show that if $E(\varepsilon,\ell) \geq H(\varepsilon,P)$, where P is a partition of ℓ, then P must be an equal partition with partition length as "close" to ℓ^ε as possible. The modification involves combining short lengths with, and splitting long lengths into, intermediate ones. The criteria for short, intermediate, and long are the two numbers $\frac{1}{16}L_0\varepsilon^{1/3}$ and $\sqrt[3]{4}L_0\varepsilon^{1/3}$.

To treat minimizers on short intervals, we need the following energy lower bound estimate whose proof will be given later in this section.

LEMMA 2.1. *There exists a positive constant c_1 (depending only on m and F) such that*

$$\rho(\varepsilon,\ell) \geqslant \min\{\, c_1\,,\ \sigma\varepsilon/(2\ell)\,\} \qquad \forall \varepsilon > 0, \ell > 0.$$

2.2. Proof of Theorem 2. Let $\ell \geq \frac{1}{\sqrt[3]{3}}L_0\varepsilon^{1/3}$ be given and consider the minimization of (1.1) under the constraint (1.2).

Step 1. Since each term that appeared in $\mathbf{E}$ defined in (1.1) is nonnegative, one can easily show that in $\mathbf{K}_N(\ell)$, $\mathbf{E}(\cdot,\cdot,\ell,\varepsilon)$ admits at least one global minimizer. In addition, any global minimizer is smooth and satisfies the Euler–Lagrange equation (1.4). Furthermore, applying a classical uniqueness result to an initial value problem of the ode system (1.2), (1.4), one can conclude that if all $v_x, v_{xx}, v_{xxx}, v_{xxxx}$ vanish at the same point, say $y \in [0,\ell]$, then $u_x(y) = u_{xx}(y) = 0$, which implies that $(u_x(x), v_x(x)) \equiv (0,0)$ for all $x \in [0,\ell]$; namely (u,v) is a constant function.

Now let $(u,v) \in \mathbf{K}_N(\ell)$ be an arbitrary global minimizer of $\mathbf{E}(\cdot,\cdot,\ell,\varepsilon)$ in $\mathbf{K}_N(\ell)$; that is, $\mathbf{E}(u,v,\varepsilon,\ell) = E(\varepsilon,\ell)$. Assume that ε is small. We see $\rho(\varepsilon,\ell)$ is small uniformly in $\ell \geq L_0\varepsilon^{1/3}/\sqrt[3]{3}$. Then (u,v) cannot be a constant function, and thus all $v_x, v_{xx}, v_{xxx}, v_{xxxx}$ cannot vanish simultaneously at any point. Hence all roots to $v_x = 0$ are isolated. We denote all the roots to $v' = 0$ in $[0,\ell]$ by $\{x_0, x_1, \ldots, x_n\}$, in increasing order with $x_0 = 0$ and $x_n = \ell$. Set $l_i = x_i - x_{i-1}$. In each interval (x_{i-1}, x_i), $i = 1, \ldots, n$, either $v_x > 0$ or $v_x < 0$. Set $(u_i(y), v_i(y)) = (u(x_{i-1}+y), v(x_{i-1}+y))$ in the former case and set $(u_i(y), v_i(y)) = (u(x_i - y), v(x_i - y))$ in the latter case. Then $(u_i, v_i) \in \mathbf{K}_N^+(l_i)$. Consequently,

$$\int_{x_{i-1}}^{x_i} \left[\tfrac{1}{2}u_x^2 + F(u) + \tfrac{1}{2}v_x^2 + \tfrac{\gamma}{2}v^2\right] dx = \mathbf{E}(u_i, v_i, \varepsilon, l_i) \geq E^+(\varepsilon, l_i).$$

Hence, for the partition $P := \{l_1, \ldots, l_n\}$ of ℓ,

$$H(\varepsilon,P) := \sum_{i=1}^n E^+(\varepsilon, l_i) \leq E(\varepsilon,\ell). \tag{2.1}$$

If each l_i is in the range $[L_0\varepsilon^{1/3}/16, \sqrt[3]{5}L_0\varepsilon^{1/3}]$, we can apply the convexity of E^+ to show that all l_i are the same (Step 5). If not, we show (Steps 2–4) that different partitions with decreased values of H could be found, and that after finitely many steps the partition found has the property that all its partition lengths lie in $[L_0\varepsilon^{1/3}/16, \sqrt[3]{5}L_0\varepsilon^{1/3}]$.

Step 2. We shall show that short partition lengths, if there are any, can combine with some others to form new partitions with smaller H. For this we need the following:

Any partition $\tilde{P} = \{\tilde{l}_1, \ldots, \tilde{l}_{\tilde{n}}\}$ of ℓ satisfying $H(\varepsilon,\tilde{P}) \leqslant E(\varepsilon,\ell)$ admits at least one $\tilde{l}_j$, $j \in \{1, \ldots, \tilde{n}\}$ such that

$$\tfrac{\tilde{l}_j}{(L_0\varepsilon^{1/3})} \in \left[\tfrac{1}{4}, \sqrt[3]{4}\right]. \tag{2.2}$$

Indeed, since $\rho(\varepsilon,\ell) \le \rho(\varepsilon,\ell/k) \le \rho^+(\varepsilon,\ell/k)$ for every integer $k \ge 1$, dividing $H(\varepsilon,\tilde{P}) \le E(\varepsilon,\ell)$ by ℓ gives us

$$\sum_{i=1}^{\tilde{n}} \frac{\tilde{l}_i}{\ell}\, \rho^+(\varepsilon,\tilde{l}_i) \le \rho(\varepsilon,\ell) \le \min_{k\ge 1} \rho^+\left(\varepsilon,\frac{\ell}{k}\right) \le \sqrt[3]{5}\, \frac{\sigma\varepsilon^{2/3}}{L_0}$$

by picking a test k such that $\ell/(kL_0\varepsilon^{1/3}) \in [1/\sqrt[3]{3}, 2/\sqrt[3]{3}]$ and using (1.7). Since $\sum \tilde{l}_i = \ell$, in view of (1.8), we see that there is at least one $j \in \{1,\dots,\tilde{n}\}$ such that (2.2) holds.

Step 3. We now combine, one at a time, the short partition lengths, if there are any, with intermediate ones.

Suppose, for some i, that $l_i/(L_0\varepsilon^{1/3}) \le \frac{1}{16}$. Pick an arbitrary l_j satisfying (2.2). Then $l_i + l_j \le (\frac{1}{16} + \sqrt[3]{4})L_0\varepsilon^{1/3} < \sqrt[3]{5}L_0\varepsilon^{1/3}$. We can calculate, by the mean value theorem,

$$E^+(\varepsilon,l_i) + E^+(\varepsilon,l_j) - E^+(\varepsilon,l_j+l_i) = l_i\left\{\rho^+(\varepsilon,l_i) - \tfrac{d}{d\ell}E^+(\varepsilon,l_j+\theta l_i)\right\}$$

for some $\theta \in (0,1)$. Using Lemma 2.1 we find that $\rho^+(\varepsilon,l_i) \ge \rho(\varepsilon,l_i) \ge 8\sigma\varepsilon^{2/3}/L_0$. On the other hand, using Theorem 1 we have $\frac{d}{d\ell}E^+(\varepsilon,l_j+\theta l_i) \le \frac{1}{2}m^2(1-m)^2(\sqrt[3]{5}L_0\varepsilon^{1/3})^2 = \frac{3\sqrt[3]{25}}{2}\sigma\varepsilon^{2/3}/L_0$, since $L_0^3 = 3\sigma/[m^2(1-m)^2]$. Thus,

$$E^+(\varepsilon,l_i) + E^+(\varepsilon,l_j) > E^+(\varepsilon,l_i+l_j).$$

Now deleting two members l_i and l_j from and adding one new member $l_i + l_j$ to P, we obtain a new partition $\tilde{P}$ satisfying $H(\varepsilon,\tilde{P}) < H(\varepsilon,P) \leqslant E(\varepsilon,\ell)$. Repeating the same process finitely many times, we then obtain a partition $P' = \{l'_1,\dots,l'_{n'}\}$ of ℓ such that $l'_i/(L_0\varepsilon^{1/3}) > \frac{1}{16}$ for all $i = 1,\dots,n'$ and $H(\varepsilon,P') \leqslant H(\varepsilon,P)$, where the strict inequality holds if $P' \ne P$.

Step 4. Here we split longer partition lengths, if there are any, into shorter ones.

Suppose $l'_i/(L_0\varepsilon^{1/3}) \ge \sqrt[3]{4}$ for some i. For each such i, let $k_i > 1$ be an integer such that $l'_i/(k_iL_0\varepsilon^{1/3}) \in [1/\sqrt[3]{3}, 2/\sqrt[3]{3}]$. Then, by (1.7) and (1.8),

$$\begin{aligned} E^+(\varepsilon,l'_i) &= l'_i\rho^+(\varepsilon,l'_i) \ge l'_i\sqrt[3]{6}\sigma\varepsilon^{2/3}/L_0 \ge l'_i\sqrt[3]{\tfrac{6}{5}}\;\rho^+(\varepsilon,l'_i/k_i) \\ &= \sqrt[3]{\tfrac{6}{5}}\sum_{j=1}^{k_i}(l'_i/k_i)\rho^+(\varepsilon,l'_i/k_i) = \sqrt[3]{\tfrac{6}{5}}\sum_{j=1}^{k_i}E^+(\varepsilon,l_i^j), \quad l_i^1 = \cdots = l_i^{k_i} := l'_i/k_i. \end{aligned}$$

Hence, splitting each such l'_i into k_i copies of l'_i/k_i ($k_i = 1$ if $l'_i < \sqrt[3]{4}L_0\varepsilon^{1/3}$) we obtain a new partition $P'' = \cup_{i=1}^{n'}\cup_{j=1}^{k_i}\{l_i^j\}$ of ℓ satisfying, after renaming the entries of P'',

(2.3)
$$H(\varepsilon,P'') \le H(\varepsilon,P'), \qquad P'' = \{l''_1,\dots,l''_{n''}\},\ \tfrac{1}{16} < l''_i/(L_0\varepsilon^{1/3}) < \sqrt[3]{4}\ \ \forall\, i = 1,\dots,n'',$$

where the strict inequality holds if $P'' \ne P'$.

Step 5. Now we can use the convexity of $E^+(\varepsilon,\cdot)$ in $[L_0\varepsilon^{1/3}/16, \sqrt[3]{5}L_0\varepsilon^{1/3}]$ to conclude that

$$E(\varepsilon,\ell) \ge H(\varepsilon,P'') = n''\sum_{i=1}^{n''}\tfrac{1}{n''}E^+(\varepsilon,l''_i) \ge n''E^+\Big(\varepsilon,\Sigma_{i=1}^{n''}\tfrac{l''_i}{n''}\Big) = n''E^+(\varepsilon,\tfrac{\ell}{n''}).$$

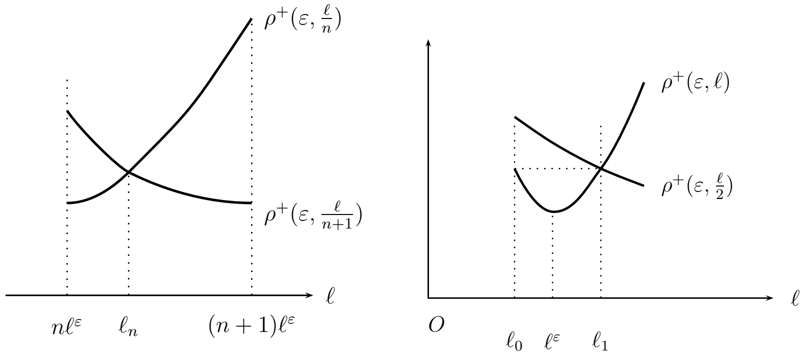


FIG. 1.

Dividing both sides by ℓ we then obtain $\rho^+(\varepsilon, \ell/n'') \leq \rho(\varepsilon, \ell) \leq \min_{k\geq 1} \rho^+(\varepsilon, \ell/k)$. This implies that

$$\rho^+(\varepsilon, \ell/n'') = \rho(\varepsilon, \ell) = \min_{k\geq 1} \rho^+(\varepsilon, \ell/k). \tag{2.4}$$

Since any one of the nontrivial reductions in Steps 3 and 4 gives a strict inequality, we conclude that $P = P' = P''$. In addition, for each i, (u_i, v_i) must be equal to the unique minimizer of $\mathbf{E}$ in $\mathbf{K}_N^+(\ell/n)$. Thus, (u, v) is even with a half-period ℓ/n (the half-period cannot be an integer fraction of ℓ/n since $v' \neq 0$ in $(0, \ell/n)$).

Step 6. It remains to identify n. Let $s \geq 1$ be the unique integer such that $\ell/(s+1) < \ell^\varepsilon \leq \ell/s$. Consider three cases:

(i) $\rho^+(\varepsilon, \ell/(s+1)) > \rho^+(\varepsilon, \ell/s)$,

(ii) $\rho^+(\varepsilon, \ell/(s+1)) < \rho^+(\varepsilon, \ell/s)$,

(iii) $\rho^+(\varepsilon, \ell/(s+1)) = \rho^+(\varepsilon, \ell/s)$.

Here we include Figure 1 to help readers to follow our argument.

In the first case we must have $n = s$. Indeed if $n < s$, then $\ell/n > \ell/s \geq \ell^\varepsilon$. As $\rho^+(\varepsilon, \cdot)$ is strictly increasing in $[\ell^\varepsilon, \sqrt[3]{5}L_0\varepsilon^{1/3}]$, we derive that $\rho^+(\varepsilon, \ell/n) > \rho^+(\varepsilon, \ell/s)$, contradicting (2.4). Similarly, if $n > s$, then $\ell/n \leq \ell/(1+s) < \ell^\varepsilon$. As $\rho^+(\varepsilon, \cdot)$ is strictly decreasing in $(\frac{1}{16}L_0\varepsilon^{1/3}, \ell^\varepsilon)$, $\rho^+(\varepsilon, \ell/n) \geq \rho^+(\varepsilon, \ell/(s+1)) > \rho^+(\varepsilon, \ell/s)$, again contradicting (2.4). Hence, we must have $n = s$.

In case (ii), we can derive in a similar manner that $n = s + 1$.

Finally, in case (iii), we can only have either $n = s$ or $n = s + 1$.

This completes the proof of Theorem 2. $\square$

Remark 2.1. Our proof is simpler than that in [15] and [24] since we do not need any a priori estimation on the distances between any successive zeros of v_x. In [24], great efforts (tens of pages) were devoted to the estimation that $C_1\varepsilon^{1/3} < x_i - x_{i-1} < C_2\varepsilon^{1/3}$, whereas in [15] such estimation was bypassed after imposing more symmetric conditions on the nonlinearity F and on m. Our proof relies on estimates (1.7) and (1.8).

Since the last assertion of Theorem 1 follows from the previous assertion of the same theorem, it is convenient to provide its proof here.

2.3. Proof of Theorem 1 part 4. (a) From the third assertion of Theorem 1, $\rho^+(\varepsilon, \cdot)$ is strictly convex in $[\frac{1}{16}L_0\varepsilon^{1/3}, \sqrt[3]{5}L_0\varepsilon^{1/3}]$ and $\frac{d}{d\ell}\rho^+$ changes sign in the interval. Hence $\rho^+(\varepsilon, \cdot)$ attains a unique strict minimum in $[\frac{1}{16}L_0\varepsilon^{1/3}, \sqrt[3]{5}L_0\varepsilon^{1/3}]$. Denote this minimum by ℓ^ε. Then $\rho^+_\ell(\varepsilon, \ell^\varepsilon) = 0$ so that $\ell^\varepsilon/(L_0\varepsilon^{1/3}) = 1 + O(1)\varepsilon^{1/3}$. In view of (1.7), and (1.8), this minimum is also global.

(b) Now we solve (1.9) for ℓ_n. Writing $\ell_n/n = \mu\ell^\varepsilon$, we need only find $\mu \in (1, 1 + 1/n)$ such that

$$\rho^+\left(\varepsilon, \frac{n}{n+1}\mu\ell^\varepsilon\right) - \rho^+(\varepsilon, \mu\ell^\varepsilon) = 0.$$

Since $\rho^+(\varepsilon, \cdot)$ is strictly increasing in $[\ell^\varepsilon, \sqrt[3]{5}L_0\varepsilon^{1/3}]$ and strictly decreasing in $[L_0\varepsilon^{1/3}/16, \ell^\varepsilon]$, as a function of $\mu \in [1, 1 + 1/n]$, the left-hand side is decreasing and has different signs at $\mu = 1$ and $\mu = 1 + 1/n$. Thus the above equation admits a unique solution $\mu \in (1, 1 + 1/n)$.

It remains to estimate μ. For $\ell \in [L_0\varepsilon^{1/3}/16, \sqrt[3]{5}L_0\varepsilon^{1/3}]$, writing $\rho^+(\varepsilon, \ell)$ as $\rho^+(\ell)$ and using $\rho_\ell(\ell^\varepsilon) = 0$, we obtain

$$\begin{aligned}
\rho^+(\ell) - \rho^+(\ell^\varepsilon) &= \int_{\ell^\varepsilon}^{\ell} \rho^+_s(s)\,ds = \int_{\ell^\varepsilon}^{\ell} (\ell - s)\rho^+_{ss}(s)\,ds \\
&= \int_{\ell^\varepsilon}^{\ell} (\ell - s)\left\{\frac{2\sigma\varepsilon}{s^3} + \frac{m^2(1-m)^2}{3} + O(s)\right\} ds \\
&= \frac{m^2(1-m^2)}{6}(\ell - \ell^\varepsilon)^2\left\{O(\ell) + 1 + \frac{2\ell^\varepsilon}{\ell}\right\} \\
&= \frac{m^2(1-m^2)}{6}\left\{\ell^2 + \frac{2\ell^{\varepsilon 3}}{\ell} - 3\ell^{\varepsilon 2} + O(\ell)(\ell - \ell^\varepsilon)^2\right\}
\end{aligned}$$

by using the expansion for $\rho^+_{\ell\ell}$ in Theorem 1 part 3, the definition of L_0, and expansion $\ell^\varepsilon = [1 + O(\varepsilon^{1/3})]L_0\varepsilon^{1/3}$.

Upon setting $\ell = \mu\ell^\varepsilon$ and $\ell = n\mu\ell^\varepsilon/(n+1)$, respectively, and noting that $\mu \in (1, 1 + 1/n)$ we obtain

$$\mu^2 + \frac{2}{\mu} = \frac{n^2\mu^2}{(n+1)^2} + \frac{2(n+1)}{n\mu} + \frac{O(\ell^\varepsilon)}{n^2}.$$

Therefore,

$$\begin{aligned}
&\mu = \left[\frac{2(n+1)^2}{(2n+1)n}\right]^{1/3}\left[1 + \left(\frac{O(\ell^\varepsilon)}{n}\right)\right], \\
&\frac{\ell_n}{\ell^\varepsilon} = n\mu = \left[\frac{n^2(n+1)^2}{(n+1/2)}\right]^{1/3}\left[1 + \frac{O(\ell^\varepsilon)}{n}\right] = \left(n + \frac{1}{2}\right)\left(\frac{n^2+n}{n^2+n+1/4}\right)^{2/3} + O(\varepsilon^{1/3}).
\end{aligned}$$

This completes the proof of Theorem 1 part 4. $\square$

2.4. Proof of Lemma 2.1. Let δ be a small positive fixed number that is independent of ε and let $(u, v) \in \mathbf{K}_N(\ell)$ be any function. We consider three cases:

(i) There exist x_1 and x_2 in $[0, \ell]$ such that $u(x_1) < \delta$ and $u(x_2) > 1 - \delta$,

(ii) $u \leq 1 - \delta$ in $[0, \ell]$, and

(iii) $u \geq \delta$ in $[0, \ell]$.

In case (i), since δ is small,

$$\mathbf{E}(u,v,\varepsilon,\ell) \geq \int_0^\ell \varepsilon |u_x| \sqrt{2F(u)} dx \geq \varepsilon \int_\delta^{1-\delta} \sqrt{2F(s)} ds \geq \tfrac{1}{2}\sigma\varepsilon.$$

In case (ii) we consider two subcases:

$$\text{(a) } |\{x \mid \delta \leq u \leq 1-\delta\}| \geq \tfrac{1}{2}m\ell; \quad \text{(b) } |\{x \mid \delta \leq u \leq 1-\delta\}| < \tfrac{1}{2}m\ell.$$

In case (a),

$$\mathbf{E}(u,v,\varepsilon,\ell) \geq \int_0^\ell F(u)\, dx \geq \frac{1}{2} m\ell c(\delta), \quad c(\delta) := \min_{\delta < s < 1-\delta} F(s).$$

In case (b), $\ell\bar{u} = \int_0^\ell u \leq \frac{1}{2}(1-\delta)m\ell + \delta\ell \leq \frac{3}{4}m\ell$ since δ is small. Thus, $|\bar{u} - m| \geq \frac{1}{4}m$ and, by (1.5),

$$\mathbf{E}(u,v,\varepsilon,\ell) \geq \frac{1}{2\gamma} \int_0^\ell (\bar{u} - m)^2 \geq \frac{1}{32\gamma} m^2 \ell.$$

Similarly, we can consider case (iii). Hence, combining all the cases we see that

$$1/\ell\, \mathbf{E}(u,v,\varepsilon,\ell) \geq \min\{\sigma\varepsilon/(2\ell), mc(\delta)/2, (1-m)c(\delta)/2, m^2/(32\gamma), (1-m)^2/(32\gamma)\}.$$

Taking $c_1 := \min\{m(1-m)c(\delta)/2, m^2(1-m)^2/(32\gamma)\}$, the assertion of Lemma 2.1 follows. $\square$

3. An energy upper bound. In what follows, we use notation

$$\psi = \mathbf{K}_N^\ell \phi \iff \begin{cases} -\psi'' + \gamma\psi = \phi & \text{in } (0,\ell), \\ \psi'(0) = \psi'(\ell) = 0. \end{cases} \tag{3.1}$$

When $\gamma = 0$, ψ is unique up to an additive constant which does not affect the energy at all.

Lemma 3.1. *For every $\varepsilon > 0$ and $\ell > 0$,*

$$\min_{(u,v)\in\mathbf{K}_N^+(\ell)} \mathbf{E}(u,v,\ell,\varepsilon) < \sigma\varepsilon + \tfrac{1}{6}m^2(1-m)^2\ell^3. \tag{3.2}$$

Consequently, (1.7) *holds for every $\varepsilon > 0$.*

Proof. (a) Test function. With $\mathbf{K}_N^\ell$ defined in (3.1), we choose a test function

$$u(x) = Q\Big(\frac{x-z}{\varepsilon}\Big), \qquad v = \mathbf{K}_N^\ell(u - m),$$

where z is a constant chosen such that the average of u over $(0,\ell)$ is m. Such a z exists and is unique since $\dot{Q} > 0$, $Q(-\infty) = 0$, and $Q(\infty) = 1$. Also, since $-(v_x)_{xx} + \gamma v_x = u_x > 0$, and $v_x(0) = v_x(\ell) = 0$, the maximum principle implies that $v_x > 0$ on $(0,\ell)$. Thus, $(u,v) \in \mathbf{K}_N^+(\ell)$.

(b) Interfacial energy. Using $\dot{Q} = \sqrt{2F(Q)} > 0$, we have $\varepsilon u_x = \sqrt{2F(u)}$ and thus

$$\int_0^\ell \left\{\tfrac{1}{2}\varepsilon^2 u_x^2 + F(u)\right\} dx = \varepsilon \int_0^\ell u_x \sqrt{2F(u)} dx < \varepsilon \int_0^1 \sqrt{2F(s)} ds = \sigma\varepsilon.$$

(c) Interaction energy. We compare the energies of v and $\tilde{v} \equiv \int_0^x \int_0^{x'} [m - u(x'')]dx''dx'$.

Note $\tilde{v}_x = \int_0^x (m - u(x'))dx'$ and $\tilde{v}_{xx} = m - u$. Since the average of u is m, $\tilde{v}_x(\ell) = 0$. Also, as $-\tilde{v}_{xxx} = u_x > 0$, the maximum principle implies that $\tilde{v}_x > 0$ in $(0, \ell)$. Furthermore, $\tilde{v}_x(x) = \int_0^x (m - u(x'))dx' < mx$, $\tilde{v}_x(x) = \int_x^\ell (u(x') - m)dx' < (1 - m)(\ell - x)$. Hence,

$$0 < \tilde{v}_x(x) < \min\{mx, (1 - m)(\ell - x)\} \quad \forall\, x \in (0, \ell).$$

Finally, using $-v_{xx} + \gamma v = u - m = -\tilde{v}_{xx}$ we calculate

$$\begin{aligned} \int_0^\ell (v_x^2 + \gamma v^2)\, dx &= \int_0^\ell v(-v_{xx} + \gamma v)\, dx = \int_0^\ell v(u - m)\, dx \\ &= -\int_0^\ell v\tilde{v}_{xx}\, dx = \int_0^\ell v_x \tilde{v}_x \le \tfrac{1}{2} \int_0^\ell (v_x^2 + \tilde{v}_x^2)\, dx. \end{aligned}$$

Thus

$$\int_0^\ell (v_x^2 + 2\gamma v^2)\, dx \le \int_0^\ell \tilde{v}_x^2 \le \int_0^\ell \Big(\min\{mx, (1 - m)(\ell - x)\}\Big)^2 dx = \tfrac{1}{3} m^2 (1 - m)^2 \ell^3.$$

Combining this with the estimation in (b), the assertion (3.2) thus follows.

(d) Proof of (1.7). Since $m^2(1 - m)^2 = 3\sigma/L_0^3$, (3.2) implies, for every $\varepsilon > 0$ and $\ell = \mu L_0 \varepsilon^{1/3}$ with $\mu \in [1/\sqrt[3]{3}, 2/\sqrt[3]{3}]$, that

$$\rho^+(\varepsilon, \ell) < \frac{\sigma\varepsilon + \sigma\ell^3/(2L_0^3)}{\ell} = \frac{\sigma\varepsilon^{2/3}}{L_0}\Big(\frac{1}{\mu} + \frac{\mu^2}{2}\Big) \le \frac{7\sqrt[3]{3}}{6}\frac{\sigma\varepsilon^{2/3}}{L_0} < \sqrt[3]{5}\frac{\sigma\varepsilon^{2/3}}{L_0}. \qquad \square$$

4. Energy lower bounds. Lemma 2.1 provides an energy lower bound that can be used to prove (1.8) for all $\ell < \frac{1}{4} L_0 \varepsilon^{1/3}$. Here we establish energy lower bounds for all $\ell \ge \frac{1}{16} L_0 \varepsilon^{1/3}$, aimed at (i) completing the proof of (1.8) and (ii) sandwiching tightly the energy with the upper bound (3.2) for $\ell \in [\frac{1}{16} L_0 \varepsilon^{1/3}, \sqrt[3]{5} L_0 \varepsilon^{1/3}]$ so that a leading order expansion of the energy minimizer can be obtained.

4.1. The idea about how to deal with arbitrarily large . When $\ell/\varepsilon^{1/3}$ is bounded uniformly in ε, estimates can be obtained via a limiting process, e.g., the Γ-convergence presented in the next subsection. However, for arbitrarily large ℓ, this process may not work. For this reason, we shall not estimate $\rho^+(\varepsilon, \ell)$ directly for large ℓ. Instead, we estimate $\rho_*(\varepsilon, \ell)$ defined as follows:

$$\begin{gathered} \mathbf{K}^+(\ell) = \{(u, v)|\ u \in H^1(0, \ell), v'' - \gamma v = m - u, v' \ge 0 \ \text{ in } \ (0, \ell)\}, \\ E_*(\varepsilon, \ell) = \inf_{(u,v)\in\mathbf{K}^+(\ell)} \mathbf{E}(u, v, \varepsilon, \ell), \quad \rho_*(\varepsilon, \ell) = \frac{1}{\ell} E_*(\varepsilon, \ell). \end{gathered}$$

There are the relations, for every $\varepsilon > 0$, $\ell > 0$, and positive integer k,

$$\rho^+(\varepsilon, \ell) \ge \rho_*(\varepsilon, \ell) \ge \rho_*(\varepsilon, \ell/k). \tag{4.1}$$

Here the first inequality follows from the fact that the class $\mathbf{K}_N^+(\ell)$ of test functions for ρ^+ is a subset of the class $\mathbf{K}^+(\ell)$ of test functions for ρ_*. For the second inequality, one observes that any portion of a minimizer in $\mathbf{K}^+$ can be extracted to be used as a test function for smaller length intervals; more precisely, for any $(u,v) \in \mathbf{K}^+(\ell)$,

$$\begin{aligned}\frac{\mathbf{E}(u,v,\varepsilon,\ell)}{\ell} &\geq \frac{k}{\ell} \min_{0\leq j\leq k-1} \int_{j\ell/k}^{(j+1)\ell/k} \left\{ \frac{1}{2}u_x^2 + F(u) + \frac{1}{2}v_x^2 + \frac{\gamma}{2}v^2 \right\} dx \geq \frac{k}{\ell} E_* \left(\varepsilon, \frac{\ell}{k} \right) \\ &= \rho_* \left(\varepsilon, \frac{\ell}{k} \right).\end{aligned}$$

Note that (4.1) gives, for each fixed ε,

$$\inf_{\mu\geq 6} \rho^+(\varepsilon, \mu L_0 \varepsilon^{1/3}) \geq \inf_{\mu \geq 6} \rho_*(\varepsilon, \mu L_0 \varepsilon^{1/3}) = \min_{6\leq\mu\leq 12} \rho_*(\varepsilon, \mu L_0 \varepsilon^{1/3}). \tag{4.2}$$

In this manner, the estimate of $\rho^+(\varepsilon,\ell)$ for all $\ell \geq 6L_0\varepsilon^{1/3}$ can be obtained by estimating $\rho_*(\varepsilon, \mu L_0\varepsilon^{1/3})$ for all $\mu \in [6,12]$.

Finally, we remark that for every positive ε and ℓ, the energy functional $\mathbf{E}(\cdot,\cdot,\varepsilon,\ell)$ admits at least one global minimizer in each of the three function classes $\mathbf{K}_N^+(\ell)$, $\mathbf{K}^+(\ell)$, and $\mathbf{K}_N(\ell)$.

4.2. The Γ-convergence. The Γ-convergence was used by Modica [13] to the interfacial energy functional in the class of functions with prescribed average in arbitrary high space dimensions. This technique applies directly to the current situation if we impose

$$\int_0^\infty \sqrt{2F(s)}\, ds = \infty, \qquad \int_{-\infty}^0 \sqrt{2F(s)}\, ds = \infty. \tag{4.3}$$

Since this theory is elegant and the result is clean, we present it here under the above assumption. Though this is a very reasonable assumption, we still want to drop it. This will be done in the next subsection with a direct approach using neither assumption (4.3) nor the Γ-convergence. It is certainly not true that the Γ-convergence theory cannot be used without (4.3).

To set up the Γ-convergence, we use the scaling change

$$x = \ell\, y, \quad \ell = L\,\varepsilon^{1/3}, \quad \varepsilon = \epsilon\ell, \quad u(x) = U(y), \quad v(x) = \ell^2 V(y). \tag{4.4}$$

Then

$$\mathbf{E}(u,v,\varepsilon,\ell) \;=\; \varepsilon\Psi(U,V,\epsilon,L), \tag{4.5}$$

$$\Psi(U,V,\epsilon,L) := \int_0^1 \left\{ \frac{\epsilon}{2}U_y^2 + \frac{1}{\epsilon}F(U) \right\} dy + \frac{L^3}{2}\int_0^1 \left\{ V_y^2 + \gamma L^3 \epsilon V^2 \right\} dy, \tag{4.6}$$

where V satisfies $V_{yy} = m - U + \gamma L^3 \epsilon V$ in $(0,1)$.

Denote by (U^ϵ, V^ϵ) a global minimizer of $\Psi(\cdot,\cdot,\epsilon,L)$ in a space analogous to any one of $\mathbf{K}_N(\ell)$, $\mathbf{K}_N^+(\ell)$, or $\mathbf{K}^+(\ell)$. The energy upper bound (3.2) translates to

$$\Psi(U^\epsilon, V^\epsilon, \epsilon, L) < \sigma + \frac{L^3 m^2(1-m)^2}{6} \qquad \forall \epsilon > 0, L > 0.$$

For fixed $L > 0$, the right-hand side is independent of ϵ, so the $\epsilon \searrow 0$ limit can be studied to derive energy lower bounds. The key here is that the function $\tilde{U}^\varepsilon$ defined by

$$\tilde{U}^\epsilon(y) := \int_0^{U^\epsilon(y)} \sqrt{2F(s)} ds \qquad \forall y \in [0,1]$$

satisfies the estimate

$$\int_0^1 \left|\widetilde{U}^\epsilon_y(y)\right|dy = \int_0^1 \sqrt{2F(U^\epsilon)}\left|U^\epsilon_y\right|dy \leqslant \int_0^1 \left\{\frac{\epsilon}{2}U_y^{\epsilon 2} + \frac{1}{\epsilon}F(U^\varepsilon)\right\}dy < \sigma + \frac{L^3m^2(1-m)^2}{6}.$$

Since (U^ϵ, V^ϵ) are energy minimizers (in the space where at least $V^\epsilon_y \geq 0$), without using assumption (4.3) one can still show that there is at least one point at which U^ϵ takes a value in $[0,1]$. Indeed, if everywhere $U^\epsilon < 0$ or everywhere $U^\epsilon > 1$, then (i) if $L \geq 6L_0$, the interaction energy will be too large to allow (U^ϵ, V^ϵ) to be an energy minimizer (see the calculation later in this section), and (ii) if $L \in (0, 6L_0]$, one derives a similar contradiction after adding $\frac{(\bar{U}^\epsilon - m)^2}{2\epsilon\gamma}$ to the interfacial energy.

Thus, $\{\tilde{U}^\epsilon\}_{0<\epsilon<1}$ is a bounded family in the function space of bounded variations $(BV(0,1))$. Hence, along a subsequence, it converges pointwisely to a limit, say $\tilde{U}^0$, a BV function. In addition, along the sequence,

$$\liminf_{\epsilon\to 0}\int_0^1 \left\{\frac{\epsilon}{2}U_y^{\epsilon 2} + \frac{1}{\epsilon}F(U^\epsilon)\right\}dy \geq \sigma \int_0^1 |D\tilde{U}^0(y)|\,dy.$$

A pointwise convergence of $\tilde{U}^\epsilon$ implies a pointwise convergence of U^ϵ to a limit U^0. Furthermore, as $\int_0^1 F(U^\epsilon(y))dy < O(1)\epsilon$, by Fatou's lemma the limit U^0 satisfies

$$\int_0^1 F(U^0(y))dy = 0, \qquad \tilde{U}^0(y) = \int_0^{U^0(y)} \sqrt{2F(s)}ds \quad \forall\, y \in [0,1].$$

Under assumption (4.3), one infers from the uniform boundedness of $\tilde{U}^\varepsilon$ that $\{U^\epsilon\}$ is also uniformly bounded, and hence along the sequence converges to U^0 in L^2. Thus, almost everywhere, U^0 takes only two values, 0 and 1; that is, for some set $\Omega \subset [0,1]$, $U_0 = \chi_\Omega$ a.e.

The uniform bound on the energy provides a uniform bound on the L^2 norm of $\{V^\epsilon_y\}$ and the L^2 limit, $\lim_{\epsilon\to 0}(\gamma L^3 \epsilon V^\epsilon) = 0$. The uniform L^∞ bound on $\{U^\epsilon\}$ gives a uniform Lipschitz continuity of $\{V^\epsilon_y\}$. Hence, along a subsequence, $V^\epsilon_y \to W$ uniformly for some Lipschitz continuous function W. In conclusion,

$$U^0 = \chi_\Omega, \qquad W_y = m - \chi_\Omega,$$

$$\liminf_{\epsilon\to 0}\Psi(U^\epsilon, V^\epsilon, \epsilon, L) \geqslant \Psi^0(\bar{\Omega}, W, L) := \sigma\int_0^1 |D\chi_\Omega|dy + \frac{L^3}{2}\int_0^1 W^2(y)dy,$$

where $\int_0^1 |D\chi_\Omega|dy$ is the number of boundary points of Ω (after it is defined in a unique way). With a related reverse inequality established (which we omit here), the functional Ψ^0 is then called the Γ-limit of the original functional sequence $\{\Psi(\cdot,\cdot,\epsilon,\cdot)\}_{0<\epsilon<1}$; see [13] for more details.

For a given L, minimizers of Ψ^0 can be calculated explicitly.

1. First, we consider the functional in the class that corresponds to the original $\mathbf{K}^+$. This is the set

$$\mathbf{J}^+ := \{(\Omega, W) \in BV(0,1) \times H^1(0,1) \mid W_y = m - \chi_\Omega, W \geq 0\}.$$

Write $\Omega = \cup_{i=1}^{k_1}[c_i, d_i]$ and $\Omega^c = \cup_{i=1}^{k_0}(a_i, b_i)$, where $d_i < c_{i+1}$ and $b_i < a_{i+1}$. Note that

$$|k_0 - k_1| \le 1, \quad \int_0^1 |D\chi_\Omega| dy = k_0 + k_1 - 1.$$

On $[a_i, b_i]$, we integrate $W_y = m$ over $[a_i, y]$ to obtain $W(y) = W(a_i) + m(y - a_i) \geqslant m(y - a_i)$, so that $\int_{a_i}^{b_i} W^2 dy \ge \frac{1}{3} m^2 (b_i - a_i)^2$.

On $[c_i, d_i]$, we integrate $W_y = m - 1$ over $[y, d_i]$ to obtain $W(y) = W(d_i) + (1 - m)(d_i - y) \geqslant (1 - m)(d_i - y)$, so that $\int_{c_i}^{d_i} W^2 dy \ge \frac{1}{3}(1 - m)^2 (d_i - c_i)^3$. Thus,

$$\begin{aligned}
\int_0^1 W^2 dy &\geqslant \frac{m^2}{3} \sum_{i=1}^{k_0} (b_i - a_i)^3 + \frac{(1-m)^2}{3} \sum_{i=1}^{k_1} (d_i - c_i)^3 \\
&\geqslant \frac{m^2}{3k_0^2} \left(\sum_{i=1}^{k_0} (b_i - a_i) \right)^3 + \frac{(1-m)^2}{3k_1^2} \left(\sum_{i=1}^{k_1} (d_i - c_i) \right)^3 \\
&= \frac{m^2 (1 - |\Omega|)^3}{3k_0^2} + \frac{(1-m)^2 |\Omega|^3}{3k_1^2} \\
&\geqslant \min_{0 \le t \le 1} \left\{ \frac{m^2 t^3}{3k_0^2} + \frac{(1-m)^2 (1-t)^3}{3k_1^2} \right\},
\end{aligned}$$

where in the second inequality we have used $\sum e_i \le (\sum e_i^3)^{1/3} (\sum 1)^{2/3}$. The minimum is attained at $t = (1 - m)k_0/[mk_1 + (1 - m)k_0]$ so that

$$\int_0^1 W^2(y)\, dy \ge \frac{m^2 (1-m)^2}{3[mk_1 + (1-m)k_0]^2}. \tag{4.7}$$

Note that this estimate holds even if $k_0 = 0$ (e.g., $\Omega = [0, 1]$) or $k_1 = 0$ (e.g., $\Omega = \emptyset$). Thus,

$$\Psi^0(\Omega, W, L) \ge \sigma(k_0 + k_1 - 1) + \frac{L^3 m^2 (1-m)^2}{6[mk_1 + (1-m)k_0]^2}.$$

We remark that here the equality can be attained for any given k_0, k_1 satisfying $|k_0 - k_1| \leqslant 1 \leqslant k_1 + k_0$. Indeed, set $t = (1 - m)k_0/[mk_1 + (1 - m)k_0]$ and take (1) $a_1 = 0$ if $k_0 \ge k_1$ and $c_1 = 0$ if $k_0 < k_1$; (2) $b_i = a_i + l_0$, $l_0 := t/k_0$ for all i; (3) $d_i = c_i + l_1$, $l_1 := (1 - t)/k_1$ for all i; (4) $W = m(y - a_i)$ for all $y \in [a_i, b_i]$; and (5) $W(y) = (1 - m)(d_i - y)$ for all $y \in [c_i, d_i]$. Then W is a continuous function since $ml_0 = (1 - m)l_1$. For this particular choice, $(\Omega, W) \in \mathbf{J}^+$ and the above inequality is indeed an equality. Thus, for any $L > 0$,

$$\min_{\mathbf{J}^+} \Psi^0 = \min_{|p-q| \le 1 \le p+q} \left(\sigma(p + q - 1) + \frac{L^3 m^2 (1-m)^2}{6[mp + (1-m)q]^2} \right).$$

2. Next we consider the space that corresponds to the original space $\mathbf{K}_N^+$. This is the space

$$\mathbf{J}_N^+ := \{(\Omega, W) \in \mathbf{J}^+ \mid W(0) = W(1) = 0\}.$$

Suppose $(\Omega, W) \in \mathbf{J}_N^+$. Use the same notation as in the previous case. Note that $W(0) = 0$, $W \ge 0$, and $W_y = m - \chi_\Omega$ imply that $\Omega \subset [b, 1]$ for some $b > 0$. Similarly,

$W(1) = 0$ implies that $[c, 1] \subset \Omega$ for some $c < 1$. Hence, we must have $k_0 = k_1 \geq 1$. Thus, for every $L > 0$,

$$\min_{\mathbf{J}_N^+} \Psi^0 = \min_{k \geq 1} \left(\sigma(2k-1) + \frac{L^3 m^2 (1-m)^2}{6k^2} \right).$$

Remark 4.1. With a little extra work, i.e., constructing test functions for upper bounds, one can indeed prove the following: For every $\mu > 0$,

$$\lim_{\varepsilon \to 0} \frac{L_0}{\sigma \varepsilon^{2/3}} \rho_*(\varepsilon, \mu L_0 \varepsilon^{1/3}) = \min_{|p-q| \leq 1 \leq p+q} \left(\frac{p+q-1}{\mu} + \frac{\mu^2}{2[pm + (1-m)q]^2} \right),$$
$$\lim_{\varepsilon \to 0} \frac{L_0}{\sigma \varepsilon^{2/3}} \rho^+(\varepsilon, \mu L_0 \varepsilon^{1/3}) = \min_{k \geq 1} \left(\frac{2k-1}{\mu} + \frac{\mu^2}{2k^2} \right),$$
$$\lim_{\varepsilon \to 0} \frac{L_0}{\sigma \varepsilon^{2/3}} \rho(\varepsilon, \mu L_0 \varepsilon^{1/3}) = \min_{n \geq 1} \left(\frac{n}{\mu} + \frac{\mu^2}{2n^2} \right).$$

We leave the details to interested readers.

When condition (4.3) is dropped, we face technical difficulties on the convergence of $\{U^\epsilon\}$ and the equation $V^\epsilon_{yy} = m - U^\epsilon + \gamma L^3 \epsilon V^\epsilon$. Though these difficulties could be overcome by establishing a uniform L^∞ bound of $\{U^\epsilon\}$, we prefer to present another approach.

4.3. Assumption (4.3) dropped. Let $\delta \in (0, 1/2)$ be any small positive number. Define

$$\sigma^\delta := \int_\delta^{1-\delta} \sqrt{2F(s)} ds, \qquad A(\delta) = \sup_{0<s<1-\delta} \frac{s}{\sqrt{F(s)}} + \sup_{\delta<s<1} \frac{1-s}{\sqrt{F(s)}}.$$

Since $F''(0) > 0 = F(0) = F'(0)$ and $F''(1) > 0 = F(1) = F'(1)$, $A(\delta)$ is bounded.

LEMMA 4.1. *Assume* (1.3) *only. Suppose $U \in H^1(0,1)$, V satisfies $V_{yy} = m - U + \gamma L^3 \epsilon V$, and $V_y \geq 0$ on $[0,1]$. Then, for every $\delta \in (0, 1/2)$ there exist non-negative integers k_0, k_1, and k such that $k_0 + k_1 \leqslant k+1$, $|k_0 - k_1| \leqslant 1$, and*

$$(k+1)\sigma^\delta \geqslant \int_0^1 \left\{ \frac{\epsilon}{2} U_y^2 + \frac{1}{\epsilon} F(U) \right\} dy > k\sigma^\delta, \tag{4.8}$$

$$\int_0^1 (V_y^2 + \gamma \epsilon L^3 V^2) dy + \frac{1}{4} \gamma L^3 \epsilon + A\sqrt{(k+1)\sigma\epsilon} \geqslant \frac{m^2(1-m)^2}{3[mk_1 + (1-m)k_0]^2}. \tag{4.9}$$

Proof. The idea is to approximate the equation for V by $V_{yy} \geqslant m -$ error in one set Ω_0, and $V_{yy} \leqslant m - 1 +$ error in the other set Ω_1, where the error terms are either positively small or negative. The numbers k_0 and k_1 of disjoint pieces of Ω_0 and Ω_1 are taken as small as possible so that after integration V_y will not be underestimated too much.

We are only interested in the case $U \not\equiv 0$ and $U \not\equiv 1$. Then (4.8) defines a unique nonnegative integer k. Set $y_0 = 0$, $y_{k+1} = 1$. When $k > 0$, define $y_1, \ldots, y_k$ by

$$j\sigma^\delta = \int_0^{y_j} \left\{ \frac{\epsilon}{2} U_y^2 + \frac{1}{\epsilon} F(U) \right\} dy \qquad \forall\, j = 1, \ldots, k.$$

Then for each $j = 0, \dots, k$, either $U \geq \delta$ on $[y_j, y_{j+1}]$ or $U \leq 1-\delta$ on $[y_j, y_{j+1}]$, since

$$\int_\delta^{1-\delta} \sqrt{2F(s)}ds = \sigma^\delta \geqslant \int_{y_j}^{y_{j+1}} |U_y|\sqrt{2F(U)}dy.$$

Now define

$$\Omega_0 = \Big\{[y_j, y_{j+1}] \mid U \leq 1-\delta \text{ on } \ [y_j, y_{j+1}]\Big\} =: \cup_{i=1}^{k_0}[a_i, b_i], \quad \Big(b_i < a_{i+1}\Big),$$
$$\Omega_1 = [0,1] \setminus \Big(\cup_{i=1}^{k_0}(a_i, b_i)\Big) =: \cup_{i=1}^{k_1}[c_i, d_i], \quad \Big(d_i < c_{i+1}\Big).$$

Note that each maximal connected component of Ω_0 or Ω_1 is one interval or a union of several consecutive intervals of $[y_j, y_{j+1}]$ so that $k_0 + k_1 \leq k+1$. Also, maximal connected components of Ω_0 and Ω_1 interlace each other so that $|k_0 - k_1| \leqslant 1$. Furthermore, $U \leq 1-\delta$ on Ω_0 and $U \geqslant \delta$ on Ω_1. Hence,

$$U \leq A(\delta)\sqrt{F(U)} \ \text{ on } \ \Omega_0, \qquad 1 - U \leq A(\delta)\sqrt{F(U)} \ \text{ on } \ \Omega_1.$$

Pick any maximal connected component $[a_i, b_i]$ of Ω_0. Integrating $V_{yy} = m - U + \gamma L^3 \epsilon V$ over $[a_i, y]$, $y \in [a_i, b_i]$ gives

$$V_y(y) = m(y - a_i) + \Bigg\{V_y(a_i) + \int_{a_i}^{y}(\gamma L^3\epsilon V - U)\Bigg\}. \tag{4.10}$$

First squaring both sides and then using $V_y(a_i) \geqslant 0$ and $U \leqslant A(\delta)\sqrt{F(U)}$ we obtain

$$\begin{aligned} V_y^2(y) &\geqslant m^2(y-a_i)^2 + 2m(y-a_i)\left(V_y(a_i) + \int_{a_i}^{y}(\gamma L^3\epsilon V - U)\right) \\ &\geqslant m^2(y-a_i)^2 - 2m(y-a_i)\int_{a_i}^{b_i}\Big\{\gamma L^3\epsilon|V| + A(\delta)\sqrt{F(U)}\Big\}, \end{aligned}$$
$$\int_{a_i}^{b_i} V_y^2 dy \geqslant \frac{m^2}{3}(b_i - a_i)^3 - m\int_{a_i}^{b_i}\Big\{\gamma L^3\epsilon|V| + A\sqrt{F(U)}\Big\}dy.$$

Similarly, for any interval $[c_i, d_i]$ of Ω_1, we integrate the equation for V over $[y, d_i]$ for $y \in [c_i, d_i]$ to obtain $V_y(y) = (1-m)(d_i - y) + V_y(d_i) + \int_y^{d_i}(U - 1 - \gamma L^3\epsilon V)$ and

$$\int_{c_i}^{d_i} V_y^2 dy \geqslant \frac{(1-m)^2}{3}(d_i - c_i)^3 - (1-m)\int_{c_i}^{d_i}\Big\{\gamma L^3\epsilon|V| + A\sqrt{F(U)}\Big\}.$$

Now adding them up we obtain

$$\int_0^1 V_y^2 dy \geq \frac{m^2}{3}\sum_{i=1}^{k_0}(b_i - a_i)^3 + \frac{(1-m)^2}{3}\sum_{i=1}^{k_1}(d_i - c_i)^3 - \int_0^1\Big\{\gamma L^3\epsilon|V| + A\sqrt{F(U)}\Big\}dy.$$

The first two sums can be estimated from below by the right-hand side of (4.7), whereas the integral can be estimated by

$$\int_0^1\Big\{\gamma L^3\epsilon|V| + A\sqrt{F(U)}\Big\}dy \leqslant \tfrac{1}{4}\epsilon\gamma L^3 + \int_0^1 \gamma\epsilon L^3 V^2 + \sqrt{\epsilon}A\sqrt{(k+1)\sigma},$$

since $\int_0^1 \sqrt{F}dy \leq \sqrt{\int_0^1 F dy} \leq \sqrt{(k+1)\sigma\epsilon}$. This completes the proof. $\square$

4.4. Energy lower bounds. Now we are ready to calculate explicitly energy lower bounds. Depending on the size of $\mu := L/L_0$ we consider three cases.

Case I. $\mu = L/L_0 = \ell/[L_0\varepsilon^{1/3}] \in [6, 12]$.

Let (u, v) be a minimizer of $\mathbf{E}(\cdot, \cdot, \varepsilon, \ell)$ in $\mathbf{K}^+(\ell)$ so that $\mathbf{E}(u, v, \varepsilon, \ell) = \rho_*(\varepsilon, \ell)$.

Define (U, V) as in (4.4). Applying Lemma 4.1, we obtain (4.8) and (4.9) for some integers k, k_1, k_0 satisfying $k_0 + k_1 \leqslant k + 1$, $|k_0 - k_1| \leqslant 1$. Translating to (u, v) (recalling $L = \mu L_0$, $m^2(1-m)^2 = 3\sigma/L_0^3$, $\epsilon = \varepsilon^{2/3}/(\mu L_0)$) we obtain

$$\frac{L_0\mathbf{E}(u,v,\varepsilon,\ell)}{\sigma\ell\varepsilon^{2/3}} + \frac{\mu^2 L_0^3}{2\sigma}\left\{A(\delta)\sqrt{\tfrac{(k+1)\sigma}{\mu L_0}}\varepsilon^{1/3} + \frac{\gamma\mu^2 L_0^2\varepsilon^{2/3}}{4}\right\} + \frac{k(\sigma-\sigma^\delta)}{\sigma\mu}$$

$$\geqslant \frac{k}{\mu} + \frac{\mu^2}{2[mk_1 + (1-m)k_0]^2}\ . \tag{4.11}$$

There are two cases to consider: (i) $k \le 5$, (ii) $k \ge 6$.

(i) $k \le 5$. Then $k_0 \le 3$ and $k_1 \le 3$, so that the right-hand side of (4.11) is $\ge \mu^2/18 \ge 2$.

(ii) $k \ge 6$. Then using $\max\{k_0, k_1\} \le \frac{k}{2}+1$ and $a + \frac{1}{2}b \ge \frac{3}{2}(a^2b)^{1/3}$, the right-hand side of (4.11) is no smaller than $\frac{3}{2}l(\frac{k^2}{(k+1/2)^2})^{1/3} \geqslant (\frac{243}{32})^{1/3}$.

Thus, in both cases the right-hand side of (4.11) is $\ge (\frac{243}{32})^{1/3} > \sqrt[3]{7}$. Taking δ small we see that for all sufficiently small positive ε,

$$\frac{L_0}{\sigma\varepsilon^{2/3}}\rho_*(\varepsilon, \mu L_0\varepsilon^{1/3}) > \sqrt[3]{7} \quad \forall \mu \in [6, 12]. \tag{4.12}$$

Case II. $\mu = L/L_0 = \ell/(L_0\varepsilon^{1/3}) \in [\sqrt[3]{4}, 6]$.

Let (u, v) be a minimizer of $\mathbf{E}$ in $\mathbf{K}_N^+(\ell)$. We claim that in (4.11) we can have $k_1 = k_0$ (and $k \ge k_0 + k_1 - 1 = 2k_0 - 1$), at a cost of at most an $O(\epsilon)$ term added.

Suppose $k_0 > k_1$. Then $k_0 = k_1 + 1$. This implies that $b_{k_0} = 1$. Since $V'(1) = 0$, (4.10) with $y = 1$ gives

$$\begin{aligned} 0 &= m(1 - a_{k_0}) + V_y(a_{k_0}) + \int_{a_{k_0}}^1 (\gamma L^3\epsilon V - U) \\ &\geqslant m(1 - a_{k_0}) - \int_{a_{k_0}}^1 \Big\{\gamma L^3\epsilon|V| + A(\delta)\sqrt{F(U)}\Big\}. \end{aligned}$$

After using the Hölder inequality

$$\int_{a_{k_0}}^1 (\gamma L^3\epsilon|V| + A\sqrt{F}) \le \sqrt{(1 - a_{k_0})\int_{a_{k_0}}^1 \{\gamma L^3\epsilon|V| + A\sqrt{F}\}^2},$$

we then obtain

$$m^2(1 - a_{k_0}) \le \int_{a_{k_0}}^1 \Big\{\gamma L^3\epsilon|V| + A\sqrt{F}\Big\}^2 \le 2\epsilon\int_0^1 \Big\{\gamma^2 L^6\epsilon V^2 + \frac{A^2}{\epsilon}F(U)\Big\}\,dy = O(1)\epsilon.$$

Thus, the component $[a_{k_0}, b_{k_0}] = [a_{k_0}, 1]$ has length of size $O(1)\epsilon$. In the estimation in deriving the right-hand side of (4.7), we redo the estimate for the first term by

$$\begin{aligned} \sum_{i=1}^{k_0}(b_i - a_i)^3 &\ge \sum_{i=1}^{k_0-1}(b_i - a_i)^3 \ge \frac{1}{(k_0-1)^2}\left(\sum_{i=1}^{k_0-1}(b_i - a_i)\right)^3 \\ &= \frac{m^2(|\Omega_0| - (1 - a_{k_0}))^3}{3(k_0-1)^2} = \frac{m^2|\Omega_0|^3}{3(k_0-1)^2} + O(1)\epsilon. \end{aligned}$$

Hence, k_0 in (4.11) can be replaced by $k_0 - 1$, with the cost of adding a term of order $O(1)\epsilon = O(1)\varepsilon^{2/3}$. The case $k_1 > k_0$ can be similarly treated.

We remark that we are working on a finite range of $\mu \in [1/2, 12]$, so all k_0 and k_1 cannot be too large, e.g., ≥ 12, since otherwise interfacial energy $k\sigma^\delta\epsilon$ alone will be large enough to eliminate the possibility. Hence, all $O(1)$ may depend on k, but it is irrelevant.

Once we know $k_0 = k_1$, (4.11) can be expressed simply as

$$\frac{L_0}{\sigma\varepsilon^{2/3}}\rho^+(\varepsilon, \mu L_0\varepsilon^{1/3}) + O(1)\varepsilon^{1/3} + O(1)(\sigma - \sigma^\delta) \geq \frac{2k_0 - 1}{\mu} + \frac{\mu^2}{2k_0^2}.$$

If $k_0 = 1$, then the right-hand side is $\geq \frac{1}{\mu} + \frac{\mu^2}{2} \geq \sqrt[3]{\frac{27}{4}}$ since $\mu \geq \sqrt[3]{4}$.

If $k_0 \geq 2$, then by $a + \frac{1}{2}b \geq \frac{3}{2}(a^2b)^{1/3}$, the right-hand side is $\geqslant \frac{3}{2}(1 - 1/k_0)^{2/3} \geq \sqrt[3]{\frac{243}{32}} > \sqrt[3]{7}$.

In any case, the right-hand side is $\geq \sqrt[3]{27/4} > \sqrt[3]{6}$. Thus, for all sufficiently small positive ε,

$$\frac{L_0}{\sigma\varepsilon^{2/3}}\rho^+(\varepsilon, \mu L_0\varepsilon^{1/3}) > \sqrt[3]{6} \qquad \forall\, \mu \in [\sqrt[3]{4}, 6]. \tag{4.13}$$

Proof of (1.8). The estimate for $\mu \leq 1/4$ follows from Lemma 2.1 since we have

$$\frac{L_0}{\sigma\varepsilon^{2/3}}\rho^+(\varepsilon, \ell) \geq \frac{L_0}{\sigma\varepsilon^{2/3}}\rho(\varepsilon, \ell) \geq \min\left\{\frac{c_1 L_0}{\sigma\varepsilon^{2/3}}, \frac{1}{2\mu}\right\},$$

where c_1 is a constant depending only on m and F. The estimate for $\mu \geq \sqrt[3]{4}$ follows from (4.2), (4.12), and (4.13).

Case III. $\mu = L/L_0 = \ell/(L_0\varepsilon^{1/3}) \in [\frac{1}{16}, \sqrt[3]{5}]$.

Let (u, v) be a minimizer of $\mathbf{E}$ in $\mathbf{K}_N^+(\ell)$. The same consideration as above leads to the conclusion that (4.11) holds with $k_0 = k_1$. From the energy upper bound, we conclude that we must have $k = k_0 = k_1 = 1$, and the total energy is $\mathbf{E}(u, v, \varepsilon, \ell) \geq [\sigma\varepsilon + \ell^3 m^2(1 - m)^2/6](1 - o(1))$. As the energy upper bound differs from the energy lower bound by a tiny fraction, the separate lower bound estimates on the interfacial and potential energy in Lemma 4.1 then give us

$$\int_0^\ell \left\{\frac{\varepsilon}{2}u_x^2 + \frac{1}{\varepsilon}F(u) + \frac{1}{2\gamma\varepsilon}(\bar{u} - m)^2\right\} dx = \sigma + o(1), \tag{4.14}$$

$$\frac{1}{\ell^3}\int_0^\ell \left\{v_x^2 + \gamma(v - \bar{v})^2\right\} dx = \frac{1}{3}m^2(1 - m)^2 + o(1). \tag{4.15}$$

Furthermore, after checking the lower bound estimation that leads to the right-hand side of (4.7), we conclude that we must have $b_1 = c_1 = 1 - m + o(1)$. Namely,

$$u = \begin{cases} o(1) & \text{in } [0, (1 - m - o(1))\ell], \\ 1 - o(1) & \text{in } [(1 - m + o(1))\ell, \ell], \end{cases} \tag{4.16}$$

$$v_x = \begin{cases} [m + o(1)]x & \text{in } [0, (1 - m)\ell], \\ -(1 - m + o(1))(x - \ell) & \text{in } [(1 - m)\ell, \ell]. \end{cases} \tag{4.17}$$

5. Certain quantitative estimates for minimizers. To complete the proof of Theorem 1, it remains to show its parts 2 and 3. The rest of this paper is devoted to this task. Hence, in what follows we always assume that

$$0 < \varepsilon \ll 1, \qquad \tfrac{1}{16}\, L_0 \varepsilon^{1/3} \le \ell \le \sqrt[3]{5}\, L_0 \varepsilon^{1/3}. \tag{5.1}$$

Our plan is as follows. First, in this section we provide certain quantitative estimates for energy minimizers in $\mathbf{K}_N^+(\ell)$. In particular, we show that any two minimizers are at most $O(\ell^4)$ apart. In the next section, we study the principal eigenvalue of the linearized operator around any minimizer. We show that the eigenvalue is bigger than $c\ell^2$ for some c independent of ε. As a consequence, energy minimizers are unique. Once we have the uniqueness, we can differentiate the solution with respect to ℓ to calculate the derivatives of E^+ and ρ^+ with respect to ℓ, and show that E^+ and ρ^+ are convex in ℓ in the range specified in (5.1).

5.1. The Euler–Lagrange equation. In what follows, differentiation with respect to x will be denoted by $'$. Also, (u, v) is always a global minimizer of $\mathbf{E}(\cdot,\cdot,\varepsilon,\ell)$ in $\mathbf{K}_N^+(\ell)$ with ℓ in the range specified in (5.1). Since $\mathbf{K}_N^+(\ell)$ is not an open set, that (u, v) satisfies the Euler–Lagrange equation is not obvious.

LEMMA 5.1. *Let (u, v) be a minimizer of $\mathbf{E}$ in $\mathbf{K}_N^+(\ell)$. Then $v' > 0$ in $(0,\ell)$, $v''(0) > 0 > v''(\ell)$, and (u, v) satisfies the Euler–Lagrange system*

$$\begin{cases} \varepsilon^2 u'' = f(u) + v \quad \text{in} \quad (0,\ell),\\ v'' = m + \gamma v - u \quad \text{in} \quad (0,\ell),\\ u'(0) = v'(0) = u'(\ell) = v'(\ell) = 0. \end{cases}$$

Furthermore, for all $\phi \in H^1(0,\ell)$ and $\psi = \mathbf{K}_N^\ell \phi$,

$$\int_0^\ell \left\{ \varepsilon^2 \phi'^2 + f_u(u)\phi^2 + \frac{1}{\gamma}\bar{\phi}^2 + \psi'^2 + \gamma(\psi - \bar{\psi})^2 \right\} dx \ge 0. \tag{5.2}$$

Proof. We shall start from the elementary estimates (4.14)–(4.17).

(a) Note that γv is small since (i) (4.17) gives $v' = O(\ell)$, (ii) $\gamma \bar{v} = \bar{u} - m$ (obtained by integrating $v'' - \gamma v = m - u$), and (iii) the energy upper bound and (1.5) give $m - \bar{u} = O(\sqrt{\varepsilon/\ell})$.

(b) As long as we know that γv is small, we immediately conclude that $v''(0) = m + \gamma v - u = m + o(1)$ and $v''(\ell) = m + \gamma v - u = m - 1 + o(1)$. Together with (4.17), we have $v''(0) > 0 > v''(\ell)$ and $v' > 0$ in $(0,\ell)$. This will allow us to do the standard calculus of variation.

(c) For every $\phi \in H^1(0,\ell)$, consider the test function

$$\tilde{u} = u + t\phi, \quad \tilde{v} = v + t\psi, \quad \psi := \mathbf{K}_N^\ell \phi = \tfrac{1}{\gamma}\bar{\phi} + \mathbf{K}_N^\ell(\phi - \bar{\phi}).$$

Here $\mathbf{K}_N^\ell$ is the operator defined in (3.1). This operator is uniquely defined for $\gamma > 0$. When $\gamma = 0$, it is defined only for functions of zero mean, and ψ is unique up to an additive constant. Hence, here we use $\bar{\phi}/\gamma$ to denote such an additive constant when $\gamma = 0$.

Note that $\|\psi\|_{C^2}$ is bounded and $\psi'(0) = \psi'(\ell) = 0$. Since $v_x > 0$ in $(0, \ell)$ and $v''(0) > 0 > v''(\ell)$, it follows that $(\tilde{u}, \tilde{v}) \in \mathbf{K}_N^+(\ell)$ if $|t|$ is small enough. Hence,

$$0 = \frac{d}{dt}\mathbf{E}(u + t\phi, v + t\psi, \varepsilon, \ell)\Big|_{t=0} = \int_0^\ell \Big\{\varepsilon^2 u'\phi' + f(u)\phi + v'\psi' + \gamma v\psi\Big\}$$
$$= \varepsilon^2 u'\phi\Big|_0^\ell + \int_0^\ell \Big\{-\varepsilon^2 u'' + f(u) + v\Big\}\phi\, dx$$

after integrating by parts and using equation $-\psi'' + \gamma\psi = \phi$. This integral identity implies that u is a weak solution to $\varepsilon^2 u'' = f(u) + v$ in $(0, \ell)$ with the natural boundary condition $u'(0) = u'(\ell) = 0$. Since (u, v) is bounded, (u, v) is a classical solution.

(d) Finally, since (u, v) is a minimizer, $\frac{d^2}{dt^2}\mathbf{E}(u + t\phi, v + t\psi, \varepsilon, \ell)|_{t=0} \geq 0$, which is equivalent to (5.2). □

Once we have the Euler–Lagrange equation, we can estimate the L^∞ bound of v. First of all, from the equation for u we derive that

$$\bar{v} = -\frac{1}{\ell}\int_0^\ell f(u)dx = \frac{1}{\ell}\int_0^\ell O(\sqrt{F(u)})dx$$
$$= O(1)\ell^{-1/2}\left(\int_0^\ell F(u)dx\right)^{1/2} = O(1)\ell^{-1/2}\varepsilon^{1/2} = O(\ell).$$

Here we used the fact that $u \in [-o(1), 1 + o(1)]$ and that

$$\sup_{s\in[-1,2]\setminus\{0,1\}} \frac{|f(s)|}{\sqrt{F(s)}} < \infty.$$

Since $\|v'\|_{L^2}^2 = O(\varepsilon)$, we conclude that $\|v\|_{L^\infty} = O(\ell)$. Improvements will be made later.

5.2. The profile of **.** Note that

$$\frac{d}{dx}\Big(\tfrac{1}{2}\varepsilon^2 u'^2 - F(u) - \tfrac{\gamma}{2}v^2 + vv'' - \tfrac{1}{2}v'^2\Big) = 0. \tag{5.3}$$

After integration we see that $\frac{1}{2}\varepsilon^2 u'^2 = F(u) + o(1)$. Hence, for any fixed small positive δ and ε sufficiently small, $|u'| > 0$ whenever $\delta < u < 1 - \delta$. Since u has only one layer, it follows that $u' > 0$ whenever $\delta < u < 1 - \delta$. In summary, we have the following basic estimate.

LEMMA 5.2. *In* $[0, \ell]$, $v = o(1)$, $\varepsilon u' = \sqrt{2F(u)} + o(1)$, *and there exists a unique* $z = (1 - m + o(1))\ell$ *such that*

$$u(z) = Q(0), \qquad u(x) = Q\Big(\frac{x - z}{\varepsilon}\Big) + o(1) \quad \forall x \in [0, \ell].$$

5.3. A relation between the tails of **and** **.** Away from the interface $x = z$, we expect $f(u) + v = \varepsilon^2 u'' = O(\varepsilon^2)$. Thus, formally $u \sim f^{-1}(-v)$, where f^{-1} represents the inverse of f near 0 or 1. To make this statement more precise, we introduce below functions h_0 and h_1, which will later be shown to be exponentially $O(e^{-c|x-z|/\varepsilon})$ close to $u - Q$.

LEMMA 5.3. *For each* $i = 0$ *and* $i = 1$ *there exists a unique solution* h_i *to*

$$\begin{cases} \varepsilon^2 h_i'' = f(h_i) + v & \text{in } (0, \ell), \\ h_i'(0) = h_i'(\ell) = 0, \ h_i = i + O(v). \end{cases} \tag{5.4}$$

In addition, h_i satisfies the following estimates:

$$f(h_i) + v = O(\varepsilon^2), \quad h_i' = O(\ell), \quad h_i'' = O(1) \qquad \text{on } [0, \ell].$$

Proof. Denote by f_i^{-1} the inverse of f near $i = 0$ or $i = 1$. As $v = o(1)$, $v'(0) = v'(\ell) = 0$, and $v'' = O(1)$, it is easy to show that for some large enough constant M, $f_i^{-1}(-v) \pm M\varepsilon^2$ is a super- or sub-solution. As $f'(0) > 0$ and $f'(1) > 0$, it then follows from a standard elliptic theory that (5.4) admits a unique solution and the solution satisfies $|h_i - f_i^{-1}(-v)| \le M\varepsilon^2$ on $[0, \ell]$. Consequently, $\varepsilon^2 h_i'' = f(h_i) + v = O(\varepsilon^2)$ so that $h_i'' = O(1)$, which implies, as $h_i'(0) = 0$, that $h_i' = O(\ell)$. ◻

LEMMA 5.4. *Let θ, h be defined as*

$$\theta(x) = u(x) - \begin{cases} h_0(x), & x \in (0, z), \\ h_1(x) - 1, & x \in (z, \ell], \end{cases} \qquad h(x) = \begin{cases} h_0(x), & x \in (0, z), \\ h_1(x), & x \in (z, \ell]. \end{cases}$$

Then

$$\text{(5.5)} \qquad \begin{cases} \varepsilon^2 \theta'' - f(\theta) = [O(v) + O(\varepsilon^2)]\theta(1 - \theta) & \text{in } [0, \ell] \setminus \{z\}, \\ \theta'(0) = 0 = \theta'(\ell), \\ \theta(z\pm) = Q(0) + O(v(z)) + O(\varepsilon^2). \end{cases}$$

Proof. The boundary conditions of θ at 0 and ℓ follow from that of u and h_i. The condition at $z\pm$ follows from the fact that $u(z) = Q(0)$ and $h_i = f_i^{-1}(-v) + O(\varepsilon^2) = i + O(v) + O(\varepsilon^2)$.

For $x \in [0, z-]$, $\theta = u - h_0$ and thus $\varepsilon^2\theta'' = \varepsilon^2(u'' - h_0'') = f(\theta + h_0) - f(h_0)$. Note that

$$f(\theta + h_0) - f(h_0) - f(\theta) = \int_0^\theta \int_0^{h_0} f_{uu}(\xi + \eta)\, d\eta\, d\xi = O(1) h_0 \theta.$$

The assertion for $x \in [0, z-]$ thus follows. The case $x \in [z+, \ell]$ is similar. ◻

Note that (5.5) shows that $\theta(x)$ is very close to $Q = Q((x - z)/\varepsilon)$. The following lemma shows that θ is similar to Q.

LEMMA 5.5. *For all $x \in (0, z-] \cup [z+, \ell)$, we have $0 < \theta(x) < 1$, $\theta'(x) > 0$, and*

$$\varepsilon\theta' = [1 + O(v(z)) + O(\varepsilon\ell \ln(\theta - \theta^2))]\sqrt{2(F(\theta) - F(\theta_i))} \quad \forall x \in [0, z-] \cup [z+, \ell],$$
$$\theta(1 - \theta) = O(1) e^{-c|x - z|/\varepsilon} \quad \forall x \in [0, \ell],$$

where $\theta_i = \theta(0)$ if $x < z$, $\theta_i = \theta(\ell)$ if $x > z$, and $c = \frac{1}{2}\min\{\sqrt{f'(0)}, \sqrt{f'(1)}\}$.

Note that the assertion implies that u is exponentially $O(e^{-c|x-z|/\varepsilon})$ close to h, which has expansion $h = f^{-1}(-v) + O(\varepsilon^2)$.

Proof. Since $f'(0) > 0$ and $f'(1) > 0$, it is easy to see from equation $\varepsilon^2\theta'' = f(\theta) + o(1)\theta(1 - \theta)$ that $0 < \theta < 1$ and $\theta' > 0$ in $(0, z-] \cup [z+, \ell)$.

For $x \in [0, z-]$, we have

$$\frac{1}{2}\varepsilon^2\theta'^2 = \int_0^x \Big\{ f(\theta) + [O(v) + O(\varepsilon^2)]\theta \Big\} \theta' = [1 + o(1)][F(\theta) - F(\theta(0))]$$

by the mean value theorem $\int_a^b A(x)B(x)dx = A(\xi)\int_a^b B(x)dx$ (for $B \ge 0$) and the fact that there exist positive constants c_1 and c_2 such that $c_1^2(s^2 - \hat{s}^2) \le 2(F(s) -$

$F(\hat{s})) \le c_2^2(s^2 - \hat{s}^2)$ for all $\hat{s} \in (0, \delta]$ and $s \in (\hat{s}, Q(0) + \delta]$ (for fixed small positive δ). Consequently,

$$c_1\sqrt{\theta^2 - \theta(0)^2} \le \varepsilon\theta' \le c_2\sqrt{\theta^2 - \theta(0)^2}.$$

Dividing both sides by $\sqrt{\theta^2 - \theta(0)^2}$ and integrating over $[x, z]$ then gives

$$c_1(z - x) \le \varepsilon \ln \frac{\theta(z) + \sqrt{\theta^2(z) - \theta^2(0)}}{\theta(x) + \sqrt{\theta^2(x) - \theta^2(0)}} \le c_2(z - x).$$

This gives us

$$z - x = O(\varepsilon \ln \theta), \qquad \theta = O(e^{-c_1(z-x)/\varepsilon}).$$

Using $v(x) = v(z) + \int_z^x v' = v(z) + O(\ell)(x - z) = v(z) + O(\varepsilon\ell \ln \theta)$, we then obtain

$$\begin{aligned}\tfrac{1}{2}\varepsilon^2\theta'^2 &= \int_{\theta(0)}^{\theta(x)} \Big\{ f(\theta) + O(v(z))\theta + O(\varepsilon\ell\theta \ln \theta) \Big\}\, d\theta \\ &= [1 + O(v(z)) + O(\varepsilon\ell \ln \theta)][F(\theta) - F(\theta(0))].\end{aligned}$$

After taking the square root and using $O(\ln \theta) = O(\ln[\theta - \theta^2])$, we obtain the assertion of the lemma for $x \in [0, z-]$. The proof for the case $x \in [z+, \ell]$ is similar and is omitted. ∎

5.4. The value ( ). Usually one writes $u = Q + \phi$ and solves ϕ iteratively from the equation $-\varepsilon^2\phi'' + f_u(Q)\phi = -v + O(\phi^2)$. This equation has a small amplitude solution if and only if the right-hand side is (almost) orthogonal to Q', which renders to a small value of $v(z)$ since $Q' = 1/\varepsilon\dot{Q}((x - z)/\varepsilon)$ is similar to a Delta function; see, e.g., Ren and Wei [24]. In this approach, an accurate enough, say $O(\varepsilon^{1/2})$, a priori estimate on ϕ is needed to control the $O(\phi^2)$ term; also, rigorous higher order expansions are needed for more accurate results, which requires higher order differentiability of F. Here we replace Q by θ and take an approach where the a priori estimate on ϕ is replaced by that of $h \approx f^{-1}(-v)$. The key is to carry out the computation that v is almost orthogonal to $Q' \sim \theta'$.

LEMMA 5.6. *For C_0 defined as in (5.8) below, $v(z) = C_0\varepsilon^2 + O(\varepsilon^2\ell^2)$.*

Proof. For $x \ne z$ we have $\theta' = u' - h'$ and $v = \varepsilon^2 u'' - f(u) = \varepsilon^2 h'' - f(h)$. Hence,

$$\begin{aligned}\int_{[0,\ell]\setminus\{z\}} \theta' v &= \int_{[0,\ell]\setminus\{z\}} \{u'(\varepsilon^2 u'' - f(u)) - h'(\varepsilon^2 h'' - f(h))\} \\ &= [F(h) - F(u)]\Big|_0^\ell + \Big(\tfrac{1}{2}\varepsilon^2 h'^2 - F(h)\Big)\Big|_{z-}^{z+} \\ &= O(\varepsilon^2\ell^2 + v^2(z)),\end{aligned} \tag{5.6}$$

since $1 - \theta(\ell) = O(e^{-c\ell/\varepsilon})$, $\theta(0) = O(e^{-c\ell/\varepsilon})$, $h_i' = O(\ell)$, and $F(h(z\pm)) = O((|v(z)| + \varepsilon^2)^2)$.

On the other hand, integration by parts twice gives

$$\begin{aligned}\int_{[0,\ell]\setminus\{z\}} \theta' v = v(z)\left\{ \theta\Big|_0^{z-} + (\theta - 1)\Big|_{z+}^{\ell} \right\} - v'(z)\left\{ \int_0^z \theta + \int_z^\ell (\theta - 1) \right\} \\ + \int_0^z v'' \int_0^x \theta + \int_z^\ell v'' \int_\ell^x (\theta - 1) + \theta(0)v\Big|_0^z + (\theta(\ell) - 1)v\Big|_z^\ell.\end{aligned} \tag{5.7}$$

First, $\{\theta|_0^{z-} + (\theta-1)|_{z+}^{\ell}\} = 1 + O(v(z)) + O(\varepsilon^2)$. Second,

$$\int_0^z \theta = \varepsilon \int_{\theta(0)}^{\theta(z-)} \frac{\theta}{\varepsilon\theta'} d\theta = \varepsilon \int_{\theta(0)}^{\theta(z-)} \frac{[1 + O(v(z)) + O(\varepsilon\ell \ln(s-s^2))]s}{\sqrt{2F(s) - 2F(\theta(0))}} ds$$
$$= \varepsilon \int_0^{Q(0)} \frac{s}{\sqrt{2F(s)}} ds + O(\varepsilon v(z)) + O(\varepsilon^2 \ell),$$

since $\theta(z-) = Q(0) + O(v(z)) + O(\varepsilon^2)$. After a similar estimate for $\int_z^\ell (1-\theta)$ we derive that

$$\int_0^z \theta(x)dx + \int_z^\ell (\theta(x)-1)dx = \varepsilon \left(\int_0^{Q(0)} \frac{s}{\sqrt{2F(s)}} ds + \int_{Q(0)}^1 \frac{s-1}{\sqrt{2F(s)}} ds \right)$$
$$+ O(\varepsilon v(z)) + O(\varepsilon^2 \ell) = O(\varepsilon v(z)) + O(\varepsilon^2 \ell),$$

since the sum of integrals on the right-hand side equals $\int_{-\infty}^0 Q\, d\xi + \int_0^\infty (Q-1)\, d\xi = \int_{\mathbb{R}} \xi \dot{Q}\, d\xi = 0$.

At this moment, we can compare the two results from (5.6) and (5.7) to conclude that $v(z) = O(\varepsilon^2)$ since for $x \in [0, z-]$, $\int_0^x \theta = O(\varepsilon\theta(x))$ and $\int_0^x \int_0^{\hat{x}} \theta = O(\varepsilon^2\theta(x))$ after using $\varepsilon\theta' = [1 + o(1)]\sqrt{2F(\theta) - 2F(\theta(0))}$ and $dx = d\theta/\theta'$. Consequently, $v = O(\ell^2)$, $u = \theta + O(v) + O(\varepsilon^2) = \theta + O(\ell^2)$, and $v'' = m + \gamma v - u = m - \theta + O(\ell^2)$. It then follows that

$$\int_0^z v'' \int_0^x \theta = \int_0^z (m - \theta + O(\ell^2)) \int_0^x \theta$$
$$= \varepsilon^2 \int_0^{Q(0)} \frac{(m-s)\, ds}{\sqrt{2F(s)}} \int_0^s \frac{\hat{s}}{\sqrt{2F(\hat{s})}} d\hat{s} + O(\varepsilon^2 \ell^2).$$

After a similar calculation for the integral on $[z, \ell]$, we then derive that

$$[1 + O(v(z)) + O(\varepsilon)]v(z) = C_0\varepsilon^2 + O(\varepsilon^2\ell^2),$$

where

$$C_0 = \int_0^{Q(0)} \frac{(s-m)\, ds}{\sqrt{2F(s)}} \int_0^s \frac{\hat{s}}{\sqrt{2F(\hat{s})}} d\hat{s} + \int_{Q(0)}^1 \frac{(s-m)\, ds}{\sqrt{2F(s)}} \int_s^1 \frac{1-\hat{s}}{\sqrt{2F(\hat{s})}} d\hat{s}. \tag{5.8}$$

The assertion of the lemma thus follows. $\square$

5.5. The location of interface. The rough estimate of z originates from integrating the equation $v'' = m - u + \gamma v \approx m - u$ so the boundary conditions of v give $\bar{u} \approx m$. That $u \approx 0$ in $[0, z)$ and $u \approx 1$ in $(z, 1]$ gives the first order estimate $z \approx (1-m)\ell$. Equipped with the better estimate that $u = \theta + h_0$ for $x < z$ and $u = \theta + (h_1 - 1)$ for $x > z$, where $f(h_i) + v = O(\varepsilon^2)$, we can perform such integration in a much more accurate manner. Here we notice that

$$\int_0^z \int_0^x \theta \approx \varepsilon^2 \int_0^{Q(0)} \frac{1}{\sqrt{F(s)}} \int_0^s \frac{\hat{s}}{\sqrt{2F(\hat{s})}} d\hat{s} = O(\varepsilon^2).$$

A similar integral on $[z, \ell]$ can also be estimated.

As $v = O(\ell^2)$, by Taylor's expansion, with

$$A_0 = \gamma + 1/f'(0) \quad \text{and} \quad B_0 = f''(0)/(2[f'(0)]^3),$$

we have

$$\gamma v - h_0 = \gamma v - f_0^{-1}(-v) + O(\varepsilon^2) = A_0 v + B_0 v^2 + O(\ell^6) \quad \forall x \in [0, z].$$

Hence, when $x \in [0, z]$,

$$\begin{aligned} v'(x) &= \int_0^x (m + \gamma v - h_0 - \theta) \\ &= mx + \int_0^x (A_0 v + B_0 v^2) - \int_0^x \theta + O(\ell^7), \\ v(x) &= v(z) + \int_z^x v' \\ &= v(z) + \tfrac{1}{2} m(x^2 - z^2) + A_0 \int_z^x d\hat{x} \int_0^{\hat{x}} v + O(\varepsilon^2 + \ell^6). \end{aligned}$$

This gives $v = \frac{1}{2}m(x^2 - z^2) + O(\ell^4 + \varepsilon^2)$, and thus inserting it into the integral gives

$$v(x) = \frac{m}{2}(x^2 - z^2) + \frac{A_0 m}{24}(x^2 - z^2)(x^2 - 5z^2) + O(\ell^6 + \varepsilon^2).$$

Substituting this back into the equation for v' gives

$$v'(z) = mz - \frac{A_0 m}{3} z^3 + \frac{2A_0^2 m + 2B_0 m^2}{15} z^5 - \int_0^z \theta + O(\varepsilon^2 \ell).$$

In a similar manner, working on $[z, \ell]$ and denoting $A_1 = \gamma + 1/f'(1)$ and $B_1 = f''(1)/(2[f'(1)]^3)$, we can derive

$$\begin{aligned} v'(z) = (1 - m)(\ell - z) &- \frac{A_1(1 - m)}{3}(\ell - z)^3 \\ &+ \frac{2A_1^2(1 - m) - 2B_1(1 - m)^2}{15}(\ell - z)^5 - \int_z^\ell (1 - \theta) + O(\varepsilon^2 \ell). \end{aligned}$$

After equating both expressions and using $\int_0^z \theta + \int_z^\ell (\theta - 1) = O(\varepsilon^2 \ell)$, we obtain

$$\begin{aligned} z = (1 - m)\ell &+ \frac{1}{3}\Big\{ m A_0 z^3 - (1 - m) A_1 (\ell - z)^3 \Big\} \\ &+ \frac{2}{15}\Big\{ [A_1^2(1 - m) - B_1(1 - m)^2](\ell - z)^5 - [A_0^2 m + B_0 m^2] z^5 \Big\} + O(\varepsilon^2 \ell). \end{aligned}$$

This "algebraic" equation can be solved iteratively. We state the result as follows.

LEMMA 5.7. *There exist constants c_1, c_2 that depend only on f, m, and γ such that*

$$z = (1 - m)\ell + c_1 \ell^3 + c_2 \ell^5 + O(\varepsilon^2 \ell).$$

Remark 5.1. Here we implicitly assumed that $f \in C^3$, i.e., $F \in C^4$. If we assumed only $F \in C^3$, namely (1.3), then we simply add an $o(\ell^5)$ error.

Here we keep separate traces of error terms in ε and ℓ. They can be used to consider $\ell \in [\pi\sqrt{\varepsilon}, \sqrt[3]{5} L_0 \varepsilon^{1/3}]$.

5.6. Conclusion. Finally, we provide an estimate for u. Writing $Q = Q((x - z)/\varepsilon)$ and $\dot{Q} = \varepsilon Q' = \sqrt{2F(Q)}$, we have

$$\frac{d\theta}{dQ} = \frac{[1 + O(\varepsilon^2) + O(\varepsilon\ell \ln(\theta - \theta^2))]\sqrt{2F(\theta)}}{\sqrt{2F(Q)}}.$$

This provides a relation between θ and Q by

$$\int_{\theta(z\pm)}^{\theta} \frac{d\theta}{[1 + O(\varepsilon^2) + O(\varepsilon\ell \ln(\theta - \theta^2))]\sqrt{2F(\theta)}} = \int_{Q(0)}^{Q} \frac{dQ}{\sqrt{2F(Q)}}.$$

Since $\theta(z\pm) = Q(0) + O(\varepsilon^2)$, we can integrate the above equation to obtain

$$\theta(x) = Q((x - z)/\varepsilon) + O(\varepsilon\ell), \qquad \varepsilon\theta'(x) = \dot{Q}((x - z)/\varepsilon) + O(\varepsilon\ell).$$

Hence we have the following lemma.

LEMMA 5.8. *For all $x \in [0, \ell]$,*

$$\begin{aligned} v(x) &= O(\varepsilon\ell) + \begin{cases} \frac{m}{2}(x^2 - z^2), & x < z, \\ \frac{(1-m)}{2}[(\ell - z)^2 - (\ell - x)^2], & x > z, \end{cases} \\ u(x) &= O(\varepsilon\ell) + Q\Big(\frac{x - z}{\varepsilon}\Big) - \begin{cases} \frac{v(x)}{f_u(0)}, & x \le z, \\ \frac{v(x)}{f_u(1)}, & x \ge z, \end{cases} \\ \varepsilon u'(x) &= O(\varepsilon\ell) + \dot{Q}\Big(\frac{x - z}{\varepsilon}\Big). \end{aligned}$$

In addition, $u''(0) = O(1)$ and $u''(\ell) = O(1)$.

Here the conclusion for $u''(0)$ follows by $u''(0) = h''(0) + \theta''(0) = O(1) + O(e^{-c\ell/\varepsilon})$. The estimate for $u''(\ell)$ is analogous.

6. An eigenvalue estimate. In this section, we study the linear operator

$$\begin{aligned} \mathcal{L}\phi &:= -\varepsilon^2\phi'' + f_u(u)\phi + \mathbf{K}_N^{\ell}(\phi) \\ &= -\varepsilon^2\phi'' + f_u(u)\phi + \gamma^{-1}\,\bar{\phi} + \mathbf{K}_N^{\ell}(\phi - \bar{\phi}), \end{aligned}$$

where (u, v) is a minimizer in $\mathbf{K}_N^{+}(\ell)$. As mentioned earlier, $\bar{\phi}/\gamma$ is written here to remind the reader that when $\gamma = 0$, we work in the class of functions with zero mean value. The additive constant is indeed a Lagrange multiplier which makes $\mathcal{L}$ map zero mean functions to zero mean functions. More precisely,

$$\frac{\bar{\phi}}{\gamma} := -\int_0^1 f_u(u)\phi dx \quad \text{when } \bar{\phi} = \gamma = 0.$$

Such a convention is always imposed when $\gamma = 0$.

The eigenvalue problem has been studied by Nishiura in [16], who showed that the principal (smallest) eigenvalue λ_1 is positive and small. For the case $\gamma = 0$, Ren and Wei [24] demonstrated that $\lambda_1 > c\varepsilon/\ell$ for some positive constant c. The precise value of λ_1 in the $\gamma = 0$ case was given in [23] assuming that F is symmetric, i.e., $F(s) = F(1 - s)$. Here we present a complete analysis giving the precise value of λ_1. The idea [2, 5] is to construct approximate eigenfunctions.

In what follows $\|\cdot\|$ is the L^2 norm and $(\cdot,\cdot)$ the L^2 inner product. Also, $\|\cdot\|_p$ stands for the L^p norm.

6.1. The approximate eigenvalue pair. Denote by Q_1 the solution to

$$-\ddot{Q}_1 + f_u(Q)Q_1 = \sigma - \dot{Q} \quad \text{in } \mathbb{R}, \quad Q_1 \in L^\infty(\mathbb{R}), \quad \int_{\mathbb{R}} Q_1 \dot{Q} d\xi = 0.$$

It is easy to show that Q_1 exists, is unique, and satisfies

$$\int_{-\infty}^0 \Big|Q_1(\xi) - \frac{\sigma}{f_u(0)}\Big| d\xi + \int_0^\infty \Big|Q_1(\xi) - \frac{\sigma}{f_u(1)}\Big| d\xi < \infty.$$

In what follows, we write

$$Q = Q\Big(\frac{x-z}{\varepsilon}\Big), \quad \dot{Q} = \varepsilon Q', \quad Q_1 = Q_1\Big(\frac{x-z}{\varepsilon}\Big), \quad \bar{Q}_1 = \frac{1}{\ell}\int_0^\ell Q_1 dx.$$

We define

$$\hat{\phi}_1(x) = \frac{\sqrt{\varepsilon}}{\sqrt{\sigma}}\Big(u'(x) - \frac{\alpha}{\ell} Q_1\Big), \qquad \hat{\psi}_1 = \mathbf{K}_N^\ell \hat{\phi}_1, \qquad \hat{\lambda}_1 = \frac{\alpha\varepsilon}{\ell}, \tag{6.1}$$

where

$$\alpha := \frac{u(\ell) - u(0)}{\sigma\gamma + \bar{Q}_1} = \frac{1 + O(\ell^2)}{\sigma[\gamma + (1-m)/f_u(0) + m/f_u(1)]}.$$

We shall show that $(\hat{\lambda}_1, \hat{\phi}_1)$ is a good approximation to the principal eigenpair. Here the coefficient is taken so that

$$\|\hat{\phi}_1\|_1 = O(\sqrt{\varepsilon}), \quad \|\hat{\phi}_1\| = 1 + o(1), \quad \|\hat{\phi}_1\|_\infty = O(1/\sqrt{\varepsilon}).$$

6.2. An estimation of .

LEMMA 6.1. *For $\psi = \mathbf{K}_N^\ell \phi$, there holds*

$$\|\psi - \bar{\psi}\|_\infty \le \sqrt{\ell}\|\psi_x\| \le \ell\|\phi - \bar{\phi}\|_1 \le \ell^{3/2}\|\phi - \bar{\phi}\| \le \ell^2\|\phi - \bar{\phi}\|_\infty.$$

Proof. We need only prove the second inequality. Multiplying the equation for ψ by $\psi - \bar{\psi}$, integrating the resulting equation over $(0, \ell)$, and using integration by parts we obtain

$$\int_0^\ell (\psi_x^2 + \gamma(\psi - \bar{\psi})^2) dx = \int_0^\ell \phi(\psi - \bar{\psi}) = \int_0^\ell (\phi - \bar{\phi})(\psi - \bar{\psi})$$
$$\leqslant \|\psi - \bar{\psi}\|_\infty \|\phi - \bar{\phi}\|_1 \leqslant \sqrt{\ell}\|\psi_x\| \, \|\phi - \bar{\phi}\|_1.$$

The assertion thus follows. $\square$

From the lemma, we see that

$$\int_0^\ell \Big\{\hat{\psi}_1'^2 + \gamma(\hat{\psi}_1 - \bar{\hat{\psi}}_1)^2\Big\} \le \ell\|\hat{\phi}_1 - \bar{\hat{\phi}}_1\|_1^2 = O(\varepsilon\ell).$$

6.3. The approximate equation for the principal eigenvalue. Using $-\varepsilon^2 u''' + f_u(u)u' = -v'$, $\dot{Q} = \varepsilon Q'$, and the definition of α, we can calculate

$$\begin{aligned}\mathcal{L}\hat{\phi}_1 &= \frac{\sqrt{\varepsilon}}{\sqrt{\sigma}}\left(-v' + \frac{\alpha}{\ell}(\varepsilon Q' - \sigma) + \frac{u(\ell) - u(0) - \alpha\bar{Q}_1}{\gamma\ell}\right)\\ &\qquad + \hat{\psi}_1 - \bar{\hat{\psi}}_1 + \frac{\alpha\sqrt{\varepsilon}}{\ell\sqrt{\sigma}}(f_u(Q) - f_u(u))Q_1\\ &= \frac{\alpha\varepsilon\sqrt{\varepsilon}}{\ell\sqrt{\sigma}}Q' + \hat{\psi}_1 - \bar{\hat{\psi}}_1 - \frac{\sqrt{\varepsilon}}{\sqrt{\sigma}}v' + \frac{\alpha\sqrt{\varepsilon}}{\ell\sqrt{\sigma}}(f_u(Q) - f_u(u))Q_1 = \hat{\lambda}_1\hat{\phi}_1 + R,\end{aligned}$$

where

$$\hat{\lambda}_1 := \frac{\alpha\varepsilon}{\ell}, \quad R = \hat{\lambda}_1\frac{\sqrt{\varepsilon}}{\sqrt{\sigma}}\Big(Q' - u' + \frac{\alpha}{\ell}Q_1\Big) - \frac{\sqrt{\varepsilon}}{\sqrt{\sigma}}v' + \hat{\psi}_1 - \bar{\hat{\psi}}_1 + \frac{\alpha\sqrt{\varepsilon}}{\ell\sqrt{\sigma}}(f_u(Q) - f_u(u))Q_1. \tag{6.2}$$

Using $Q' - u' = O(\ell^2/\varepsilon)$, $v' = O(\ell)$, one sees that

$$\|R\|_\infty = O(\ell^{5/2}), \qquad \|R\| = O(\ell^3).$$

LEMMA 6.2. *Let* $(\hat{\phi}_1, \hat{\lambda}_1)$ *be defined as in* (6.1). *Then*

$$\mathcal{L}\hat{\phi}_1 = \hat{\lambda}_1\hat{\phi}_1 + R, \quad \|\hat{\phi}_1\| = 1 + o(1), \quad \|R\| = O(\ell^3). \tag{6.3}$$

6.4. Positivity of second eigenvalue. As in [2, 5], a successful analysis on the sign of the tiny principal eigenvalue relies on the fact that the second eigenvalue, if it exists, is positive, uniformly in ε. This allows one to extract information on the principal eigenvalue from an approximate eigenequation, such as (6.3).

LEMMA 6.3. *There exists a positive constant* ν *that is independent of* ε *such that if* $\phi \perp \hat{\phi}_1$, *then with* $\psi = \mathbf{K}_N^\ell(\phi)$,

$$\begin{aligned}\mathrm{L}(\phi, \phi) &:= \int_0^\ell \Big(\varepsilon^2\phi'^2 + f_u(u)\phi^2 + \gamma^{-1}\bar{\phi}^2 + \psi'^2 + \gamma(\psi - \bar{\psi})^2\Big)\\ &\geq \nu\left\{\varepsilon|\phi(0)|^2 + \varepsilon|\phi(\ell)|^2 + \int_0^\ell (\varepsilon^2\phi'^2 + \phi^2 + \psi'^2 + \psi^2)\right\}.\end{aligned}$$

Proof. (a) One need only show the existence of a positive constant ν_1 such that

$$\int_0^\ell (\varepsilon^2\phi'^2 + f_u(Q)\phi^2)\,dx \geq \nu_1\int_0^\ell \phi^2\,dx. \tag{6.4}$$

The reason is as follows. Suppose this is true. Then, for $a \in (0, 1)$ to be chosen at the end,

$$\begin{aligned}&\int_0^\ell (\varepsilon^2\phi'^2 + f_u(u)\phi^2)\,dx \geqslant \int_0^\ell \Big\{\varepsilon^2\phi'^2 + f_u(Q)\phi^2 - C\|u - Q\|_\infty\phi^2\Big\}\\ &\geqslant (1 - a)\int_0^\ell \Big\{\varepsilon^2\phi'^2 + f_u(Q)\phi^2\Big\} + a\{\varepsilon^2\|\phi'\|^2 - \|f_u(Q)\|_\infty\|\phi\|^2\} - C\|u - Q\|_\infty\|\phi\|^2\\ &\geqslant a\varepsilon^2\|\phi'\|^2 + \Big\{(1 - a)\nu_1 - a\|f_u(Q)\|_\infty - C\|u - Q\|_\infty\Big\}\|\phi\|^2\\ &= a\{\varepsilon^2\|\phi'\|^2 + \|\phi\|^2\}\end{aligned}$$

if we take $a = (\nu_1 - C\|u-Q\|_\infty)/(1+\nu_1+\|f_u(Q)\|_\infty)$. Upon observing that $\varepsilon\|\phi\|_\infty^2 \le C\{\varepsilon^2\|\phi'\|^2 + \|\phi\|^2\}$, one obtains the required estimate.

(b) The proof of (6.4) is quite standard; see, for example, de Mottoni and Schatzman [14] or Chen [5]. It is based on the fact that the operator $\mathcal{L}_0 := -\frac{d^2}{d\xi^2} + f_u(Q(\xi))$ on $\mathbb{R}$ has zero as its principal eigenvalue, with a positive eigenfunction $\dot{Q}(\xi)$. Namely, there exists a positive constant $\lambda_2 > 0$ such that for any $\Phi \perp \dot{Q}$

$$L_0(\Phi,\Phi) := \int_{\mathbb{R}} \Big\{\dot{\Phi}^2 + f_u(Q)\Phi^2\Big\} d\xi \ge \lambda_2 \int_{\mathbb{R}} \Phi^2\, d\xi\ .$$

This inequality can be generalized in two ways.

First, the orthogonal condition $\Phi \perp \dot{Q}$ can be relaxed by the assumption that the angle between Φ and $\dot{Q}$ is larger than $\frac{\pi}{4}$, e.g., $\sqrt{2}|(\Phi,\dot{Q})| \le \|\Phi\|\,\|\dot{Q}\|$. Indeed, decomposing $\Phi = c\dot{Q} + \Phi^\perp$ and using $L_0(\dot{Q},\dot{Q}) = 0$ and $L_0(\Phi^\perp,\dot{Q}) = 0$, one has $L_0(\Phi,\Phi) = L_0(\Phi^\perp,\Phi^\perp) \ge \lambda_2\|\Phi^\perp\|^2 \ge \frac{1}{2}\lambda_2\|\Phi\|^2$.

Second, since $f_u(Q(-\infty)) = f_u(0) > 0$ and $f_u(Q(\infty)) = f_u(1) > 0$, the whole space $\mathbb{R}$ can be replaced by a finite interval large enough. Namely, there exists an $M > 0$ such that for any interval $[a,b]$ that contains $[-M,M]$ and any Φ satisfying $\sqrt{2}|(\Phi,\dot{Q})| \le \|\Phi\|\,\|\dot{Q}\|$ (in the $L^2([a,b])$ metric),

$$\int_a^b \Big\{\dot{\Phi}^2 + f_u(Q)\Phi^2\Big\}\, d\xi \ge \frac{\lambda_2}{3}\int_a^b \Phi^2\, d\xi.$$

Translating it into the current situation via $\xi = (x-z)/\varepsilon$, we obtain (6.4) with $\nu_1 = \lambda_2/3$. Here one observes that $\hat{\phi}_1$ is almost parallel to Q' so that $\phi \perp \hat{\phi}_1$ implies that the angle between Q' and ϕ is larger than $\frac{\pi}{4}$. This completes the proof. ∎

6.5. A boundary value problem. To finish our eigenvalue analysis and, more important, for our later applications, we consider the boundary value problem

$$\mathcal{L}\phi = g \quad \text{in } (0,\ell), \quad \phi'(0) \text{ and } \phi'(\ell) \text{ given.}$$

Suppose there is a solution. We write it as

$$\phi = c\hat{\phi}_1 + \phi_\perp, \qquad \phi_\perp \perp \hat{\phi}_1.$$

Then

$$\begin{aligned}(g,\hat{\phi}_1) &= (\mathcal{L}\phi,\hat{\phi}_1) = (-\varepsilon^2\phi'\hat{\phi}_1 + \varepsilon^2\phi\hat{\phi}_1')\Big|_0^\ell + (\phi,\mathcal{L}\hat{\phi}_1)\\ &= \Big(-\varepsilon^2\phi'\hat{\phi}_1 + \varepsilon^2(c\hat{\phi}_1+\phi_\perp)\hat{\phi}_1'\Big)\Big|_0^\ell + \hat{\lambda}_1 c\|\hat{\phi}_1\|^2 + c(R,\hat{\phi}_1) + (R,\phi_\perp).\end{aligned}$$

Since $\|R\| = O(\ell^3)$, $\hat{\lambda}_1 = \alpha\varepsilon/\ell \sim \ell^2$, and $\hat{\phi}_1' = O(\sqrt{\varepsilon})$ at $x=0$ and ℓ, we see that

$$\hat{\lambda}_1\|\hat{\phi}_1\|^2 c = O(1)\left\{(g,\hat{\phi}_1) + \varepsilon^2\phi'\hat{\phi}_1\Big|_0^\ell\right\} + O(\ell^3)\Big\{\|\phi_\perp\| + \sqrt{\varepsilon}|\phi_\perp(0)| + \sqrt{\varepsilon}|\phi_\perp(\ell)|\Big\}.$$

On the other hand,

$$\begin{aligned}(g,\phi_\perp) &= (\mathcal{L}\phi,\phi_\perp) = -\varepsilon^2\phi'\phi_\perp\Big|_0^\ell + c\varepsilon^2\hat{\phi}_1'\phi_\perp\Big|_0^\ell + \mathrm{L}(\phi_\perp,\phi_\perp) + c(\mathcal{L}\hat{\phi}_1,\phi_\perp)\\ &= -\varepsilon^2\phi'\phi_\perp\Big|_0^\ell + O(\ell^3)c\Big\{\|\phi_\perp\| + \sqrt{\varepsilon}|\phi_\perp(0)| + \sqrt{\varepsilon}|\phi_\perp(\ell)|\Big\} + \mathrm{L}(\phi_\perp,\phi_\perp).\end{aligned}$$

Substituting the estimate for c, we then find that

$$\mathrm{L}(\phi_\perp,\phi_\perp)=O(1)\Big(\|g\|+\varepsilon^{3/2}|\phi'(0)|+\varepsilon^{3/2}|\phi'(\ell)|\Big)\Big(\|\phi_\perp\|+O(\sqrt{\varepsilon})[|\phi_\perp(0)|+|\phi_\perp(\ell)|]\Big).$$

After using Lemma 6.3 for $\mathrm{L}(\phi_\perp,\phi_\perp)$, we conclude that

$$\varepsilon\phi_\perp^2(0)+\varepsilon\phi_\perp^2(\ell)+\int_0^\ell\Big(\varepsilon^2{\phi'_\perp}^2+\phi_\perp^2\Big)=O(1)\Big\{\|g\|^2+\varepsilon^3|\phi'(0)|^2+\varepsilon^3|\phi'(\ell)|^2\Big\},\tag{6.5}$$

$$\hat\lambda_1\|\hat\phi_1\|^2c=O(1)\Big\{(g,\hat\phi_1)+\varepsilon^2\phi'\hat\phi_1\Big|_0^\ell\Big\}+O(\ell^3)\Big\{\|g\|+\varepsilon^{3/2}|\phi'(0)|+\varepsilon^{3/2}|\phi'(\ell)|\Big\}.\tag{6.6}$$

We can summarize our calculation as follows.

Lemma 6.4. *Assume that $\mathcal{L}\phi=g$ in $(0,\ell)$. Write $\phi=c\hat\phi_1+\phi_\perp$, where $\phi_\perp\perp\hat\phi_1$. Then* (6.5) *and* (6.6) *hold.*

We remark that the combined quantity $(g,\hat\phi_1)+\varepsilon^2\phi'\hat\phi_1|_0^\ell$ measures the main contribution of the boundary value problem toward the orthogonal projection of the solution in the principal eigenfunction direction. That is, if this quantity is of order 1, then the solution will be of order $1/\hat\lambda_1>c\ell^{-2}$.

6.6. The principal eigenvalue. We can now complete our eigenvalue analysis.

Lemma 6.5. *The principal (smallest) eigenvalue λ_1 of the self-adjoint operator $\mathcal{L}$ is given by*

$$\lambda_1=\hat\lambda_1+O(\ell^3)=\frac{\varepsilon}{\sigma\ell[\gamma+(1-m)/f_u(0)+m/f_u(1)]}+O(\varepsilon).$$

Proof. Note that

$$\lambda_1=\inf_{\|\phi\|=1}\mathrm{L}(\phi,\phi).$$

Taking $\phi=\hat\phi_1/\|\hat\phi_1\|$ as a test function we see that $\lambda_1\le O(\ell^2)$. From (5.2), we see that $\lambda_1\ge 0$. Thus, $\lambda_1=O(\ell^2)$. Let ϕ_1 be the eigenfunction of unit length, $\mathcal{L}\phi_1=\lambda_1\phi_1$. Write $g=\lambda_1\phi_1$ and $\phi_1=c\hat\phi_1+\phi_\perp$, and apply the previous lemma. Since $\phi_1'(0)=\phi_1'(\ell)=0$ and $\|g\|=|\lambda_1|=O(\ell^2)$, the estimate (6.5) then gives $\|\phi_\perp\|=O(\ell^2)$. This implies that $c^2\|\hat\phi_1\|^2=1-\|\phi_\perp\|^2=1-O(\ell^4)$. The equation $\mathcal{L}\phi_1=\lambda_1\phi_1$ can be written as

$$\mathcal{L}\phi_\perp-\lambda_1\phi_\perp=(\lambda_1-\hat\lambda_1)c\hat\phi_1-cR.$$

Taking the inner product with $\hat\phi_1$ and using $(\mathcal{L}\phi_\perp,\hat\phi_1)=-\varepsilon^2\phi'_\perp\hat\phi_1\big|_0^\ell+\varepsilon^2\phi_\perp\hat\phi_1'\big|_0^\ell+(\phi_\perp,\mathcal{L}\hat\phi_1)=\varepsilon^2O(\varepsilon/\ell)+\varepsilon^2O(\sqrt{\varepsilon\ell})+(\phi_\perp,\hat\lambda_1\hat\phi_1+R)=O(\ell^5)$ (note $\phi'_\perp(0)=-c\hat\phi_1'(0)=O(\sqrt{\varepsilon})$, $\hat\phi_1(0)=O(\sqrt{\varepsilon}/\ell)$, $\phi_\perp(0)=O(\sqrt{\ell})$ by (6.5), etc.), we then obtain the required result. ∎

7. Proof of Theorem 1. Now we complete the proof of Theorem 1 parts 2 and 3.

7.1. Uniqueness of the minimizers. Suppose (u,v) and $(\tilde u,\tilde v)$ are two minimizers in $\mathbf{K}_N^+(\ell)$. Denote their interfaces by z and $\tilde z$. From the estimate in section 5 we see that

$$\phi:=u-\tilde u=O(\ell^4)+O(1)\frac{z-\tilde z}{\varepsilon}=o(1)\ell^2\qquad\text{in } L^\infty((0,\ell)).$$

Note that ϕ satisfies

$$\mathcal{L}\phi = f(u+\phi) - f(u) - f_u(u)\phi = O(\phi^2) = o(\ell^2)\phi.$$

Hence,

$$\lambda_1\|\phi\|^2 \le (\mathcal{L}\phi, \phi) = o(\ell^2)\|\phi\|^2.$$

This implies that $\phi \equiv 0$. Hence, minimizers are unique.

7.2. The first order derivative $^+$**.** Once we know the uniqueness of the minimizer, we can then show that minimizers are smooth in the parameter $\ell \in [L_0\varepsilon^{1/3}/16, L_0\varepsilon^{1/3}\sqrt[3]{5}]$. For simplicity, if possible, we use $(u(x), v(x)) = (u(x,\ell), v(x,\ell))$ to denote the energy minimizer in $\mathbf{K}_N^+(\ell)$. Also, we use (u_ℓ, v_ℓ) to denote the derivative of (u, v) with respect to ℓ. Then (u_ℓ, v_ℓ) satisfies

$$\begin{cases} -\varepsilon^2 u_\ell'' + f_u(u)u_\ell + v_\ell = 0 & \text{in } (0,\ell), \\ -v_\ell'' + \gamma v_\ell = u_\ell & \text{in } (0,\ell), \\ v_\ell'(0) = u_\ell'(0) = 0, \quad u_\ell'(\ell) = -u''(\ell), & v_\ell'(\ell) = -v''(\ell), \end{cases}$$

where the boundary conditions are obtained by differentiating $u'(0,\ell) = v'(0,\ell) = u'(\ell,\ell) = v'(\ell,\ell) = 0$.

First we calculate

$$\begin{aligned} E_\ell^+(\varepsilon,\ell) &= F(u) + \tfrac{\gamma}{2}v^2|^{x=\ell} + \int_0^\ell (\varepsilon^2 u' u_\ell' + f(u)u_\ell + v'v_\ell' + \gamma v v_\ell) \\ &= F(u) + \tfrac{\gamma}{2}v^2 - vv''\Big|^{x=\ell} =: c(\varepsilon,\ell) \end{aligned}$$

after integrating by parts and using the equation for u and v_ℓ. Using (5.3) we find that

$$\begin{aligned} E_\ell^+ = c(\varepsilon,\ell) &= \frac{1}{\ell}\int_0^\ell \left\{-\frac{1}{2}\varepsilon^2 u'^2 + \frac{1}{2}v'^2 + F(u) + \frac{\gamma}{2}v^2 - vv''\right\} \\ &= \frac{E^+(\varepsilon,\ell)}{\ell} + \frac{1}{\ell}\int_0^\ell (v'^2 - \varepsilon^2 u'^2)dx. \end{aligned}$$

Hence, as $E^+(\varepsilon,\ell) = \ell\rho^+(\varepsilon,\ell)$,

$$\ell^2\rho_\ell^+(\varepsilon,\ell) = \ell E_\ell^+ - E^+ = \int_0^\ell (v'^2 - \varepsilon^2 u'^2)dx.$$

Using the estimate on the interaction and interfacial energy, we then see that

$$\ell^2\rho_\ell^+ = \frac{m^2(1-m)^2}{3}\ell^3 - \sigma\varepsilon + O(\ell^4).$$

Here one observes an interesting phenomenon: ρ_ℓ^+ does not depend on the first order derivative of (u,v) with respect to ℓ. That is, it does not need equations involving the derivatives of f.

7.3. The second order derivative $^+$. To calculate the second order derivative, we need to use (u_ℓ, v_ℓ). For this purpose, we make the change of variables

$$v_\ell = \psi(x) + A[v''(0)(x-z) - v'(x)], \qquad u_\ell = \phi(x) + A[\gamma v''(0)(x-z) - u'],$$

where

$$A = \frac{-v''(\ell)}{v''(0) - v''(\ell)} = 1 - m + O(\ell^2).$$

As one shall see in a moment, ϕ and ψ are small. That is, $v_\ell \approx A[v''(0)(x-z) - v'(x)]$. This function is obtained from the observation that $v \approx \frac{1}{2}m(x^2 - z^2)$ for $x < z$ and $v = \frac{1}{2}(1-m)[(\ell - z)^2 - (\ell - x)^2]$ for $x > z \approx (1-m)\ell$, and thus $v_\ell \approx (1-m)[-m(1-m) + (x-z)\chi_{[z,\ell]}]$. We found that a good approximation to this function is $A[v''(0)(x-z) - v'(x)]$.

It is easy to verify that $\psi'(0) = \psi'(\ell) = 0$ and $-\psi'' + \gamma\psi = \phi$; namely, $\psi = \mathbf{K}_N^\ell \phi = \gamma^{-1}\bar{\phi} + \mathbf{K}_N^\ell(\phi - \bar{\phi})$. In addition, ϕ satisfies

$$\begin{gathered}\mathcal{L}\phi = g := Av''(0)(z-x)[1 + \gamma f_u(u)],\\ \phi'(0) = A[u''(0) - \gamma v''(0)] = O(1),\\ \phi'(\ell) = (A-1)u''(\ell) - A\gamma v''(0) = O(1).\end{gathered}$$

Now we can calculate

$$\begin{aligned}\frac{d}{d\ell}\Big(\ell^2 \rho_\ell^+(\varepsilon,\ell)\Big) &= 2\int_0^\ell (v'v'_\ell - \varepsilon^2 u'u'_\ell)dx\\ &= I + 2\int_0^\ell (v'\psi' - \varepsilon^2 u'\phi')dx,\end{aligned}$$

where

$$\begin{aligned}I &= 2A\int_0^\ell \Big\{v'[v''(0) - v''] - \varepsilon^2 u'[\gamma v''(0) - u'']\Big\}\\ &= 2Av''(0)[v(\ell) - v(0)] - 2A\gamma v''(0)\varepsilon^2[u(\ell) - u(0)] = m^2(1-m)^2\ell^2 + O(\ell^4).\end{aligned}$$

To estimate the integral involving ϕ and ψ, we write

$$\phi = c\hat{\phi}_1 + \phi_\perp, \quad \psi = c\hat{\psi}_1 + \psi_\perp, \quad \phi_\perp \perp \hat{\phi}_1.$$

Then, by Lemma 6.4,

$$\begin{aligned}&\int_0^\ell \Big\{\varepsilon^2 {\phi'_\perp}^2 + \phi_\perp^2 + {\psi'_\perp}^2 + \gamma\psi_\perp^2\Big\}\\ &\qquad = O(1)\Big\{\|g\|^2 + \varepsilon^3|\phi'(0)|^2 + \varepsilon^3|\phi'(\ell)|^2\Big\} = O(\ell^3).\end{aligned}$$

Since $\|\varepsilon u'\| + \|v'\| = O(\ell^{3/2})$, it then follows that

$$\int_0^\ell (v'\psi'_\perp - \varepsilon^2 u'\phi'_\perp) = O(\ell^3).$$

Also, by Lemma 6.4,

$$
\begin{aligned}
\hat{\lambda}_1\|\hat{\phi}_1\|^2 c &= O(1)\left\{\int_0^\ell g\hat{\phi}_1 dx + \varepsilon^2\hat{\phi}_1\phi'\Big|_0^\ell\right\} + O(\ell^4)\\
&= O(\ell^4) + O(\sqrt{\varepsilon})\int_0^\ell (x-z)[1+\gamma f_u(u)]\Big\{u' - \frac{\alpha}{\ell}Q_1\Big\}\\
&= O(\sqrt{\varepsilon}\ell) + O(\sqrt{\varepsilon})\left\{\int_0^z (u+\gamma f(u))dx + \int_z^\ell (u-1+\gamma f(u))dx\right\}\\
&= O(\sqrt{\varepsilon}\ell) = O(\ell^{5/2}).
\end{aligned}
$$

Thus, $c = O(\sqrt{\ell})$. Consequently, since $\|v'\|^2 = O(\ell^3)$ and $\|\hat{\psi}_1\|^2 = O(\varepsilon\ell) = O(\ell^4)$,

$$
c\int_0^\ell v'\hat{\psi}_1' dx = O(\sqrt{\ell})\|v'\|\,\|\hat{\psi}_1'\| = O(\ell^4).
$$

Using the definition of $\hat{\phi}_1$,

$$
c\int_0^\ell \varepsilon^2 u'\hat{\phi}_1' = \frac{c\sqrt{\varepsilon}}{\sqrt{\sigma}}\int_0^\ell \varepsilon^2 u'\left(u'' - \frac{\alpha}{\ell}Q_1'\right) = -\frac{c\alpha\sqrt{\varepsilon}}{\ell\sqrt{\sigma}}\int_0^\ell \varepsilon^2 u' Q_1' = O\Big(\frac{c\varepsilon\sqrt{\varepsilon}}{\ell}\Big) = O(\ell^4),
$$

since $\|u'\|^2 = O(1/\varepsilon)$ and $\|Q_1'\|^2 = O(1/\varepsilon)$. Thus,

$$
\frac{d}{d\ell}\Big(\ell^2\rho_\ell^+(\varepsilon,\ell)\Big) = m^2(1-m)^2\ell^2 + O(\ell^3).
$$

From this, assertion 3 of Theorem 1 follows. Here note that the estimates of $E_\ell^+(\varepsilon,\ell)$ and $E_{\ell\ell}^+(\varepsilon,\ell)$ follow from the ones of $\rho_\ell^+(\varepsilon,\ell)$ and $\rho_{\ell\ell}^+(\varepsilon,\ell)$ since $E^+(\varepsilon,\ell) = \ell\rho^+(\varepsilon,\ell)$. This also completes the whole proof of Theorem 1. $\square$

8. A final remark for the case $\in (0 \quad {}^{\varepsilon}]$. To make our result more complete, here we provide a final analysis on minimizers for small ℓ.

First of all, one can follow the analysis in the previous sections to conclude that *for all sufficiently small positive ε, $\rho^+(\varepsilon,\cdot) = \rho(\varepsilon,\cdot)$ is strictly decreasing and convex on $[\sqrt{\varepsilon},\ell^\varepsilon]$. In addition,*

$$
\lim_{\varepsilon\to 0,\ell\in[\sqrt{\varepsilon},2L_0\varepsilon^{1/3}/\sqrt[3]{3}]} \frac{E(\varepsilon,\ell)-\sigma\varepsilon}{\ell^3} = \frac{m^2(1-m)^2}{6}.
$$

We omit the proof here.

For $\ell \in (0,\pi\sqrt{\varepsilon}]$ we provide the following as a complement.

LEMMA 8.1. *For every $\varepsilon > 0$, $\rho(\varepsilon,\cdot)$ is a decreasing function on $(0,\pi\sqrt{\varepsilon}]$. More precisely, for every $\varepsilon > 0$, there exists a number $\hat{l}(\varepsilon) \in (0,\pi\sqrt{\varepsilon}]$ such that*

(i) *when $\ell \in (0,\hat{l}(\varepsilon))$, every global minimizer is a constant function, and thus*

$$
\rho(\varepsilon,\ell) = \rho_0 := \min_{s\in\mathbb{R}}\left\{F(s) + \frac{1}{2\gamma}(s-m)^2\right\} \qquad \forall\ell\in(0,\hat{l}(\varepsilon)];
$$

(ii) *in $[\hat{l}(\varepsilon),\pi\sqrt{\varepsilon}]$, $\rho(\varepsilon,\cdot)$ is a strictly decreasing function.*

Proof. If for every $\ell \in (0,\pi\sqrt{\varepsilon}]$ any minimizer is a constant function, then the assertion holds with $\hat{l}(\varepsilon) = \pi\sqrt{\varepsilon}$. Hence, we consider the case that there are non-constant minimizers for some $\ell \in (0,\pi\sqrt{\varepsilon})$.

Suppose $\ell \in (0, \pi\sqrt{\varepsilon})$ and (u, v) is a global nonconstant minimizer of $\mathbf{E}(\cdot,\cdot,\varepsilon,\ell)$ in $\mathbf{K}_N(\ell)$.

Take any $\ell_1 \in (\ell, \pi\sqrt{\varepsilon}]$. Set $\eta = \ell/\ell_1$. Let $\tilde{v}$ be the solution to

$$-\eta^2 \tilde{v}_{xx} + \gamma \tilde{v} = u - m \quad \text{in} \ \ (0,\ell), \qquad \tilde{v}_x(0) = 0 = \tilde{v}_x(\ell).$$

Define

$$(u_1(y), v_1(y)) = (\, u(\eta y), \tilde{v}(\eta y)) \qquad \forall\, y \in (0, \ell_1).$$

Then one can verify that $(u_1, v_1) \in \mathbf{K}_N(\ell_1)$. We shall show that $\frac{1}{\ell_1}\mathbf{E}(u_1, v_1, \varepsilon, \ell_1) < \rho(\varepsilon, \ell)$.

Note that the interaction energy can be written as

$$\int_0^\ell (v_x^2 + \gamma v^2) = \int_0^\ell v(-v_{xx} + \gamma v) = \int_0^\ell v(u - m).$$

A similar calculation for the interaction energy of v_1 then leads to

$$\rho(\varepsilon,\ell) - \frac{\mathbf{E}(u_1, v_1, \varepsilon, \ell_1)}{\ell_1} = \frac{1}{2\ell}\int_0^\ell \Big\{\varepsilon^2[1-\eta^2]u_x^2 + (u-m)(v-\tilde{v})\Big\}dx.$$

The difference $v - \tilde{v}$ can be estimated as follows:

$$\int_0^\ell \Big\{\eta^2(v_x - \tilde{v}_x)^2 + \gamma(v-\tilde{v})^2\Big\}\,dx = \int_0^\ell (v-\tilde{v})\Big\{(-\eta^2 v_{xx} + \gamma v) - (-\eta^2 \tilde{v}_{xx} + \gamma\tilde{v})\Big\}$$

$$= \int_0^\ell (v-\tilde{v})(1-\eta^2)v_{xx} = (1-\eta^2)\int_0^\ell (\tilde{v}_x - v_x)v_x \le (1-\eta^2)\|v_x - \tilde{v}_x\|\,\|v_x\|.$$

This implies that

$$\|v_x - \tilde{v}_x\| \le \frac{1-\eta^2}{\eta^2}\|v_x\|.$$

To continue, we recall the Sobolev inequality $\|g - \bar{g}\| \le \frac{\ell}{\pi}\|g_x\|$ for every $g \in H^1((0,\ell))$. Hence,

$$\int_0^\ell v_x^2 + \gamma(v-\bar{v})^2 = \int_0^\ell (u-m)(v-\bar{v}) = \int_0^\ell (u-\bar{u})(v-\bar{v}) \le \frac{\ell^2}{\pi^2}\,\|u_x\|\,\|v_x\|.$$

It then follows that $\|v_x\| \le \frac{\ell^2}{\pi^2}\|u_x\|$. Consequently, since the averages of v and $\tilde{v}$ are the same,

$$\int_0^\ell (m-u)(v-\tilde{v}) \ = \int_0^\ell (\bar{u}-u)(v-\tilde{v}) \le \frac{\ell^2}{\pi^2}\|u_x\|\,\|v_x - \tilde{v}_x\| \le \frac{\ell^4(1-\eta^2)}{\pi^4\eta^2}\|u_x\|^2.$$

Thus, recalling $\eta = \ell/\ell_1$,

$$\rho(\varepsilon,\ell) - \frac{\mathbf{E}(u_1, v_1, \varepsilon, \ell_1)}{\ell_1} \ge \frac{1-\eta^2}{2\ell}\left(\varepsilon^2 - \frac{\ell^2\ell_1^2}{\pi^4}\right)\int_0^\ell u_x^2\,dx > 0.$$

This implies that $\rho(\varepsilon,\ell) > \rho(\varepsilon,\ell_1)$. This also implies that any minimizer of $\mathbf{E}(\cdot,\cdot,\varepsilon,\ell_1)$ in $\mathbf{K}_N(\ell_1)$ cannot be a constant function (since if it is, then $\rho = \rho_0$).

Now let $\hat{l}(\varepsilon) = \sup\{\ell \mid \rho(\varepsilon,\ell) = \rho_0\}$. We see that when $\ell \in (0, \hat{l})$, any minimizer has to be a constant. On the other hand, $\rho(\varepsilon,\cdot)$ is strictly decreasing in $(\hat{l}, \pi\sqrt{\varepsilon}]$ (if $\hat{l} < \pi\sqrt{\varepsilon}$). ∎

One observes that the above analysis is for any $\varepsilon \in (0,\infty)$.

Acknowledgment. The authors thank the anonymous referee for giving many references and, more important, providing very useful suggestions on making this paper more readable than its earlier version.

REFERENCES

[1] G. Alberti and S. Müller, *A new approach to variational problems with multiple scales*, Comm. Pure Appl. Math., 54 (2001), pp. 761–825.
[2] N. D. Alikakos, X. Chen, and G. Fusco, *Motion of a droplet by surface tension along the boundary*, Calc. Var. Partial Differential Equations, 11 (2000), pp. 233–305.
[3] M. Bahiana and Y. Oono, *Cell dynamical system approach to block copolymers*, Phys. Rev. A, 41 (1990), pp. 6763–6771.
[4] J. Carr, M. E. Gurtin, and M. Slemrod, *Structured phase transitions on a finite interval*, Arch. Ration. Mech. Anal., 86 (1984), pp. 317–351.
[5] X. Chen, *Spectrum for the Allen-Cahn, Cahn-Hilliard, and phase field equations for generic interfaces*, Comm. Partial Differential Equations, 19 (1994), pp. 1371–1395.
[6] R. Choksi, *Scaling laws in microphase separation of diblock copolymers*, J. Nonlinear Sci., 11 (2001), pp. 223–236.
[7] E. N. Dancer and S. Yan, *A minimization problem associated with elliptic systems of FitzHugh-Nagumo type*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 21 (2004), pp. 237–253.
[8] P. C. Fife, *Boundary and interior transition layer phenomena for pairs of second order differential equations*, J. Math. Anal. Appl., 54 (1976), pp. 497–521.
[9] P. C. Fife and D. Hilhorst, *The Nishiura–Ohnishi free boundary problem in the 1D case*, SIAM J. Math. Anal., 33 (2001), pp. 589–606.
[10] I. W. Hamley, *The Physics of Block Copolymers*, Oxford University Press, Oxford, 1998.
[11] M. Henry, *Singular limit of a fourth-order problem arising in the micro-phase separation of diblock copolymers*, Adv. Differential Equations, 6 (2001), pp. 1049–1114.
[12] M. Mimura, M. Tabata, and Y. Hosono, *Multiple solutions of two-point boundary value problems of Neumann type with a small parameter*, SIAM J. Math. Anal., 11 (1980), pp. 613–631.
[13] L. Modica, *The gradient theory of phase transition and the minimal interface criterion*, Arch. Ration. Mech. Anal., 98 (1987), pp. 357–383.
[14] P. de Mottoni and M. Schatzman, *Development of interfaces in R^N*, Proc. Roy. Soc. Edinburgh Sect. A, 116 (1990), pp. 207–220.
[15] S. Müller, *Singular perturbations as a selection criterion for periodic minimizing sequences*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 169–204.
[16] Y. Nishiura, *Dynamics Reported (New Series)*, Vol. 3, Springer-Verlag, Berlin, 1994.
[17] Y. Nishiura and H. Fujii, *Stability of singularly perturbed solutions to systems of reaction-diffusion equations*, SIAM J. Math. Anal., 18 (1987), pp. 1726–1770.
[18] Y. Nishiura and I. Ohnishi, *Some mathematical aspects of the micro-phase separation in diblock copolymers*, Phys. D, 84 (1995), pp. 31–39.
[19] I. Ohnishi, Y. Nishiura, M. Imai, and Y. Matsushita, *Analytical solution describing the phase separation driven by a free energy functional containing a long-range interaction term*, Chaos, 9 (1999), pp. 329–341.
[20] T. Ohta and K. Kawasaki, *Equilibrium morphology of block copolymer melts*, Macromolecules, 19 (1986), pp. 2621–2632.
[21] Y. Oshita, *On stable nonconstant stationary solutions and mesoscopic patterns for FitzHugh-Nagumo equation in higher dimension*, J. Differential Equations, 188 (2003), pp. 110–134.
[22] X. Ren and J. Wei, *On the multiplicity of solutions of two non-local variational problems*, SIAM J. Math. Anal., 31 (2000), pp. 909–924.
[23] X. Ren and J. Wei, *On the spectra of three-dimensional lamellar solutions of the diblock copolymer problem*, SIAM J. Math. Anal., 35 (2003), pp. 1–32.
[24] X. Ren and J. Wei, *On energy minimizers of the diblock copolymer problem*, Interfaces Free Bound., 5 (2003), pp. 193–238.

SIAM J. MATH. ANAL.
Vol. 37, No. 4, pp. 1333–1370

MULTISCALE MODELING FOR THE BIOELECTRIC ACTIVITY OF THE HEART*

MICOL PENNACCHIO†, GIUSEPPE SAVARÉ‡, AND PIERO COLLI FRANZONE‡

Abstract. This paper deals with the mathematical models for the electrical activity of the heart at the micro- and macroscopic levels. By using the tools of the Γ-convergence theory, a rigorous mathematical derivation of the macroscopic model, called "bidomain" and derived directly from the microscopic properties of the tissue, is presented.

Key words. homogenization, Γ-convergence, degenerate evolution equations, reaction-diffusion systems, FitzHugh–Nagumo dynamic, cardiac electric field, bidomain model

AMS subject classifications. 35B27, 35K57, 35K65, 92C30, 93A30

DOI. 10.1137/040615249

1. Introduction.

The aim of this work is to study, in the framework of Γ-convergence theory, the asymptotic behavior of a microscopic-level (i.e., cellular level) modeling problem for the bioelectric activity of the heart.

The cardiac tissue is composed of an arrangement of elongated cells physically interconnected by specialized membrane structures of densely packed channels called *gap junctions*; see, e.g., [30, 22]. The *gap junction* channels connect the cytoplasmatic compartments of adjacent cells and allow the intercellular flow of ionic currents. The intercellular communication between cardiac myocytes occurs in an end-to-end orientation and in a side-to-side apposition. At a cellular level the tissue can be viewed as composed of two conducting volumes, the intra- and extracellular spaces, separated by the cellular membrane. The two spaces are considered ohmic conducting media and since the junctional resistance between two adjoining cells is different from the myoplasm of either cell, the intracellular conductivity is space dependent. The microscopic mathematical model consists of a system of two partial differential equations of elliptic type and the unknown functions are the intra- and extracellular electric potentials. These equations are coupled by means of a distinctive evolutive boundary condition in the potential jump at the interface separating the two media, i.e., the cellular membrane.

The problem, written in a nondimensional form, contains a small parameter ε related to the microstructure.

In spite of the discrete cellular structure, it is well known that the cardiac tissue can be represented by a continuous model, called the *bidomain model*, which attempts to describe the averaged electric potentials and current flows inside (intracellular space) and outside (extracellular space) the cardiac cells. Despite the widespread use of the macroscopic bidomain model, its rigorous derivation directly from the microscopic properties of the tissue is still lacking. Formally, the macroscopic equations

*Received by the editors September 17, 2004; accepted for publication (in revised form) May 24, 2005; published electronically January 6, 2006. This work was partially supported by grants of M.I.U.R. (PRIN 2003011441) and of Progetti Intergruppo Istituto Nazionale di Alta Matematica, Roma, Italy.

http://www.siam.org/journals/sima/37-4/61524.html

†Istituto di Matematica Applicata e Tecnologie Informatiche del C.N.R., via Ferrata, 1–27100 Pavia, Italy (micol@imati.cnr.it).

‡Dipartimento di Matematica "F. Casorati", Università di Pavia, Via Ferrata, 1–27100 Pavia, Italy (giuseppe.savare@unipv.it, colli@imati.cnr.it).

can be obtained from the microscopic ones by multiple scale expansion and averaging. For instance, a first formal derivation, based on current balances and expressed by averages of integral identities, was obtained in [28]. By standard multiscale arguments of homogenization the same formal derivation can be found also in the appendix of [16] and in [25, 26].

We investigate the homogenization limit when $\varepsilon \to 0$ under the simplifying assumption that cardiac cells are arranged in a periodic box structure. In this work the micro- and the macroscopic structures of the cardiac tissue are studied by using the tools of the Γ-convergence theory, and a rigorous mathematical derivation of the limit problem at the tissue level (i.e., the bidomain model) is presented.

The microscopic model of the cardiac tissue.

Intra- and extracellular regions. At a microscopic level the cardiac tissue Ω (a bounded Lipschitz open subset of $\mathbb{R}^d$, $d = 3$) is composed of a collection of elongated cardiac cells, connected end-to-end and/or side-to-side by junctions, surrounded by the extracellular fluid. The end-to-end contacts form the fiber structure of the cardiac muscle, whereas the presence of lateral junctions establishes a connection between the elongated fibers.

We can consider the cardiac tissue Ω as composed of two *connected* regions, the *intracellular* (inside the cells) Ω_i^ε, separated from the *extracellular* (fluid outside the cells) Ω_e^ε by a membrane surface Γ^ε; thus $\Omega = \Omega_i^\varepsilon \cup \Omega_e^\varepsilon \cup \Gamma^\varepsilon$, and $\Gamma^\varepsilon = \partial\Omega_i^\varepsilon \cap \partial\Omega_e^\varepsilon$ is the common part of the two boundaries of $\Omega_{i,e}^\varepsilon$. Here $\varepsilon > 0$ is a small dimensionless parameter (whose precise definition in terms of the various physical constants will be discussed in the appendix) which is proportional to the ratio between the *"micro" scale* of the length of the cells and the *"macro" scale* of the length of the cardiac fibers.

The periodic lattice of the cells. Following the standard approach of the homogenization theory, we are assuming that the cells are distributed according to an ideal periodic organization similar to a regular lattice of interconnected cylinders.

If ${}_1, \ldots, {}_d$ is an orthogonal basis of $\mathbb{R}^d$, we denote by

$$E_i, \quad E_e := \mathbb{R}^d \setminus \overline{E}_i \quad \text{with common boundary } \Gamma := \partial E_i \cap \partial E_e,$$

two reference open, *connected, and periodic* subsets of $\mathbb{R}^d$ with Lipschitz boundary, i.e., satisfying

$$E_{i,e} + {}_k = E_{i,e}, \quad k = 1, \ldots, d. \tag{1.1}$$

The elementary periodicity region

$$Y := \left\{ \sum_{k=1}^{d} \alpha_k \, {}_k : 0 \le \alpha_k < 1, \quad k = 1, \ldots, d \right\} \tag{1.2}$$

where its intra- and extracellular parts $Y_{i,e} = Y \cap E_{i,e}$ represents a reference unit volume box containing a single cell Y_i. The main geometrical assumption is that the physical intra- or extracellular regions are the ε-dilation of the reference lattices $E_{i,e}$, defined as

$$\varepsilon E_{i,e} = \big\{ \varepsilon\xi : \xi \in E_{i,e} \big\} \quad \text{with} \quad \varepsilon\Gamma := \big\{ \varepsilon\xi : \xi \in \Gamma \big\}, \tag{1.3}$$

and therefore the decomposition of the physical region Ω occupied by the heart into the intra- and extracellular domains $\Omega_{i,e}^\varepsilon$ (see, e.g., Figure 1.1) can be obtained simply

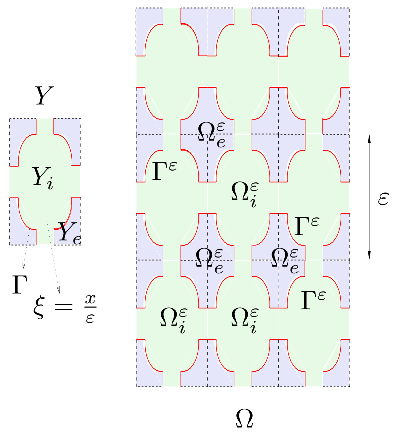


FIG. 1.1. *Right: The ideal periodic geometry in a bidimensional section of the simplified three-dimensional periodic network of interconnected cells. Left: Unit cell in the microscopic variable* $\xi = x/\varepsilon$.

by intersecting Ω with $\varepsilon E_{i,e}$, i.e.,

$$\Omega_i^\varepsilon = \Omega \cap \varepsilon E_i, \quad \Omega_e^\varepsilon = \Omega \cap \varepsilon E_e, \quad \Gamma^\varepsilon = \Omega \cap \left(\partial\Omega_i^\varepsilon \cap \partial\Omega_e^\varepsilon\right) = \Omega \cap \varepsilon\Gamma. \tag{1.4}$$

Unknowns and equations. The electric properties of the tissue are described by the couple,

$$\underline{u}^\varepsilon = (u_i^\varepsilon, u_e^\varepsilon), \quad u_{i,e}^\varepsilon : \Omega_{i,e}^\varepsilon \to \mathbb{R}, \tag{1.5}$$

of *intra- and extracellular potentials,* each one admitting a trace $u_{i,e}^\varepsilon|_{\Gamma^\varepsilon}$ on Γ^ε, whose difference

$$v^\varepsilon := u_i^\varepsilon|_{\Gamma^\varepsilon} - u_e^\varepsilon|_{\Gamma^\varepsilon} : \Gamma^\varepsilon \to \mathbb{R} \tag{1.6}$$

is the *transmembrane potential* (in the following, we will simply write $v^\varepsilon = u_i^\varepsilon - u_e^\varepsilon$ on Γ^ε) and satisfies a *dynamic condition on* Γ^ε involving auxiliary functions

$$w^\varepsilon : \Gamma^\varepsilon \to \mathbb{R}^h, \tag{1.7}$$

the so-called *gating (or recovery) variables.* We denote by $\sigma_{i,e}^\varepsilon$ suitably rescaled symmetric conductivity matrices,

$$\sigma_{i,e}^\varepsilon(x) = \sigma_{i,e}\left(x, \frac{x}{\varepsilon}\right), \tag{1.8}$$

obtained by continuous functions $\sigma_{i,e}(x,\xi) : \Omega \times E_{i,e} \to \mathbb{M}^{d\times d}$ satisfying the usual uniform ellipticity and periodicity conditions

$$\begin{matrix} \sigma|y|^2 \le \sigma_{i,e}(x,\xi)y\cdot y \le \sigma^{-1}|y|^2 \\ \sigma_{i,e}(x,\xi + {}_k) = \sigma_{i,e}(x,\xi) \end{matrix} \quad \forall\,(x,\xi) \in \Omega \times E_{i,e}, \qquad y \in \mathbb{R}^d, \tag{1.9}$$

for a given constant $\sigma > 0$; $\nu_{i,e}^\varepsilon$ are the exterior unit normals to the boundaries of $\Omega_{i,e}^\varepsilon$: observe that $\nu_i^\varepsilon = -\nu_e^\varepsilon$ on Γ^ε.

We can formulate the reaction-diffusion system satisfied by the vector $(u_i^\varepsilon, u_e^\varepsilon, w^\varepsilon)$, with $v^\varepsilon = u_i^\varepsilon - u_e^\varepsilon$ on Γ^ε, in the following way:

$$(P_{1a}^\varepsilon) \qquad -\operatorname{div}\left(\sigma_{i,e}^\varepsilon \nabla u_{i,e}^\varepsilon\right) = 0 \qquad \text{in } \Omega_{i,e}^\varepsilon \times (0,T) \qquad \textit{(quasi-stationary conduction),}$$

$$(P_{1b}^\varepsilon) \qquad \left.\begin{matrix} -\sigma_i^\varepsilon \nabla u_i^\varepsilon \cdot \nu_i^\varepsilon \\ \sigma_e^\varepsilon \nabla u_e^\varepsilon \cdot \nu_e^\varepsilon \end{matrix}\right\} = I_m^\varepsilon \quad \text{on } \Gamma^\varepsilon \times (0,T) \qquad \textit{(continuity equation),}$$

$$(P_2^\varepsilon) \qquad \varepsilon\left(\partial_t v^\varepsilon + I(v^\varepsilon, w^\varepsilon)\right) = I_m^\varepsilon \quad \text{on } \Gamma^\varepsilon \times (0,T) \qquad \textit{(reaction surface condition),}$$

$$(P_3^\varepsilon) \qquad \partial_t w^\varepsilon + r(v^\varepsilon, w^\varepsilon) = 0 \qquad \text{on } \Gamma^\varepsilon \times (0,T) \qquad \textit{(dynamic coupling)}$$

supplemented by the boundary and initial conditions

$$(P_4^\varepsilon) \qquad \sigma_{i,e}^\varepsilon \nabla u_{i,e}^\varepsilon \cdot \nu_{i,e} = 0 \qquad \text{on } (\partial\Omega_{i,e}^\varepsilon \setminus \Gamma^\varepsilon) \times (0,T),$$

$$(P_5^\varepsilon) \qquad v^\varepsilon(\cdot,0) = v_0^\varepsilon \quad \text{on } \Gamma^\varepsilon,$$

$$(P_6^\varepsilon) \qquad w^\varepsilon(\cdot,0) = w_0^\varepsilon \quad \text{on } \Gamma^\varepsilon,$$

where the coupling terms $I(v^\varepsilon, w^\varepsilon)$ (the *membrane ionic current*) and $r(v^\varepsilon, w^\varepsilon)$ depend on the particular model of the ionic flux through the cellular membrane chosen.

Here we are mainly concerning with the so-called *FitzHugh–Nagumo model,* first introduced as a simplified membrane kinetic of the Hodgkin–Huxley equations for the transmission of nervous electric impulses (see, e.g., [13, 20]) that requires only one scalar *recovery variable* w^ε (thus $h = 1$ in (1.7)). Therefore, I and r take the form

$$(P_7^\varepsilon) \qquad I(v^\varepsilon, w^\varepsilon) := F(v^\varepsilon) + \Theta w^\varepsilon, \quad r(v^\varepsilon, w^\varepsilon) := \gamma w^\varepsilon - \eta v^\varepsilon,$$

where Θ, γ, η are nonnegative constants, and

$$(1.10) \qquad F \in C^1(\mathbb{R}) \quad \text{is a cubic-like function with} \quad \inf_{x\in\mathbb{R}} F'(x) > -\infty.$$

If $(u_i^\varepsilon, u_e^\varepsilon, w^\varepsilon)$ is a solution of this system, it is easy to check that $(u_i^\varepsilon + c, u_e^\varepsilon + c, w^\varepsilon)$ is still a solution, where $c = c(t)$ is an arbitrary family of additive time-dependent constants. We avoid the use of quotient spaces by fixing a reference open subdomain

$$(1.11) \qquad \Omega_0 \subset\subset \Omega \quad \text{with} \quad \mathscr{L}^d(\partial\Omega_0) = 0, \quad \mathscr{L}^d(\Omega_0 \cap \Omega_e^\varepsilon) > 0,$$

(here $\mathscr{L}^d$ denotes the usual Lebesgue measure on $\mathbb{R}^d$) and imposing that

$$(P_8^\varepsilon) \qquad \int_{\Omega_e^\varepsilon \cap \Omega_0} u_e^\varepsilon(x)\,dx = 0.$$

We refer to the system $(P_{1a}^\varepsilon, P_{1b}^\varepsilon, \ldots, P_8^\varepsilon)$ as the *(microscopic or cellular) problem* $^{\varepsilon}$.

Well posedness of $^{\varepsilon}$ *and energy estimate.* The well posedness of problem $^{\varepsilon}$ in suitable function spaces has been studied in [16], whose main result we will report in section 4; here we recall only one of the basic a priori estimates, which involves the energy-like functionals

$$\underline{a}^\varepsilon\big(\underline{u}^\varepsilon\big) := \sum_{i,e} \int_{\Omega^\varepsilon_{i,e}} \sigma^\varepsilon_{i,e} \nabla u^\varepsilon_{i,e} \cdot \nabla u^\varepsilon_{i,e}\, dx, \quad \underline{u}^\varepsilon = (u^\varepsilon_i, u^\varepsilon_e), \tag{1.12a}$$

$$b^\varepsilon\big(v^\varepsilon\big) := \varepsilon \int_{\Gamma^\varepsilon} |v^\varepsilon|^2\, d\mathscr{H}^{d-1}, \quad b^\varepsilon\big(w^\varepsilon\big) := \varepsilon \int_{\Gamma^\varepsilon} |w^\varepsilon|^2\, d\mathscr{H}^{d-1}, \tag{1.12b}$$

$$\phi^\varepsilon\big(v^\varepsilon\big) := \varepsilon \int_{\Gamma^\varepsilon} \varphi(v^\varepsilon)\, d\mathscr{H}^{d-1}, \tag{1.12c}$$

$$j^\varepsilon\big(v^\varepsilon\big) := \inf\Big\{\underline{a}^\varepsilon\big(\underline{u}^\varepsilon\big) : u^\varepsilon_{i,e} \in H^1(\Omega^\varepsilon_{i,e}),\ u^\varepsilon_i - u^\varepsilon_e = v^\varepsilon \text{ on } \Gamma^\varepsilon\Big\}, \tag{1.12d}$$

where φ is a positive, convex, primitive function of $x \mapsto F(x) + \lambda_F x$ for a sufficiently big $\lambda_F > -\inf_{x\in\mathbb{R}} F'(x)$ (see (4.1)) and $\mathscr{H}^{d-1}$ denotes the usual $(d-1)$-dimensional Hausdorff measure.

For every $v^\varepsilon_0, w^\varepsilon_0 \in L^2(\Gamma^\varepsilon)$ (the L^2 space with respect to $\mathscr{H}^{d-1}$) with $j(v^\varepsilon_0) < +\infty$, there exists a unique variational solution $u^\varepsilon_{i,e}, w^\varepsilon, v^\varepsilon = u^\varepsilon_i - u^\varepsilon_e$ on Γ^ε, of problem $^\varepsilon$ satisfying the uniform energy bound

$$\sup_{t\in[0,T]} \Big(\underline{a}^\varepsilon\big(\underline{u}^\varepsilon\big) + b^\varepsilon\big(w^\varepsilon\big) + \phi^\varepsilon\big(v^\varepsilon\big)\Big) \le C\Big(j^\varepsilon\big(v^\varepsilon_0\big) + b^\varepsilon\big(w^\varepsilon_0\big) + \phi^\varepsilon\big(v^\varepsilon\big)\Big) \tag{1.13}$$

for a constant $C = C(T, \lambda_F, \eta, \Theta)$ independent of ε.

Convergence to the solution of the macroscopic bidomain model. One possible way to describe the asymptotic behavior of $u^\varepsilon_{i,e}, v^\varepsilon, w^\varepsilon$ as $\varepsilon \downarrow 0$ is to consider local averages. First, we denote by $\beta_{i,e}, \beta$ the asymptotic local ratios (uniform in space) of the intra- and extracellular volumes and of the membrane surface area to the volume occupied by the tissue, i.e.,

$$\begin{aligned} \beta_{i,e} &:= \lim_{\varepsilon\downarrow 0} \frac{\mathscr{L}^d(\Omega^\varepsilon_{i,e} \cap B_\rho(x))}{\mathscr{L}^d(\Omega \cap B_\rho(x))} = \frac{\mathscr{L}^d(Y_{i,e})}{\mathscr{L}^d(Y)} \\ \beta &:= \lim_{\varepsilon\downarrow 0} \frac{\varepsilon\mathscr{H}^{d-1}(\Gamma^\varepsilon \cap B_\rho(x))}{\mathscr{L}^d(\Omega \cap B_\rho(x))} = \frac{\mathscr{H}^{d-1}(\Gamma \cap Y)}{\mathscr{L}^d(Y)} \end{aligned} \qquad \forall\, \rho > 0, \qquad x \in \Omega. \tag{1.14}$$

Taking into account the a priori bound (1.13), we will introduce the following definition (where z^ε represents either the transmembrane potential v^ε or the recovery variable w^ε).

Definition 1.1 (a weak notion of convergence). *Let $u^\varepsilon_{i,e} \in L^1_{loc}(\Omega^\varepsilon_{i,e}), z^\varepsilon \in L^1_{loc}(\Gamma^\varepsilon)$, $\varepsilon > 0$, be given families of functions. We say that $u^\varepsilon_{i,e}$ converges to $u_{i,e} \in L^1_{loc}(\Omega)$ and z^ε converges to $z \in L^1_{loc}(\Omega)$ as $\varepsilon \downarrow 0$ if for every test function $\zeta \in C^0_c(\Omega)$ we have*

$$\lim_{\varepsilon\downarrow 0} \int_{\Omega^\varepsilon_{i,e}} u^\varepsilon_{i,e}(x)\zeta(x)\, dx = \beta_{i,e} \int_\Omega u(x)\zeta(x)\, dx, \tag{1.15}$$

$$\lim_{\varepsilon\downarrow 0} \varepsilon \int_{\Gamma^\varepsilon} z^\varepsilon(x)\zeta(x)\, d\mathscr{H}^{d-1}(x) = \beta \int_\Omega z(x)\zeta(x)\, dx. \tag{1.16}$$

We say that a vector $(u^\varepsilon_i, u^\varepsilon_e, z^\varepsilon) \in L^1_{loc}(\Omega^\varepsilon_i) \times L^1_{loc}(\Omega^\varepsilon_e) \times L^1_{loc}(\Gamma^\varepsilon)$ is converging to $(u_i, u_e, z) \in \big(L^1_{loc}(\Omega)\big)^3$ if each component is converging to the corresponding one according to (1.15) *and* (1.16).

Remark 1.2 (weak* convergence of the associated measures). The above formulae correspond to considering the local weak* convergence in the sense of (signed) measures: to clarify this point, we introduce the reference positive measures

$$\lambda^\varepsilon_{i,e} := \mathscr{L}^d|_{\Omega^\varepsilon_{i,e}}, \quad \lambda^\varepsilon := \varepsilon\mathscr{H}^{d-1}|_{\Gamma^\varepsilon}, \quad \lambda_{i,e} := \beta_{i,e}\mathscr{L}^d|_{\Omega}, \quad \lambda := \beta\mathscr{L}^d|_{\Omega} \tag{1.17}$$

and the Radon measures

$$\tilde u^\varepsilon_{i,e} := u^\varepsilon_{i,e}\cdot\lambda^\varepsilon_{i,e}, \quad \tilde z^\varepsilon := z^\varepsilon\cdot\lambda^\varepsilon \tag{1.18}$$

whose densities are $u^\varepsilon_{i,e}$ and z, respectively. The convergence introduced in Definition 1.1 is then equivalent to asking whether

$$\tilde u^\varepsilon_{i,e} \rightharpoonup^* \tilde u_{i,e} := u_{i,e}\lambda_{i,e}, \quad \tilde z^\varepsilon \rightharpoonup^* \tilde z := z\cdot\lambda \tag{1.19}$$

in the local weak* topology of the space of Radon measures [4, Definition 1.58].

Since the microscopic problem is strictly related to the energy functionals (1.12a), (1.12b), it is natural to introduce the corresponding homogenized ones, which are defined by

$$\underline{a}(\underline{u}) := \sum_{i,e}\int_\Omega M_{i,e}(x)\nabla u_{i,e}(x)\cdot\nabla u_{i,e}(x)\,dx, \quad \underline{u} := (u_i,u_e), \tag{1.20a}$$

$$b(v) := \beta\int_\Omega |v(x)|^2\,dx, \quad b(w) := \beta\int_\Omega |w(x)|^2\,dx, \tag{1.20b}$$

$$\phi(v) := \beta\int_\Omega \varphi(v(x))\,dx, \tag{1.20c}$$

$$j(v) := \inf\Big\{\underline{a}(\underline{u}) : \underline{u} = (u_i,u_e)\in H^1(\Omega)\times H^1(\Omega),\ u_i-u_e = v \text{ in } \Omega\Big\}, \tag{1.20d}$$

where $M_i(x), M_e(x)$ are the symmetric and positive definite matrices obtained by solving the cellular problems for every $y\in\mathbb{R}^d$,

$$\begin{aligned} M_{i,e}(x)\,y\cdot y := \min\Bigg\{&\frac{1}{\mathscr{L}^d(Y)}\int_{Y_{i,e}}\sigma_{i,e}(x,\xi)\big(\nabla u(\xi)+y\big)\cdot\big(\nabla u(\xi)+y\big)\,d\xi : \\ &u\in H^1_{loc}(\mathbb{R}^d),\ u\ Y\text{-periodic}\Bigg\}, \end{aligned} \tag{1.21}$$

and satisfying the usual uniform ellipticity condition for a constant $\mu>0$,

$$\mu|y|^2 \le M_{i,e}(x)y\cdot y \le \mu^{-1}|y|^2 \quad \forall\, y\in\mathbb{R}^d, \qquad x\in\Omega. \tag{1.22}$$

We have now all the elements to state our main result.

THEOREM 1.3 (convergence to the macroscopic problem). *Let us suppose that, as $\varepsilon\downarrow 0$, the initial data $v^\varepsilon_0, w^\varepsilon_0\in L^2(\Gamma^\varepsilon)$ converge to $v_0, w_0\in L^2(\Omega)$ according to* (1.16)*, and the related energies satisfy*

$$\lim_{\varepsilon\downarrow 0} b^\varepsilon(v^\varepsilon_0) = b(v_0), \quad \lim_{\varepsilon\downarrow 0} b^\varepsilon(w^\varepsilon_0) = b(w_0), \quad \limsup_{\varepsilon\downarrow 0}\Big(j^\varepsilon(v^\varepsilon_0)+\phi^\varepsilon(v^\varepsilon_0)\Big) < +\infty. \tag{1.23}$$

Then there exists $\underline{u} = (u_i,u_e)$, $v = u_i-u_e$, w with

$$u_{i,e}, v, w\in C^0(0,T;H^1(\Omega)), \quad \partial_t v,\ \partial_t w\in L^2(0,T;L^2(\Omega)),$$

such that for every time $t \in [0,T]$

$$(u^{\varepsilon}_{i,e}, v^{\varepsilon}, w^{\varepsilon}) \to (u_{i,e}, v, w) \quad \textit{as } \varepsilon \downarrow 0 \textit{ according to Definition } 1.1, \tag{1.24}$$

with

$$\underline{a}^{\varepsilon}(\underline{u}^{\varepsilon}) = j^{\varepsilon}(v^{\varepsilon}) \to \underline{a}(\underline{u}) = j(v), \quad b^{\varepsilon}(v^{\varepsilon}) \to b(v), \quad b^{\varepsilon}(w^{\varepsilon}) \to b(w). \tag{1.25}$$

$\big(u_{i,e}, w\big)$ *with* $v = u_i - u_e$ *is the (unique) variational solution of the* macroscopic reaction-diffusion system

$$(P_1) \qquad \left.\begin{array}{c} \operatorname{div}(M_i \nabla u_i) \\ -\operatorname{div}(M_e \nabla u_e) \end{array}\right\} = I_m \qquad \textit{in } \Omega \times (0,T) \qquad \textit{(continuity equation)},$$

$$(P_2) \qquad \beta\,(\partial_t v + I(v,w)) = I_m \qquad \textit{in } \Omega \times (0,T) \qquad \textit{(reaction-diffusion condition)},$$

$$(P_3) \qquad \partial_t w + r(v,w) = 0 \qquad \textit{in } \Omega \times (0,T), \qquad \textit{(dynamic coupling)}$$

supplemented by the boundary and initial conditions

$$(P_4) \qquad M_{i,e} \nabla u_{i,e} \cdot \nu_{i,e} = 0 \qquad \textit{on } \partial\Omega \times (0,T),$$

$$(P_5) \qquad v(\cdot,0) = v_0 \qquad \textit{in } \Omega,$$

$$(P_6) \qquad w(\cdot,0) = w_0 \qquad \textit{in } \Omega;$$

again, the coupling terms I *and* r *take the same form as* (P^{ε}_7),

$$(P_7) \qquad I(v,w) := F(v) + \Theta w, \quad r(v,w) := \gamma w - \eta v,$$

and a reference value for the potential u_e *is determined by imposing*

$$(P_8) \qquad \int_{\Omega_0} u_e(x)\,dx = 0.$$

Thus we find the equations of the so-called *bidomain model* (see, e.g., [21, 14, 15, 31, 24]): it describes at a macroscopic level the averaged electric potentials and current flows inside (intracellular space) and outside (extracellular space) the cardiac cells, disregarding the discrete cellular structure and representing the cardiac tissue as the superposition of two interpenetrating and superimposed continua. In this representation Ω, the physical region occupied by the heart, coincides with the intra- and extracellular domains and at every point the two media are connected by a distributed cellular membrane on Ω. The two superposed conducting media are ohmic, i.e., their current densities are given by $\mathbf{j}_{i,e} = -M_{i,e}\nabla u_{i,e}$ with $M_{i,e}$ the conductivity tensors. Thus condition P_1 is the current conservation law

$$\operatorname{div}(\mathbf{j}_i + \mathbf{j}_e) = 0 \qquad \text{and} \qquad -\ \operatorname{div}\mathbf{j}_i = \operatorname{div}\mathbf{j}_e = I_m,$$

where I_m is the current per unit volume crossing the cellular membrane.

Convergence by extensions of the data. One could also consider a different approach to capture the asymptotic behavior of $u^{\varepsilon}_{i,e}, v^{\varepsilon}, w^{\varepsilon}$ by performing a preliminary extension of them to the whole Ω and considering the limit of such extensions in a suitable (ε-independent) function space.

THEOREM 1.4 (weak convergence in H^1_{loc} for extended solutions). *Let us assume that the initial data* $v^{\varepsilon}_0, w^{\varepsilon}_0$ *converge to* v_0, w_0 *according to* (1.15), (1.16) *of Definition* 1.1, *they satisfy the energy condition* (1.23), *and there exist extensions* $\breve{w}^{\varepsilon}_0$ *of* w^{ε}_0

which are bounded in $H^1(\Omega')$*, for every* $\Omega' \subset\subset \Omega$*. Then there exist extensions* $\breve{u}^\varepsilon_{i,e}$*,* $\breve{w}^\varepsilon$ *of the microscopic solutions* $u^\varepsilon_{i,e}, w^\varepsilon$ *of problem* ${}^{\boldsymbol{\varepsilon}}$ *satisfying*

$$\sup_{t\in[0,T],\varepsilon>0} \int_{\Omega'} \Big(|\breve{u}^\varepsilon_{i,e}|^2 + |\breve{w}^\varepsilon|^2 + |\nabla \breve{u}^\varepsilon_{i,e}|^2 + |\nabla \breve{w}^\varepsilon|^2 \Big)\, dx < +\infty \quad \forall\, \Omega' \subset\subset \Omega; \tag{1.26}$$

moreover, for every $\Omega' \subset\subset \Omega$ *and* $t \in [0,T]$*, any family* $\breve{u}^\varepsilon_{i,e}, \breve{w}^\varepsilon$ *of such extensions will satisfy*

$$\breve{u}^\varepsilon_{i,e} \rightharpoonup u_{i,e}, \quad \breve{w}^\varepsilon \rightharpoonup w \quad \textit{weakly in } H^1(\Omega') \quad \textit{as } \varepsilon \downarrow 0, \tag{1.27}$$

where $(u_{i,e}, w)$ *is the solution of the macroscopic problem* .

Remark 1.5. As we will discuss in section 2 the existence of admissible extensions $\breve{u}^\varepsilon_{i,e}, \breve{w}^\varepsilon$ satisfying the uniform bounds (1.26) follows by a general result of Acerbi et al. [1]; they also show that only local a priori bounds like (1.26) are available, due to the particular geometry of this problem: in fact, the boundary of $\Omega^\varepsilon_{i,e}$ could be quite irregular and one cannot find global extension operators which preserve the H^1-norm.

Homogenization and Γ*-convergence of the associated stationary problems.* As we shall discuss in more detail in section 5, the microscopic problems ($P^\varepsilon_{1a}, \dots, P^\varepsilon_8$) can be considered as a sort of (perturbation of) gradient flows of ε-dependent energies with respect to a varying family of degenerate metrics, which are induced by nonnegative quadratic forms with a nontrivial kernel (the forms b^ε of (1.12b)).

The characterization of the asymptotic behavior of the energy functionals, in the framework of Γ-convergence theory, is one of the crucial step of the proof of Theorem 1.3 and it is naturally related to a stationary homogenization problem, which is of independent interest. Here we state the stationary homogenization result in a simplified version, obtained by neglecting the role of the recovery variable w^ε.

We first introduce the family of convex functionals defined on $H^1(\Omega^\varepsilon_i) \times H^1(\Omega^\varepsilon_e)$,

$$\mathscr{F}^\varepsilon(\underline{u}) := \frac{h}{2} b^\varepsilon(v - v^\varepsilon_0) + \frac{1}{2} \underline{a}^\varepsilon(\underline{u}) + \phi^\varepsilon(v), \quad v := u_i - u_e \quad \text{on } \Gamma^\varepsilon, \tag{1.28}$$

with $h \geq 0$ a given constant, $v^\varepsilon_0 \in L^2(\Gamma^\varepsilon)$, $b^\varepsilon, \underline{a}^\varepsilon, \phi^\varepsilon$ defined in (1.12), and the limit functional defined in $\big(H^1(\Omega)\big)^2$

$$\mathscr{F}(\underline{u}) := \frac{h}{2} b(v - v_0) + \frac{1}{2} \underline{a}(\underline{u}) + \phi(v), \quad v := u_i - u_e \quad \text{in } \Omega \tag{1.29}$$

for a given $v_0 \in L^2(\Omega)$ and $b, \underline{a}, \phi$ introduced in (1.20).

THEOREM 1.6 (coercivity and Γ-convergence). *Let us suppose that* $v^\varepsilon_0 \in L^2(\Gamma^\varepsilon)$ *converge to* $v_0 \in L^2(\Omega)$ *as* $\varepsilon \downarrow 0$ *according to* (1.16) *and satisfy*

$$\lim_{\varepsilon\downarrow 0} b^\varepsilon(v^\varepsilon_0) = b(v_0), \qquad \limsup_{\varepsilon\downarrow 0} \Big(j^\varepsilon(v^\varepsilon_0) + \phi^\varepsilon(v^\varepsilon_0) \Big) < +\infty. \tag{1.30}$$

Then the following properties hold:

(a) *Compactness. If* $\underline{u}^\varepsilon = (u^\varepsilon_i, u^\varepsilon_e) \in H^1(\Omega^\varepsilon_i) \times H^1(\Omega^\varepsilon_e)$ *satisfies*

$$\liminf_{\varepsilon\downarrow 0} \mathscr{F}^\varepsilon(\underline{u}^\varepsilon) < +\infty \quad \textit{and} \quad \int_{\Omega^\varepsilon_e \cap \Omega_0} u^\varepsilon_e(x)\, dx = 0 \tag{1.31}$$

then there exists $\underline{u} \in H^1(\Omega) \times H^1(\Omega)$ *satisfying*

$$\int_{\Omega_0} u_e(x)\,dx = 0 \tag{1.32}$$

and a vanishing subsequence ε_n *such that* $\underline{u}^{\varepsilon_n}$ *converges to* u *according to Definition* 1.1 *as* $n \to \infty$.

(b) lim inf *inequality. For every family* $\underline{u}^{\varepsilon_n}$ *converging to* $\underline{u}$ *according to Definition* 1.1,

$$\liminf_{n\to\infty} \mathscr{F}^{\varepsilon_n}(\underline{u}^{\varepsilon_n}) \geq \mathscr{F}(\underline{u}). \tag{1.33}$$

(c) lim sup *inequality. For every* $\underline{u} = (u_i, u_e) \in \big(H^1(\Omega)\big)^2$ *satisfying* (1.32) *there exist* $\underline{u}^\varepsilon \in H^1(\Omega_i^\varepsilon) \times H^1(\Omega_e^\varepsilon)$ *converging to* $\underline{u}$ *as in Definition* 1.1 *with* $v^\varepsilon = u_i^\varepsilon - u_e^\varepsilon$ *converging to* $v = u_i - u_e$ *as in* (1.16) *such that*

$$\int_{\Omega_e^\varepsilon\cap\Omega_0} u_e^\varepsilon(x)\,dx = 0, \qquad \limsup_{\varepsilon\downarrow 0} \mathscr{F}^\varepsilon(\underline{u}^\varepsilon) \leq \mathscr{F}(\underline{u}). \tag{1.34}$$

Whenever a suitable weak (and metrizable) topology is introduced in the spaces of (signed) Radon measures (we postpone the discussion of this point to section 3), the above result shows that $\mathscr{F}$ is the Γ-limit of the coercive family of functionals $\mathscr{F}^\varepsilon$.

Well-known results on Γ-convergence [19, Cor. 2.4] and a standard computation of the first variation of the functionals $\mathscr{F}^\varepsilon, \mathscr{F}$ [16] immediately yield the following.

Corollary 1.7 (homogenization of the stationary problems). *Under the assumptions of Theorem* 1.6, *each functional* $\mathscr{F}^\varepsilon$ *admits a unique minimizer* $\underline{u}^\varepsilon = (u_i^\varepsilon, u_e^\varepsilon) \in H^1(\Omega_i^\varepsilon) \times H^1(\Omega_e^\varepsilon)$ *satisfying* (1.31) *with* $v^\varepsilon = u_i^\varepsilon - u_e^\varepsilon \in H^{1/2}(\Gamma^\varepsilon)$ *and characterized by the system*

$$-\operatorname{div}\big(\sigma_{i,e}^\varepsilon \nabla u_{i,e}^\varepsilon\big) = 0 \qquad \text{in } \Omega_{i,e}^\varepsilon, \tag{1.35a}$$

$$\left.\begin{matrix} -\sigma_i^\varepsilon \nabla u_i^\varepsilon \cdot \nu_i^\varepsilon \\ \sigma_e^\varepsilon \nabla u_e^\varepsilon \cdot \nu_e^\varepsilon \end{matrix}\right\} = I_m^\varepsilon \qquad \text{on } \Gamma^\varepsilon, \tag{1.35b}$$

$$\varepsilon\Big((h+\lambda_F)v^\varepsilon + F(v^\varepsilon)\Big) = I_m^\varepsilon + \varepsilon h v_0^\varepsilon \qquad \text{on } \Gamma^\varepsilon, \tag{1.35c}$$

$$\sigma_{i,e}^\varepsilon \nabla u_{i,e}^\varepsilon \cdot \nu_{i,e} = 0 \qquad \text{on } \partial\Omega_{i,e}^\varepsilon \setminus \Gamma^\varepsilon, \tag{1.35d}$$

$$\int_{\Omega_e^\varepsilon\cap\Omega_0} u_e^\varepsilon(x)\,dx = 0. \tag{1.35e}$$

$(u_{i,e}^\varepsilon, v^\varepsilon)$ *converge to* $(u_{i,e}, v) \in \big(H^1(\Omega)\big)^3$ *as* $\varepsilon \downarrow 0$ *according to Definition* 1.1; $\underline{u} = (u_i, u_e)$ *with* $v = u_i - u_e$ *is the (unique) minimizer of* $\mathscr{F}$ *satisfying* (1.32) *and it is characterized by the system*

$$\left.\begin{matrix} \operatorname{div}(M_i \nabla u_i) \\ -\operatorname{div}(M_e \nabla u_e) \end{matrix}\right\} = I_m \qquad \text{in } \Omega, \tag{1.36a}$$

$$\beta\Big((h+\lambda_F)v + F(v)\Big) = I_m + \beta h v_0 \qquad \text{in } \Omega, \tag{1.36b}$$

$$M_{i,e}\nabla u_{i,e} \cdot \nu_{i,e} = 0 \qquad \text{on } \partial\Omega, \tag{1.36c}$$

$$\int_{\Omega_0} u_e(x)\,dx = 0. \tag{1.36d}$$

Plan of the paper. The paper is divided into two parts: the first part (sections 2 and 3) is devoted to the stationary homogenization result stated in Theorem 1.6 and can be read independently of the remaining sections. The next section contains some preliminary technical results related to the extension problem of functions defined on Γ^ε and $\Omega^\varepsilon_{i,e}$; we also present a natural generalization of the Riemann–Lebesgue lemma and some properties of the notion of convergence introduced in Definition 1.1; finally we discuss some applications to lower semicontinuity and approximation results for integral functionals defined on Γ^ε.

Section 3 will conclude the proof of Theorem 1.6 and its corollary. Here we combine very general results on Γ-convergence of noncoercive functionals [12] (related to elliptic problems with Neumann boundary conditions) with the more recent extension techniques of [1]. Our main contribution is to extend this framework to the homogenization of integral functionals defined on the interface between the two ε-domains.

The second part of the paper is more specifically devoted to the evolutive systems $^\varepsilon$, . The well posedness of their variational formulation and some preliminary accessory results are collected in section 4.

In section 5 we outline and carry out the main steps of the proof of Theorems 1.3 and 1.4. Here we adopt the point of view of the "Minimizing movement" approach to evolution equations suggested by De Giorgi in [18]: we perform an auxiliary "semi-implicit time discretization" of the microscopic reaction-diffusion system, which consists of a recursive family of variational problems depending on the step size $\tau > 0$. The final macroscopic model will thus result in two limit procedures: the first one in the time discretization, as the time step τ goes to 0 keeping ε fixed, and the second in the homogenization process as $\varepsilon \downarrow 0$.

Uniform approximation estimates for the discretized problem (which are strictly related to the *convexity* of the underlying functionals) allow us to "invert" the order of the limits: we can therefore pass first to the limit as $\varepsilon \downarrow 0$ keeping τ fixed, and in this way we obtain a family of *homogenized discrete macroscopic problems*; a final limit as $\tau \downarrow 0$ recovers the continuous form of the macroscopic problem. By this approach, the homogenization of the time-dependent problem $^\varepsilon$ is reduced to the homogenization of a finite sequence of stationary problems of elliptic type, which exhibit (up to lower order perturbation terms) the same structure we studied in section 3.

Applications of Γ-convergence to evolution problems are well known for gradient flows of *convex* (or λ-*convex*) functionals in Hilbert spaces [6]: in that case the uniform convergence of the induced evolution semigroups can be deduced from the Γ-convergence and Mosco-convergence [27, 19] of the underlying functionals.

In our case things are more difficult due to the degeneration of the parabolic structure of our systems (as was discussed in [16]) and to the lack of a "fixed" (i.e., independent of ε) Hilbert space, where the evolution can be settled.

Therefore, a general abstract result for studying the convergence of the present problem seems to be missing. Nevertheless, we tried to develop a general procedure (uniform discretization estimates and Γ-convergence of the discretized variational problems) to attack this kind of λ-convex but "degenerate" evolution problems: even if our arguments could have been presented in a more compact (but maybe more obscure) form, we decided to clarify their structure as much as possible, hoping that a better understanding of the main ideas of this approach could also be helpful for other applications in different contexts.

In section 6 we briefly sketch the rigorous derivation of the error estimates for the semi-implicit discretization of Problems $^\varepsilon$ and : here we adapt to our setting the technique introduced in [29] (but see also [7, 32, 33]) to obtain optimal a priori estimates for gradient flows of convex functionals in Hilbert spaces.

Such estimates could have also been derived by standard perturbation arguments from the general results of [9] for a fully implicit discretization scheme; here we chose a direct approach to have precise control of the various constants involved (which should be *independent of* ε) and to keep the presentation simpler and almost self-contained.

The main advantage of the semi-implicit discretization (instead of an implicit one) lies in the variational structure of the problems, which should be solved at each time step: in fact, they are associated with the minimum of convex functionals.

For the sake of completeness, in the appendix we briefly recall the derivation of the model at the cellular level from well established physical laws and introduce its dimensionless form.

2. Notation, extension results, and related convergence properties.

2.1. Vector notation, function spaces, and bilinear forms. In order to write the micro- and macroscopic problems in a compact form, we introduce a vector notation, which will also be useful for dealing with the evolution systems; thus

$$\underline{u}^{\,\varepsilon}, \underline{u} \quad \text{will denote the vectors } (u_i^\varepsilon, u_e^\varepsilon),\ (u_i, u_e), \tag{2.1}$$

and we will use underlined letters (as $\underline{u}, \underline{\phi}, \underline{\mathcal{V}}, \dots$) for vectors, functions, and spaces involving intraextracellular couples. We set

$$\begin{aligned} &\mathcal{H}^\varepsilon := L^2(\Gamma^\varepsilon), \underline{\mathcal{V}}^\varepsilon := \Big\{ \underline{u} \in H^1(\Omega_i^\varepsilon) \times H^1(\Omega_e^\varepsilon) : v = u_i^\varepsilon - u_e^\varepsilon \in L^2(\Gamma^\varepsilon) \Big\}, \\ &\mathcal{H} := L^2(\Omega),\ \underline{\mathcal{V}} := \ H^1(\Omega) \times H^1(\Omega), \quad \underline{\mathcal{V}}_{loc} := H^1_{loc}(\Omega) \times H^1_{loc}(\Omega), \end{aligned} \tag{2.2}$$

together with their closed subspaces

$$\underline{\mathcal{V}}_0^\varepsilon := \left\{ \underline{u}^\varepsilon \in \underline{\mathcal{V}}^\varepsilon : \int_{\Omega_e^\varepsilon \cap \Omega_0} u_e^\varepsilon(x)\, dx = 0 \right\}, \quad \underline{\mathcal{V}}_0 := \left\{ \underline{u} \in \underline{\mathcal{V}} : \int_{\Omega_0} u_e(x)\, dx = 0 \right\}. \tag{2.3}$$

Remark 2.1. Since $\Omega_{i,e}^\varepsilon$ could have a very irregular boundary, we do not know if the traces $u_{i,e}^\varepsilon|_{\Gamma^\varepsilon}$ of $u_{i,e}^\varepsilon \in H^1(\Omega_{i,e}^\varepsilon)$ belong to $H^{1/2}(\Gamma^\varepsilon) \subset L^2(\Gamma^\varepsilon)$: a priori we only know $u_{i,e}^\varepsilon|_{\Gamma^\varepsilon} \in H^{1/2}_{loc}(\Gamma^\varepsilon) \subset L^2_{loc}(\Gamma^\varepsilon)$. Therefore, the integrability condition on $v = u_i^\varepsilon - u_e^\varepsilon$ on Γ^ε imposed in the definition (2.2) of $\underline{\mathcal{V}}_0^\varepsilon$ is not redundant.

We now introduce some continuous and symmetric bilinear forms on $\underline{\mathcal{V}}^\varepsilon, \mathcal{H}^\varepsilon, \underline{\mathcal{V}}, \mathcal{H}$ which will play a crucial role in the following. Recalling (1.12a,b,c) and (1.20a,b,c), we set

$$\underline{a}^\varepsilon(\underline{u}, \underline{\hat{u}}) := \sum_{i,e} \int_{\Omega_{i,e}^\varepsilon} \sigma_{i,e}^\varepsilon \nabla u_{i,e} \cdot \nabla \hat{u}_{i,e}\, dx \quad \forall\, \underline{u}, \underline{\hat{u}} \in \underline{\mathcal{V}}^\varepsilon, \tag{2.4}$$

$$\underline{a}(\underline{u}, \underline{\hat{u}}) := \sum_{i,e} \int_{\Omega} M_{i,e} \nabla u_{i,e} \cdot \nabla \hat{u}_{i,e}\, dx \quad \forall\, \underline{u}, \underline{\hat{u}} \in \underline{\mathcal{V}}, \tag{2.5}$$

and we introduce the scalar products on $\mathcal{H}^\varepsilon, \mathcal{H}$,

$$\begin{aligned} b^\varepsilon(w, \hat{w}) &:= \varepsilon \int_{\Gamma^\varepsilon} w\, \hat{w}\, d\mathscr{H}^{d-1} \quad \forall\, w, \hat{w} \in \mathcal{H}^\varepsilon, \\ b(w, \hat{w}) &:= \beta \int_{\Omega} w\, \hat{w}\, dx \quad \forall\, w, \hat{w} \in \mathcal{H}. \end{aligned} \tag{2.6}$$

We also set

$$\underline{b}^\varepsilon(\underline{u},\hat{\underline{u}}) := \varepsilon \int_{\Gamma^\varepsilon} (u_i - u_e)(\hat{u}_i - \hat{u}_e)\, d\mathscr{H}^{d-1} = b^\varepsilon(\underline{B}^\varepsilon \underline{u}^\varepsilon, \underline{B}^\varepsilon \hat{\underline{u}}^\varepsilon) \quad \forall\, \underline{u},\hat{\underline{u}} \in \underline{\mathcal{V}}^\varepsilon, \tag{2.7}$$

$$\underline{b}(\underline{u},\hat{\underline{u}}) := \beta \int_{\Omega} (u_i - u_e)(\hat{u}_i - \hat{u}_e)\, dx = b(\underline{B}\underline{u}, \underline{B}\hat{\underline{u}}) \quad \forall\, \underline{u},\hat{\underline{u}} \in \underline{\mathcal{V}}, \tag{2.8}$$

$$\underline{\phi}^\varepsilon(\underline{u}) := \varepsilon \int_{\Gamma^\varepsilon} \varphi(u_i - u_e)\, d\mathscr{H}^{d-1} = \phi^\varepsilon(\underline{B}^\varepsilon \underline{u}) \quad \forall\, \underline{u} \in \underline{\mathcal{V}}^\varepsilon, \tag{2.9}$$

$$\underline{\phi}(\underline{u}) := \beta \int_{\Omega} \varphi(u_i - u_e)\, dx = \phi(\underline{B}\underline{u}) \quad \forall\, \underline{u} \in \underline{\mathcal{V}}, \tag{2.10}$$

where $\underline{B}^\varepsilon, \underline{B}$ are the linear continuous operators

$$\begin{aligned} &\underline{B}^\varepsilon : \underline{\mathcal{V}}^\varepsilon \to \mathcal{H}^\varepsilon, \underline{B}^\varepsilon \underline{u} := u_i|_{\Gamma^\varepsilon} - u_e|_{\Gamma^\varepsilon} \forall\, \underline{u} \in \underline{\mathcal{V}}_0^\varepsilon,\\ &\underline{B} : \underline{\mathcal{V}} \to \mathcal{H}, \quad \underline{B}\underline{u} := u_i - u_e \qquad \forall\, \underline{u} \in \underline{\mathcal{V}}_0. \end{aligned} \tag{2.11}$$

It is easy to check that the bilinear forms $\underline{a}^\varepsilon(\cdot,\cdot) + \underline{b}^\varepsilon(\cdot,\cdot)$ and $\underline{a}(\cdot,\cdot) + \underline{b}(\cdot,\cdot)$ are scalar products on $\underline{\mathcal{V}}_0^\varepsilon$ and $\underline{\mathcal{V}}_0$; they induce on $\underline{\mathcal{V}}_0^\varepsilon$ and $\underline{\mathcal{V}}_0$ the usual topology as closed subspaces of $\underline{\mathcal{V}}^\varepsilon$ and $\underline{\mathcal{V}}$, respectively. As in (1.12a,b,c) and (1.20a,b,c), we adopt the convention of writing the associated quadratic forms as

$$\underline{b}^\varepsilon(\underline{u}) := \underline{b}^\varepsilon(\underline{u},\underline{u}), \quad \underline{a}^\varepsilon(\underline{u}) := \underline{a}^\varepsilon(\underline{u},\underline{u}), \quad \underline{b}(\underline{u}) := \underline{b}(\underline{u},\underline{u}), \quad \underline{a}(\underline{u}) := \underline{a}(\underline{u},\underline{u}). \tag{2.12}$$

2.2. Uniform bounds for extension operators. Let us now discuss some extension results for functions defined in $\Omega^\varepsilon_{i,e}, \Gamma^\varepsilon$, which will be applied to the notion of convergence introduced in Definition 1.1.

DEFINITION 2.2 (extensions). *We say that* $\breve{\underline{u}}^\varepsilon \in \underline{\mathcal{V}}_{loc} = H^1_{loc}(\Omega) \times H^1_{loc}(\Omega)$ *is an* extension *of* $\underline{u}^\varepsilon \in \underline{\mathcal{V}}^\varepsilon$ *if*

$$\breve{u}^\varepsilon_{i,e}|_{\Omega^\varepsilon_{i,e}} = u^\varepsilon_{i,e}. \tag{2.13}$$

Analogously, we say that $\breve{w}^\varepsilon \in H^1_{loc}(\Omega)$ *is an* extension *of* $w^\varepsilon \in H^{1/2}_{loc}(\Gamma^\varepsilon)$ *if*

$$\breve{w}^\varepsilon|_{\Gamma^\varepsilon} = w^\varepsilon \quad \textit{in the sense of traces.} \tag{2.14}$$

One of the technical difficulties in the present setting is to find suitable extension operators $T^\varepsilon_{i,e}$ of functions defined only on $\Omega^\varepsilon_{i,e}$ to the whole Ω which preserve uniform bounds of the L^2 and H^1 norms. Due to the possible irregular behavior of the boundary of $\Omega_{i,e}$, only *local bounds* are available.

The following result proved by Acerbi et al. [1] is almost optimal. The crucial assumption, here is that the sets $E_{i,e}$ are Lipschitz and connected; we denote by $\Omega(\delta)$, $\delta \ge 0$, the open subset of Ω defined by

$$\Omega(\delta) := \big\{x \in \Omega : d(x, \mathbb{R}^d \setminus \Omega) > \delta\big\}. \tag{2.15}$$

THEOREM 2.3 (see [1]). *There exists linear and continuous extension operators* $T^\varepsilon_{i,e} : H^1(\Omega^\varepsilon_{i,e}) \to H^1_{loc}(\Omega)$ *and three constants* $k_0, h_0, h_1 > 0$ *independent of* $\varepsilon > 0$ *and* Ω*, such that for every* $u \in H^1(\Omega^\varepsilon_{i,e})$ *we have*

$$T^\varepsilon_{i,e} u = u \qquad \textit{a.e. in } \Omega^\varepsilon_{i,e}, \tag{2.16}$$

$$\int_{\Omega(\varepsilon k_0)} |T^\varepsilon_{i,e} u|^2\, dx \le h_0 \int_{\Omega^\varepsilon_{i,e}} |u|^2\, dx, \tag{2.17}$$

$$\int_{\Omega(\varepsilon k_0)} |\nabla T^\varepsilon_{i,e} u|^2\, dx \le h_1 \int_{\Omega^\varepsilon_{i,e}} |\nabla u|^2\, dx. \tag{2.18}$$

As usual we set $\underline{T}^\varepsilon := \big(T_i^\varepsilon, T_e^\varepsilon\big) : \underline{\mathcal{V}}^\varepsilon \to \underline{\mathcal{V}}_{loc}$.

Remark 2.4. In general, it is not possible to construct a family of extension operators $T_{i,e}^\varepsilon : H^1(\Omega_{i,e}^\varepsilon) \to H^1(\Omega)$ satisfying (2.16), (2.17), (2.18) with $\Omega(\varepsilon k_0)$ replaced by Ω, since we do not have any control of the behavior of E^ε near $\partial\Omega$. For more details and an explicit counterexample we refer to [1].

2.3. Generalized Riemann–Lebesgue lemma. Let us first recall a well-known version of the Riemann–Lebesgue lemma.

LEMMA 2.5 (generalized Riemann–Lebesgue lemma). *Let A be a bounded open subset of $\mathbb{R}^d$ with $\mathscr{L}^d(\partial A) = 0$ and let $\zeta \in C^0(\overline{A})$. Then*

$$\lim_{\varepsilon\downarrow 0} \int_{A\cap E_{i,e}^\varepsilon} \zeta(x)\,dx = \beta_{i,e} \int_A \zeta(x)\,dx, \tag{2.19}$$

$$\lim_{\varepsilon\downarrow 0} \varepsilon \int_{A\cap\Gamma^\varepsilon} \zeta(x)\,d\mathscr{H}^{d-1}(x) = \beta \int_A \zeta(x)\,dx, \tag{2.20}$$

where the coefficients $\beta_{i,e}, \beta$ are defined in (1.14).

Remark 2.6 (weak* convergence in L^∞). Limit (2.19) shows that the characteristic functions $\chi_{A\cap E_{i,e}^\varepsilon}$ of $A\cap E_{i,e}^\varepsilon$ are converging to $\beta_{i,e}\,\chi_A$ in the sense of distributions as $\varepsilon\downarrow 0$; since they are also uniformly bounded in $L^\infty(\Omega)$, an obvious weak*-compactness argument shows that

$$\chi_{A\cap E_{i,e}^\varepsilon} \rightharpoonup^* \beta_{i,e}\,\chi_A \quad \text{in } L^\infty(\Omega) \quad \text{as } \varepsilon\downarrow 0. \tag{2.21}$$

Remark 2.7 (weak* convergence in the space of measures). Lemma 2.5 also shows that the measures $\lambda_{i,e}^\varepsilon, \lambda^\varepsilon$ defined by (1.17) converge to $\lambda_{i,e}, \lambda$, respectively, as $\varepsilon\downarrow 0$ in the weak* topology of the space of finite Radon measures on Ω.

The next result reinforces Lemma 2.5 and how weak convergence in $H^1_{loc}(\Omega)$ implies the convergence in the sense of Definition 1.1. Since we will deal with functionals depending on the continuous parameter $\varepsilon > 0$ or on the discrete values of a suitable decreasing infinitesimal sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$, for notational convenience we will treat both cases in the same way by considering a general nonempty set Λ of real numbers such that

$$\Lambda \subset (0, +\infty), \quad \inf\Lambda = 0. \tag{2.22}$$

Expressions like $\lim_{\varepsilon\downarrow 0,\varepsilon\in\Lambda}$, $\liminf_{\varepsilon\downarrow 0,\varepsilon\in\Lambda}$, etc., have an obvious meaning as limits for ε going to 0 in Λ. Of course, when Λ contains an open interval $(0,\delta)$, $\delta > 0$, we will use the usual notation $\lim_{\varepsilon\downarrow 0}$.

PROPOSITION 2.8 (weak H^1_{loc}-convergence yields convergence of Definition 1.1). *Let us suppose that z^ε weakly converge to z in $H^1_{loc}(\Omega)$ for $\varepsilon\downarrow 0, \varepsilon\in\Lambda$. Then for every continuous function $\zeta \in C^0_c(\Omega)$ we have*

$$\lim_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} \int_{\Omega_{i,e}^\varepsilon} z^\varepsilon(x)\zeta(x)\,dx = \beta_{i,e} \int_\Omega z(x)\zeta(x)\,dx, \tag{2.23}$$

$$\lim_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} \varepsilon \int_{\Gamma^\varepsilon} z^\varepsilon(x)\zeta(x)\,d\mathscr{H}^{d-1}(x) = \beta \int_A z(x)\zeta(x)\,dx. \tag{2.24}$$

In particular $z^\varepsilon|_{\Gamma^\varepsilon}$ and $z^\varepsilon|_{\Omega_{i,e}^\varepsilon}$ converge to z according to Definition 1.1.

Proof. By the Rellich compactness theorem, we know that $z^\varepsilon \to z$ strongly in $L^2_{loc}(\Omega)$ as $\varepsilon\downarrow 0, \varepsilon\in\Lambda$. Equation (2.23) thus follows directly from Remark 2.6 since

$$\chi_{\Omega_{i,e}^\varepsilon} \rightharpoonup^* \beta_{i,e} \quad \text{in } L^\infty(\Omega), \quad \text{and therefore} \quad z^\varepsilon\chi_{\Omega_{i,e}^\varepsilon} \rightharpoonup \beta_{i,e} z \ \text{ weakly in } L^2_{loc}(\Omega).$$

In order to prove (2.24), we first observe that for ε sufficiently small, we can find pluricellular regions (recall (1.2))

$$R^\varepsilon = \bigcup_{m=1}^{M^\varepsilon} \varepsilon\left(Y + \sum_{k=1}^{d} j^\varepsilon_{m,k}\ {}_k\right) \quad \text{for some} \quad {}^{\varepsilon}_{m} = (j^\varepsilon_{m,1}, \cdots, j^\varepsilon_{m,d}) \in \mathbb{Z}^d,$$

and a regular open set A such that

$$\operatorname{supp}\zeta \subset R^\varepsilon \subset A \subset\subset \Omega.$$

Poincaré inequality and a rescaling argument easily yield

$$\varepsilon\mathscr{H}^{d-1}(R^\varepsilon \cap \varepsilon\Gamma) = \beta\mathscr{L}^d(R^\varepsilon), \quad \varepsilon^2\|z\|^2_{L^1(R^\varepsilon\cap\varepsilon\Gamma)} \le \varepsilon\beta\mathscr{L}^d(A)\|z\|^2_{L^2(R^\varepsilon\cap\varepsilon\Gamma)}, \tag{2.25}$$

$$\varepsilon\|z\|^2_{L^2(R^\varepsilon\cap\varepsilon\Gamma)} \le c_1\left(\|z\|^2_{L^2(R^\varepsilon_{i,e})} + \varepsilon^2\|\nabla z\|^2_{L^2(R^\varepsilon_{i,e})}\right). \tag{2.26}$$

Then, if $S := \sup_{x\in\Omega}|\zeta(x)|$, we have

$$\begin{aligned}\left|\varepsilon\int_{\Gamma^\varepsilon}\big(z^\varepsilon(x) - z(x)\big)\zeta(x)\,d\mathscr{H}^{d-1}(x)\right| &\le \varepsilon S\|z^\varepsilon - z\|_{L^1(R^\varepsilon\cap\varepsilon\Gamma)}\\ &\le C\Big(\|z^\varepsilon - z\|_{L^2(A)} + \varepsilon\|\nabla z^\varepsilon - \nabla z\|_{L^2(A)}\Big),\end{aligned} \tag{2.27}$$

where $C := S\sqrt{c_1\beta\mathscr{L}^d(A)}$; since z^ε is bounded in $H^1(A)$ and converges to z in $L^2(A)$, (2.27) vanishes as $\varepsilon\downarrow 0$. Thus we simply have to prove (2.24) for $z^\varepsilon \equiv z$.

Choosing now another arbitrary function $\eta \in H^1(A)\cap C^0(\overline{A})$ and taking into account Lemma 2.5, we get

$$\begin{aligned}&\limsup_{\varepsilon\downarrow 0,\varepsilon\in\Lambda}\left|\varepsilon\int_{\Gamma^\varepsilon} z(x)\zeta(x)\,d\mathscr{H}^{d-1}(x) - \beta\int_\Omega z(x)\zeta(x)\,dx\right|\\ &\qquad\le \limsup_{\varepsilon\downarrow 0,\varepsilon\in\Lambda}\left|\varepsilon\int_{\Gamma^\varepsilon}\big(z(x)-\eta(x)\big)\zeta(x)\,d\mathscr{H}^{d-1}(x)\right| + \left|\beta\int_\Omega\big(z(x)-\eta(x)\big)\zeta(x)\,dx\right|\\ &\qquad\le (1+\beta)Z\Big(\varepsilon\|z-\eta\|_{L^1(R^\varepsilon\cap\varepsilon\Gamma)} + \|z-\eta\|_{L^1(R^\varepsilon)}\Big) \le C'\|z-\eta\|_{H^1(A)}\end{aligned}$$

for a constant C' independent of z and η. Being η arbitrary and A regular, a standard density result yields (2.24). □

COROLLARY 2.9. *Suppose that $w^\varepsilon \in H^{1/2}_{loc}(\Gamma^\varepsilon)$ converge to $w \in L^1_{loc}(\Omega)$ according to Definition* 1.1 *and let $\breve{w}^\varepsilon \in H^1_{loc}(\Omega)$ be an extension of w^ε which is uniformly bounded in $H^1(A)$ for every open subset $A \subset\subset \Omega$. Then $w \in H^1_{loc}(\Omega)$ and $\breve{w}^\varepsilon \rightharpoonup w$ in $H^1_{loc}(\Omega)$.*

Proof. By Proposition 2.8 w is the unique limit point of any weakly convergent subsequence of $\breve{w}^\varepsilon$ in $H^1_{loc}(\Omega)$. □

2.4. Compactness properties. Recall that in (1.11) we introduced an open subset $\Omega_0 \subset\subset \Omega$ which induces closed subspaces of $H^1(\Omega^\varepsilon_e)$ and $H^1(\Omega)$ through the integral conditions (2.3). By Lemma 2.5 with $A := \Omega_0$ and $\zeta \equiv 1$ we have that

$$\lim_{\varepsilon\downarrow 0}\mathscr{L}^d(\Omega_0\cap\varepsilon E_e) = \beta_e\mathscr{L}^d(\Omega_0), \quad \lim_{\varepsilon\downarrow 0}\varepsilon\mathscr{H}^{d-1}(\Omega_0\cap\varepsilon\Gamma) = \beta\mathscr{L}^d(\Omega_0), \tag{2.28}$$

so that we can always assume that

$$\mathscr{L}^d(\Omega_0 \cap \varepsilon E_e) \geq \frac{\beta_e}{2}\mathscr{L}^d(\Omega_0), \quad \varepsilon\mathscr{H}^{d-1}(\Omega_0 \cap \varepsilon\Gamma) \geq \frac{\beta}{2}\mathscr{L}^d(\Omega_0) \quad \forall\, \varepsilon \in \Lambda. \tag{2.29}$$

LEMMA 2.10 (uniform local H^1 bounds and compactness). *Let $\underline{u}^\varepsilon \in \underline{\mathcal{V}}_0^\varepsilon$, $\varepsilon \in \Lambda$, be a family of functions satisfying*

$$\sup_{\varepsilon\in\Lambda}\Big(\underline{a}^\varepsilon(\underline{u}^\varepsilon) + \underline{b}^\varepsilon(\underline{u}^\varepsilon)\Big) = S < +\infty. \tag{2.30}$$

Then for every open subset $A \subset\subset \Omega$ we have

$$\sup_{\varepsilon\in\Lambda}\|\underline{T}^\varepsilon\underline{u}^\varepsilon\|_{H^1(A)} < +\infty. \tag{2.31}$$

In particular, there exists an infinitesimal subsequence $\Lambda' = (\varepsilon_j)_{j\in\mathbb{N}} \subset \Lambda$ and a limit function $\underline{u} \in \underline{\mathcal{V}}_0$ such that

$$\lim_{\varepsilon\downarrow 0,\varepsilon\in\Lambda'} \underline{T}^\varepsilon\underline{u}^\varepsilon = \underline{u} \quad \textit{weakly in } \underline{\mathcal{V}}_{loc} = H^1_{loc}(\Omega)\times H^1_{loc}(\Omega). \tag{2.32}$$

Proof. Let $\underline{u}^\varepsilon = (u_i^\varepsilon, u_e^\varepsilon) \in \underline{\mathcal{V}}_0^\varepsilon, \underline{\check{u}}^\varepsilon := \underline{T}^\varepsilon\underline{u}^\varepsilon \in \underline{\mathcal{V}}_{loc}$, and let us choose $\delta > 0$ sufficiently small such that (see (2.15) and, e.g., [34])

$$A \subset \Omega(\delta), \quad \Omega_0 \subset \Omega(2\delta), \quad \Omega(\delta) \text{ is Lipschitz.} \tag{2.33}$$

We can suppose that $(k_0 + \ell)\varepsilon < \delta$; by using the properties (2.16), (2.17), (2.18) of the extension operators $T^\varepsilon_{i,e}$, we get

$$\|\nabla\check{u}^\varepsilon_{i,e}\|^2_{L^2(A)} \leq \|\nabla\check{u}^\varepsilon_{i,e}\|^2_{L^2(\Omega(\delta))} \leq h_1\|\nabla u^\varepsilon_{i,e}\|^2_{L^2(\Omega^\varepsilon_{i,e})} \leq h_1\sigma^{-1}\underline{a}^\varepsilon(\underline{u}^\varepsilon). \tag{2.34}$$

Poincaré inequality yields constants $c^\varepsilon_{i,e}$ (depending on $u^\varepsilon_{i,e}$) and c_P (depending only on Ω) satisfying

$$\|\check{u}^\varepsilon_{i,e} - c^\varepsilon_{i,e}\|_{L^2(\Omega(\delta))} \leq c_P\|\nabla\check{u}^\varepsilon_{i,e}\|_{L^2(\Omega(\delta))}. \tag{2.35}$$

Setting $\Omega^\varepsilon_{0,e} := \Omega_0 \cap \varepsilon E_e$, by the properties of the extension operator we know that

$$\int_{\Omega_0^\varepsilon}\check{u}^\varepsilon_e\,dx = 0, \tag{2.36}$$

so that

$$|c^\varepsilon_e|\,\mathscr{L}^d(\Omega^\varepsilon_{0,e}) \leq \int_{\Omega^\varepsilon_{0,e}}|c^\varepsilon_e - \check{u}^\varepsilon_e|\,dx \leq \mathscr{L}^d(\Omega^\varepsilon_{0,e})^{1/2}\|\check{u}^\varepsilon_e - c^\varepsilon_e\|_{L^2(\Omega^\varepsilon_{0,e})}$$

and therefore by (2.29) we have

$$|c^\varepsilon_e| \leq \Big(\frac{\beta_e}{2}\mathscr{L}^d(\Omega_0)\Big)^{-1/2}\|\check{u}^\varepsilon_e - c^\varepsilon_e\|_{L^2(\Omega(\delta))},$$

which, together with (2.35) and (2.34), shows that c^ε_e is uniformly bounded with respect to ε. In order to get an analogous bound for c^ε_i we will use the estimate

$$\varepsilon\int_{\Gamma^\varepsilon}|u^\varepsilon_i - u^\varepsilon_e|^2\,d\mathscr{H}^{d-1} = \underline{b}^\varepsilon(\underline{u}^\varepsilon) \leq S, \tag{2.37}$$

observing that by (2.33) and (2.26),

$$\begin{aligned}\varepsilon\int_{\Omega_0\cap\varepsilon\Gamma}|u^\varepsilon_{i,e}-c^\varepsilon_{i,e}|^2\,d\mathscr{H}^{d-1}\le& c_1\int_{\Omega(\delta)\cap\varepsilon E_{i,e}}\Big((u^\varepsilon_{i,e}-c^\varepsilon_{i,e})^2+\varepsilon^2\left|\nabla u^\varepsilon_{i,e}\right|^2\Big)\,dx\\ \le& c_1({c_P}^2+\varepsilon^2)\|\nabla\breve{u}^\varepsilon_{i,e}\|^2_{L^2(\Omega(\delta))}\le c_2:=c_1(c_P+\varepsilon^2)h_1\sigma^{-1}S.\end{aligned}$$

Thus we get

$$\begin{aligned}\frac{\beta}{2}\mathscr{L}^d(\Omega_0)|c^\varepsilon_i-c^\varepsilon_e|^2&\le\varepsilon\int_{\Omega_0\cap\varepsilon\Gamma}|c^\varepsilon_i-c^\varepsilon_e|^2\,d\mathscr{H}^{d-1}\\&\le\varepsilon\int_{\Omega_0\cap\varepsilon\Gamma}\left|c^\varepsilon_i-c^\varepsilon_e-(u^\varepsilon_i-u^\varepsilon_e)+(u^\varepsilon_i-u^\varepsilon_e)\right|^2d\mathscr{H}^{d-1}\\&\le3\varepsilon\int_{\Gamma^\varepsilon}|u^\varepsilon_i-u^\varepsilon_e|^2\,d\mathscr{H}^{d-1}+6c_2=3\underline{b}^\varepsilon(\underline{u}^\varepsilon)+6c_2\le3S+6c_2.\end{aligned}$$

It follows that also c^ε_i is uniformly bounded with respect to ε, so that $\breve{u}^\varepsilon_{i,e}$ are bounded in $H^1(A)$.

A standard diagonal argument yields (2.32) for some $\underline{u}\in\underline{\mathcal{V}}$; the fact that $\underline{u}$ belongs to $\underline{\mathcal{V}}_0$, too, follows from Proposition 2.8. ∎

COROLLARY 2.11. *Let us consider $\underline{u}\in L^2_{loc}(\Omega)\times L^2_{loc}(\Omega)$ and let us suppose that $\underline{u}^\varepsilon\in\underline{\mathcal{V}}^\varepsilon_0$, $\varepsilon\in\Lambda$, satisfy*

$$\sup_{\varepsilon\in\Lambda}\Big(\underline{a}^\varepsilon(\underline{u}^\varepsilon)+\underline{b}^\varepsilon(\underline{u}^\varepsilon)\Big)<+\infty. \tag{2.38}$$

Then $\underline{u}\in\underline{\mathcal{V}}_0$ and $\underline{u}^\varepsilon\to\underline{u}$ according to Definition 1.1 *if and only if there exist extensions $\breve{\underline{u}}^\varepsilon\in\underline{\mathcal{V}}_{loc}$ of $\underline{u}^\varepsilon$ such that*

$$\breve{\underline{u}}^\varepsilon\rightharpoonup\underline{u}\quad\textit{weakly in }\underline{\mathcal{V}}_{loc}=H^1_{loc}(\Omega)\times H^1_{loc}(\Omega)\quad\textit{as }\varepsilon\downarrow0,\varepsilon\in\Lambda. \tag{2.39}$$

Moreover, if $\underline{u}^\varepsilon\to\underline{u}$ according to Definition 1.1, *then* every *extension $\breve{\underline{u}}^\varepsilon$ bounded in $\underline{\mathcal{V}}_{loc}$ is weakly convergent to $\underline{u}$ in $\underline{\mathcal{V}}_{loc}$; in particular, we always have*

$$\underline{T}^\varepsilon(\underline{u}^\varepsilon)\rightharpoonup\underline{u}\quad\textit{weakly in }\underline{\mathcal{V}}_{loc}=H^1_{loc}(\Omega)\times H^1_{loc}(\Omega), \tag{2.40}$$

and setting $v^\varepsilon:=\underline{B}^\varepsilon\underline{u}^\varepsilon=u^\varepsilon_i|_{\Gamma^\varepsilon}-u^\varepsilon_e|_{\Gamma^\varepsilon}$ we have

$$v^\varepsilon\to u_i-u_e\quad\textit{according to Definition }1.1. \tag{2.41}$$

Proof. If $\breve{\underline{u}}^\varepsilon$ is an extension of $\underline{u}^\varepsilon$ which is bounded in $\underline{\mathcal{V}}_{loc}$, then any weak limit point $\breve{\underline{u}}$ in $\underline{\mathcal{V}}_{loc}$ should coincide with $\underline{u}$ and belongs to $\underline{\mathcal{V}}_0$ by Proposition 2.8. Lemma 2.10 and (2.38) show that $\underline{T}^\varepsilon$ provides such an extension, so that the equivalence between the two notions of convergence is proved. ∎

Remark 2.12. Observe that for a general family $\underline{u}^\varepsilon\in\underline{\mathcal{V}}^\varepsilon_0$ converging to $\underline{u}$ according to Definition 1.1, such that $v^\varepsilon=\underline{B}^\varepsilon\underline{u}^\varepsilon$ converges to v as $\varepsilon\downarrow0,\varepsilon\in\Lambda$, it may happen that $v\ne u_i-u_e$ in Ω. The above corollary shows that this inconvenience can be avoided if $\underline{u}$ satisfies the equibounded energy condition (2.38).

PROPOSITION 2.13 (compactness for the convergence of Definition 1.1). *For $\varepsilon\in\Lambda$ let $\underline{u}^\varepsilon\in\underline{\mathcal{V}}^\varepsilon_0$ (resp., $w^\varepsilon\in L^2(\Gamma^\varepsilon)$) satisfy* (2.38) *(resp., $\sup_{\varepsilon\in\Lambda}b^\varepsilon(w^\varepsilon)<+\infty$). Then there exists a decreasing vanishing subsequence $\Lambda'=(\varepsilon_j)_{j\in\mathbb{N}}\subset\Lambda$ and an element $\underline{u}\in\underline{\mathcal{V}}_0$ (resp., $w\in\mathcal{H}$) such that $\underline{u}^{\varepsilon_j}$ converges to $\underline{u}$ as $j\to\infty$ according to Definition* 1.1 *(resp., $w^{\varepsilon_j}\to w$).*

Proof. The compactness of $\underline{u}^\varepsilon$ follows directly from Lemma 2.10 and Corollary 2.11. In the case of w^ε we observe that the total variation of the measures $\tilde{w}^\varepsilon$ introduced in (1.18) is easily bounded by

$$|\tilde{w}^\varepsilon|(\Omega) = \varepsilon \int_{\Gamma^\varepsilon} |w^\varepsilon(x)| \, d\mathscr{H}^{d-1}(x) \le C b^\varepsilon(w^\varepsilon)^{1/2}$$

thanks to (2.28). Therefore we can extract a subsequence $\Lambda' = (\varepsilon_j)_{j\in\mathbb{N}}$ and a limiting Radon measure $\tilde{w}$ in Ω such that $\tilde{w}^\varepsilon \rightharpoonup^* \tilde{w}$, i.e.,

$$\lim_{\varepsilon\to 0, \varepsilon\in\Lambda'} \varepsilon \int_{\Gamma^\varepsilon} w^\varepsilon(x)\zeta(x) \, d\mathscr{H}^{d-1}(x) = \int_\Omega \zeta(x) \, d\tilde{w}(x) \quad \forall\, \zeta \in C^0_c(\Omega). \tag{2.42}$$

On the other hand, keeping the same notation of (1.17), we have

$$w^\varepsilon = \frac{d\tilde{w}^\varepsilon}{d\lambda^\varepsilon}, \quad b^\varepsilon(w^\varepsilon) = \int_\Omega \left| \frac{d\tilde{w}^\varepsilon}{d\lambda^\varepsilon}(x)\right|^2 \, d\lambda^\varepsilon(x).$$

Since $\lambda^\varepsilon \rightharpoonup^* \lambda = \beta\mathscr{L}^d$, as shown by Remark 2.7, general lower semicontinuity results on integral functionals defined on measures [4, Thm. 2.34] show that for every convex function $\psi : \mathbb{R} \to [0, +\infty]$ with superlinear growth (in our case $\psi(s) := s^2$)

$$\liminf_{\varepsilon\downarrow 0, \varepsilon\in\Lambda'} \int_\Omega \psi\Big(\frac{d\tilde{w}^\varepsilon}{d\lambda^\varepsilon}(x)\Big) \, d\lambda^\varepsilon(x) < +\infty \quad \Rightarrow \quad \tilde{w} = w\cdot\lambda << \lambda \tag{2.43}$$

and

$$\int_\Omega \psi\big(w(x)\big) \, d\lambda(x) \le \liminf_{\varepsilon\to 0, \varepsilon\in\Lambda'} \int_\Omega \psi\Big(\frac{d\tilde{w}^\varepsilon}{d\lambda^\varepsilon}\Big) \, d\lambda^\varepsilon. \tag{2.44}$$

It follows that $\tilde{w} = \beta w \mathscr{L}^d$ for $w \in \mathcal{H} = L^2(\Omega)$ and (2.42) yields (1.16). □

2.5. Lower semicontinuity and convergence results for integral functionals on Γ^ε. Arguing as in the last part of the proof of Proposition 2.13 and taking into account (2.43) and (2.44), we have the following.

PROPOSITION 2.14 (lower semicontinuity for the convergence of Definition 1.1). *Let $v^\varepsilon \in L^1_{loc}(\Gamma^\varepsilon)$, $\varepsilon \in \Lambda$, converge to $v \in L^1_{loc}(\Omega)$ according to Definition 1.1 and let $\psi : \mathbb{R} \to [0, +\infty]$ be a convex, lower semicontinuous function with superlinear growth. We have*

$$\liminf_{\varepsilon\downarrow 0, \varepsilon\in\Lambda} \varepsilon \int_{\Gamma^\varepsilon} \psi(v^\varepsilon(x)) \, d\mathscr{H}^{d-1}(x) \ge \beta \int_\Omega \psi(v(x)) \, dx. \tag{2.45}$$

When ψ is locally Lipschitz and v^ε are uniformly bounded with uniformly bounded extensions in $H^1_{loc}(\Omega)$, we can prove a convergence result.

LEMMA 2.15. *If $\psi : \mathbb{R} \to \mathbb{R}$ is a locally Lipschitz function and*

$$v^\varepsilon \rightharpoonup v \quad \textit{weakly in } H^1_{loc}(\Omega) \textit{ as } \varepsilon \downarrow 0, \varepsilon \in \Lambda, \quad \sup_{\varepsilon\in\Lambda} \|v^\varepsilon\|_{L^\infty(\Omega)} = S < +\infty,$$

then

$$\lim_{\varepsilon\downarrow 0, \varepsilon\in\Lambda} \varepsilon \int_{\Gamma^\varepsilon} \psi(v^\varepsilon) \, d\mathscr{H}^{d-1} = \beta \int_\Omega \psi(v) \, dx. \tag{2.46}$$

Proof. Up to a possible modification of ψ outside $[-S,S]$, it is not restrictive to assume that ψ is globally Lipschitz; since $\psi(v^\varepsilon) \rightharpoonup \psi(v)$ in $H^1_{loc}(\Omega)$, it is not restrictive to assume that ψ is the identity in (2.46).

If $\zeta \in C^0_c(\Omega)$, Proposition 2.8 yields

$$\begin{aligned}\limsup_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} &\left|\varepsilon\int_{\Gamma^\varepsilon} v\, d\mathscr{H}^{d-1} - \beta\int_\Omega v\,dx\right| \\ &\le \limsup_{\varepsilon\downarrow 0}\varepsilon\int_{\Gamma^\varepsilon}|v^\varepsilon|\,|1-\zeta|\,d\mathscr{H}^{d-1} + \beta\int_\Omega |v|\,|1-\zeta|\,dx \\ &\le S\limsup_{\varepsilon\downarrow 0}\varepsilon\int_{\Gamma^\varepsilon}|1-\zeta|\,d\mathscr{H}^{d-1} + S\beta\int_\Omega|1-\zeta|\,dx \\ &\le 2S\beta\int_\Omega |1-\zeta|\,dx.\end{aligned}$$

Taking the infimum of the last integral with respect to ζ, we conclude the proof. ◻

Combining Proposition 2.14, Lemma 2.15, and the equivalence property stated by (2.41) of Corollary 2.11, we obtain the following.

COROLLARY 2.16. *If $\underline{u}^\varepsilon \in \underline{\mathcal{V}}^\varepsilon_0$, $\varepsilon \in \Lambda$, is a family satisfying the bounded energy condition* (2.30) *and converging to $\underline{u} \in \underline{\mathcal{V}}_0$ according to Definition* 1.1, *then for every convex functional $\psi : \mathbb{R} \to [0,+\infty)$ we have*

$$\liminf_{\varepsilon\downarrow 0,\varepsilon\in\Lambda}\varepsilon\int_{\Gamma^\varepsilon}\psi(u^\varepsilon_i - u^\varepsilon_e)\,d\mathscr{H}^{d-1} \ge \beta\int_\Omega \psi(u_i-u_e)\,dx. \tag{2.47}$$

Moreover, if $\sup_{\varepsilon\in\Lambda}\|u^\varepsilon_i - u^\varepsilon_e\|_{L^\infty(\Gamma^\varepsilon)} < +\infty$, *then*

$$\lim_{\varepsilon\downarrow 0,\varepsilon\in\Lambda}\int_{\Gamma^\varepsilon}\psi(u^\varepsilon_i - u^\varepsilon_e)\,d\mathscr{H}^{d-1} = \beta\int_\Omega\psi(u_i-u_e)\,dx. \tag{2.48}$$

We conclude this section with a final auxiliary result.

LEMMA 2.17 (bulk energy approximation). *Let $\psi : \mathbb{R} \to [0,+\infty)$ be a locally Lipschitz function such that $\psi' \ge 0$ in $(a,+\infty)$, $\psi' \le 0$ in $(-\infty,-a)$ for some $a > 0$, and let $\underline{u} \in \underline{\mathcal{V}}_0$ such that*

$$\int_\Omega \psi(u_i-u_e)\,dx < +\infty. \tag{2.49}$$

There exists a sequence $(\underline{u}_k)_{k\in\mathbb{N}} \subset \underline{\mathcal{V}}_0$ such that

$$\begin{gathered} u_{i,k} - u_{e,k} \in L^\infty(\Omega), \quad \lim_{k\uparrow\infty}\underline{u}_k \to \underline{u} \quad \textit{strongly in } \underline{\mathcal{V}}_0, \\ \lim_{k\uparrow\infty}\int_\Omega \psi(u_{i,k}-u_{e,k})\,dx = \int_\Omega\psi(u_i-u_e)\,dx. \end{gathered} \tag{2.50}$$

Proof. Recalling that $v := u_i - u_e$, we set

$$v_k := (v\wedge k)\vee(-k), \quad s = (u_i+u_e)/2, \quad \gamma_k := \frac{1}{\mathscr{L}^d(\Omega_0)}\int_{\Omega_0}\big(s(x) - v_k(x)/2\big)\,dx$$

and

$$u_{i,k} := s - \gamma_k + v_k/2, \quad u_{e,k} := s - \gamma_k - v_k/2.$$

By construction, $\underline{u}_k \in \underline{\mathcal{V}}_0$ and

$$u_{i,k} - u_{e,k} = v_k \in L^\infty(\Omega), \quad \lim_{k\uparrow\infty} v_k = v \quad \text{strongly in } H^1(\Omega).$$

In particular,

$$\lim_{k\uparrow\infty} \gamma_k = \frac{1}{\mathscr{L}^d(\Omega_0)} \int_{\Omega_0} \big(s(x) - v(x)/2\big)\, dx = \frac{1}{\mathscr{L}^d(\Omega_0)} \int_{\Omega_0} u_e(x)\, dx = 0,$$

so that

$$\underline{u}_k \to \underline{u} \quad \text{strongly in } \underline{\mathcal{V}}_0 \quad \text{as } k \uparrow \infty.$$

Finally, since $k \mapsto \psi(v_k(x))$ is (definitively) nondecreasing and converges pointwise to $\psi(v(x))$, Levi's theorem yields

$$\lim_{k\uparrow+\infty} \int_\Omega \psi(v_k(x))\, dx = \int_\Omega \psi(v(x))\, dx. \qquad \square$$

3. Γ-convergence results.

3.1. Γ-convergence. For the reader's convenience, we include hereafter a few definitions and theorems used in what follows [17, 19].

DEFINITION 3.1 (Γ-convergence). *Let (X, d) be a metric space and $\mathscr{F}^\varepsilon$, $\mathscr{F}$, $\varepsilon \in \Lambda$, be functionals from X into $[-\infty, +\infty]$. We say that $(\mathscr{F}^\varepsilon)_{\varepsilon\in\Lambda}$ $\Gamma(X)$-converges to $\mathscr{F}$ as $\varepsilon \downarrow 0, \varepsilon \in \Lambda$, i.e.,*

$$\mathscr{F} = \Gamma(X)\text{-}\lim_{\varepsilon\downarrow 0, \varepsilon\in\Lambda} \mathscr{F}^\varepsilon,$$

if for every $x \in X$ the following conditions are fulfilled:

$$\forall\, x^\varepsilon \in X: \quad \lim_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} x^\varepsilon = x \quad \Rightarrow \quad \liminf_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} \mathscr{F}^\varepsilon(x^\varepsilon) \geq \mathscr{F}(x), \tag{3.1}$$

$$\exists \hat{x}^\varepsilon \in X: \quad \lim_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} \hat{x}^\varepsilon = x, \quad \limsup_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} \mathscr{F}^\varepsilon(\hat{x}^\varepsilon) \leq \mathscr{F}(x). \tag{3.2}$$

Remark 3.2. Notice that by (3.1) the "lim sup" in (3.2) is in fact a limit. We will sometimes use a slight variant of this property, when $\mathscr{F}^\varepsilon, \mathscr{F}$ admit the decomposition $\mathscr{F}^\varepsilon = \mathscr{F}_1^\varepsilon + \mathscr{F}_2^\varepsilon$, $\mathscr{F} = \mathscr{F}_1 + \mathscr{F}_2$, and $\mathscr{F}_1^\varepsilon, \mathscr{F}_2^\varepsilon$ satisfy condition (3.1) with respect to $\mathscr{F}_1, \mathscr{F}_2$. In this case, every "optimal" family $\hat{x}^\varepsilon$ for $\mathscr{F}^\varepsilon$ satisfies

$$\hat{x}^\varepsilon \to x, \ \limsup_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} \mathscr{F}^\varepsilon(\hat{x}^\varepsilon) \leq \mathscr{F}(x) \quad \Rightarrow \quad \lim_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} \mathscr{F}_j^\varepsilon(\hat{x}^\varepsilon) = \mathscr{F}_j(x), \qquad j = 1, 2. \tag{3.3}$$

THEOREM 3.3 (see [19, Cor. 2.4]). *Let (X, d) be a metric space, $\mathscr{F}^\varepsilon, \mathscr{F}$, $\varepsilon \in \Lambda$, be functionals from X into $(-\infty, +\infty]$ such that $\mathscr{F} = \Gamma(X)\text{-}\lim_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} \mathscr{F}^\varepsilon$, and let $x^\varepsilon \in X$ be a minimizer for $\mathscr{F}^\varepsilon$, i.e.,*

$$\mathscr{F}^\varepsilon(x^\varepsilon) = \min\left\{\mathscr{F}^\varepsilon(x) \,:\, x \in X\right\}.$$

If the family $(x^\varepsilon)_{\varepsilon\in\Lambda}$ is relatively compact *in X and x^0 is the* unique minimizer *for $\mathscr{F}$, then*

$$\lim_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} x^\varepsilon = x^0, \quad \lim_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} \mathscr{F}^\varepsilon(x^\varepsilon) = \mathscr{F}(x^0). \tag{3.4}$$

The following useful criterion [11, p. 97] allows us to check (3.2) on a smaller $\mathscr{F}$-dense subset $D \subset X$.

PROPOSITION 3.4 (a density argument for Γ-$\limsup$). *Let $X, \mathscr{F}^\varepsilon, \mathscr{F}$ as in Definition* 3.1 *and let $D \subset X$ satisfy*

$$\forall\, x \in X \quad \exists x^\varepsilon \in D : \quad \lim_{\varepsilon \downarrow 0, \varepsilon \in \Lambda} x^\varepsilon = x, \quad \lim_{\varepsilon \downarrow 0, \varepsilon \in \Lambda} \mathscr{F}(x^\varepsilon) = \mathscr{F}(x). \tag{3.5}$$

If

$$\forall\, x \in D, \quad \exists x^\varepsilon \in X : \quad \lim_{\varepsilon \downarrow 0, \varepsilon \in \Lambda} x^\varepsilon = x, \quad \limsup_{\varepsilon \downarrow 0, \varepsilon \in \Lambda} \mathscr{F}_\varepsilon(x^\varepsilon) \le \mathscr{F}(x), \tag{3.6}$$

then the Γ-$\limsup$ *condition* (3.2) *for* Γ*-convergence is satisfied.*

Combining the results of Braides ([10]; see also [12, Thm. 14.8]) and of Acerbi et al. [1] we obtain the following homogenization result for noncoercive integral functionals.

THEOREM 3.5. *Let us consider the family of integral functionals in $L^2_{loc}(\Omega)$,*

$$a^\varepsilon_{i,e}(u) := \begin{cases} \displaystyle\int_{\Omega^\varepsilon_{i,e}} \sigma^\varepsilon_{i,e}(x)\nabla u(x)\cdot\nabla u(x)\,dx & \text{if } u \in L^2_{loc}(\Omega), u|_{\Omega_{i,e}} \in H^1(\Omega_{i,e}),\\ +\infty & \text{otherwise,}\end{cases} \tag{3.7}$$

where $\sigma^\varepsilon_{i,e}$ were introduced in (1.8) *and* (1.9)*, and let us define*

$$a_{i,e}(u) := \begin{cases} \displaystyle\int_\Omega M_{i,e}(x)\nabla u(x)\cdot\nabla u(x)\,dx & \text{if } u \in H^1(\Omega),\\ +\infty & \text{if } u \in L^2_{loc}(\Omega)\setminus H^1(\Omega)\end{cases} \tag{3.8}$$

with $M_{i,e}$ defined as in (1.21)*. Then we have*

$$a_{i,e}(u) = \Gamma(L^2_{loc}(\Omega))\text{-}\lim_{\varepsilon\downarrow 0} a^\varepsilon_{i,e}(u) = \Gamma(L^\infty(\Omega))\text{-}\lim_{\varepsilon\downarrow 0} a^\varepsilon_{i,e}(u). \tag{3.9}$$

3.2. Γ-convergence of $\mathscr{F}^\varepsilon$ and proof of Theorem 1.6. In this section we want to prove Theorem 1.6. The natural topology for this variational approach should be the one introduced by Definition 1.1. Therefore, in order to apply the Γ-convergence technique, we have to imbed the domain of the functionals $\mathscr{F}^\varepsilon, \mathscr{F}$ in a fixed underlying metric space, whose converging sequences with equibounded energy coincide with those characterized by Definition 1.1.

To this aim, we consider the space of finite signed Radon measures on Ω [4, Section 1.57],

$$\mathcal{M} := \mathcal{M}(\Omega) = \big(C_c(\Omega)\big)' = \big(C_0(\Omega)\big)', \tag{3.10}$$

endowed with the dual norm

$$\|\mu\|_{\mathcal{M}} := |\mu|(\Omega) \tag{3.11}$$

and the (weaker) continuous distance

$$\begin{aligned} d(\mu,\nu) := \sup\Big\{ \int_\Omega \zeta(x)\,d(\mu-\nu)(x) : \zeta \in C_c(\Omega)\cap \mathrm{Lip}(\Omega),\\ \sup_{x\in\Omega}|\zeta(x)| \le 1, \quad \mathrm{Lip}(\zeta,\Omega)\le 1\Big\}, \end{aligned} \tag{3.12}$$

where $\mathrm{Lip}(\Omega)$ (resp., $\mathrm{Lip}(\zeta, \Omega)$) is the space of the Lipschitz real functions defined in Ω (resp., the Lipschitz constant of ζ).

Since $C_0(\Omega)$ is a separable Banach space, the dual unit ball of $\mathcal{M}$ is weakly* compact and separable, so that the distance d induces the weak* topology of $\mathcal{M}$ on each norm-bounded subset of $\mathcal{M}$; in particular, $(\mathcal{M}, d)$ is a separable metric space and norm-bounded sequences are relatively compact with respect to the weaker topology induced by the distance d.

We then identify vectors $\underline{u}^\varepsilon = (u_i^\varepsilon, u_e^\varepsilon) \in \underline{\mathcal{V}}_0^\varepsilon, \underline{u} = (u_i, u_e) \in \underline{\mathcal{V}}_0$ with measures $\tilde{\underline{u}}^\varepsilon = (\tilde{u}_i^\varepsilon, \tilde{u}_e^\varepsilon), \tilde{\underline{u}} = (\tilde{u}_i, \tilde{u}_e) \in \mathcal{M}^2$ as in (1.18) and (1.19), denoting by $\underline{\mathsf{m}}^\varepsilon : \underline{\mathcal{V}}_0^\varepsilon \to \mathcal{M}^2$, $\underline{\mathsf{m}} : \underline{\mathcal{V}}_0 \to \mathcal{M}^2$ the corresponding maps.

This operator allows us to extend all the functionals on $\underline{\mathcal{V}}_0^\varepsilon$ to $\mathcal{M}^2$; e.g., in the case of $\mathscr{F}^\varepsilon$ we set

$$\tilde{\mathscr{F}}^\varepsilon(\tilde{\underline{u}}) := \begin{cases} \mathscr{F}^\varepsilon(\underline{u}) & \text{if } \tilde{\underline{u}} = \mathsf{m}^\varepsilon(\underline{u}), \quad \text{in } \mathcal{M}^2, \\ +\infty & \text{otherwise.} \end{cases} \tag{3.13}$$

We can thus consider coercivity and Γ-convergence of $\tilde{\mathscr{F}}^\varepsilon$ in $\mathcal{M}^2$ as $\varepsilon \downarrow 0$ which are in fact equivalent to statements (a), (b), (c) of Theorem 1.6. We split the proof of these properties in three steps. Recall that by (1.30) we can assume that $v_0^\varepsilon = \underline{B}^\varepsilon \underline{u}_0^\varepsilon$ with

$$\limsup_{\varepsilon \downarrow 0} \Big(\underline{a}^\varepsilon(\underline{u}_0^\varepsilon) + \underline{b}^\varepsilon(\underline{u}_0^\varepsilon) \Big) < +\infty, \tag{3.14}$$

v_0^ε converges to v_0 according to Definition 1.1, and $b^\varepsilon(v_0^\varepsilon) \to b(v_0)$; the functional $\mathscr{F}^\varepsilon$ takes the form

$$\mathscr{F}^\varepsilon(\underline{u}^\varepsilon) := \frac{h}{2} b^\varepsilon(v^\varepsilon - v_0^\varepsilon) + \frac{1}{2} \underline{a}^\varepsilon(\underline{u}^\varepsilon) + \phi^\varepsilon(v^\varepsilon), \quad v^\varepsilon := \underline{B}^\varepsilon \underline{u}^\varepsilon. \tag{3.15}$$

(a) *Compactness.* It follows directly from Proposition 2.13, thanks to (1.31).

(b) lim inf *inequality.* Suppose that $\underline{u}^\varepsilon \in \underline{\mathcal{V}}_0^\varepsilon$, $\varepsilon \in \Lambda$, converges to $\underline{u} \in \underline{\mathcal{V}}_0$ as $\varepsilon \downarrow 0, \varepsilon \in \Lambda$ according to Definition 1.1 and satisfies

$$\limsup_{\varepsilon \downarrow 0, \varepsilon \in \Lambda} \Big(\underline{a}^\varepsilon(\underline{u}^\varepsilon) + \underline{b}^\varepsilon(\underline{u}^\varepsilon) \Big) < +\infty. \tag{3.16}$$

Corollary 2.11 and Theorem 3.5 yield

$$\liminf_{\varepsilon \downarrow 0, \varepsilon \in \Lambda} \underline{a}^\varepsilon(\underline{u}^\varepsilon) \geq \underline{a}(\underline{u}), \tag{3.17}$$

whereas (2.47) of Corollary 2.16 and (3.14) show that

$$\liminf_{\varepsilon \downarrow 0, \varepsilon \in \Lambda} b^\varepsilon(v^\varepsilon - v_0^\varepsilon) \geq b(\underline{u} - \underline{u}_0), \quad \liminf_{\varepsilon \downarrow 0, \varepsilon \in \Lambda} \phi^\varepsilon(v^\varepsilon) \geq \phi(v), \qquad v := \underline{B}\underline{u}. \tag{3.18}$$

(c) lim sup *inequality.* We introduce the set

$$D := \Big\{ \underline{u} = (u_i, u_e) \in \underline{\mathcal{V}}_0 : \underline{B}\underline{u} = v = u_i - u_e \in L^\infty(\Omega) \Big\},$$

which satisfies the density condition (3.5) thanks to Lemma 2.17. By Proposition 3.4 it is then sufficient to prove the lim sup inequality for $\underline{u} \in D$.

By Theorem 3.5 we find a uniformly bounded family $\breve{\underline{u}}^\varepsilon = (\breve{u}_i^\varepsilon, \breve{u}_e^\varepsilon)$ converging to $\underline{u}$ in $L^\infty(\Omega) \times L^\infty(\Omega)$, whose restriction $\underline{u}^\varepsilon$ to $\Omega_i^\varepsilon \times \Omega_e^\varepsilon$ belongs to $\underline{\mathcal{V}}_0^\varepsilon$ (we can add a vanishing constant to $\breve{u}_e^\varepsilon$ as in Lemma 2.17) such that

$$\lim_{\varepsilon\downarrow 0} \underline{a}^\varepsilon(\underline{u}^\varepsilon) = \underline{a}(\underline{u}).$$

The boundedness and the regularity of Ω show that $\underline{b}^\varepsilon(\underline{u}^\varepsilon)$ is bounded so that (3.16) holds. Therefore, a simple application of (2.48) of Corollary 2.16 yields

$$\lim_{\varepsilon\downarrow 0} b^\varepsilon(v^\varepsilon) = b(v), \quad \lim_{\varepsilon\downarrow 0} \phi^\varepsilon(v^\varepsilon) = \phi(v), \qquad v^\varepsilon := \underline{B}^\varepsilon \underline{u}^\varepsilon, \quad v = \underline{B}\underline{u}. \tag{3.19}$$

In order to conclude the proof we have to pass to the limit in the term $b^\varepsilon(v^\varepsilon - v_0^\varepsilon)$. We invoke the next lemma.

LEMMA 3.6. *For every couple $v_0^\varepsilon, v^\varepsilon \in \mathcal{H}^\varepsilon$ converging to v_0, v as in Definition* 1.1, *we have*

$$\left.\begin{aligned} \lim_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} b^\varepsilon(v_0^\varepsilon) = b(v_0) \\ \limsup_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} b^\varepsilon(v^\varepsilon) = S < +\infty \end{aligned}\right\} \quad \Rightarrow \quad \lim_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} b^\varepsilon(v_0^\varepsilon, v^\varepsilon) = b(v_0, v). \tag{3.20}$$

Proof. Let us recall that

$$\liminf_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} b^\varepsilon(z^\varepsilon) \ge b(z) \tag{3.21}$$

for every family $z^\varepsilon \in \mathcal{H}^\varepsilon$ converging to z according to Definition 1.1. For every positive scalar $\rho > 0$ we have

$$2b^\varepsilon(v_0^\varepsilon, v^\varepsilon) = 2b^\varepsilon(\rho^{-1}v_0^\varepsilon, \rho v^\varepsilon) = b^\varepsilon(\rho^{-1}v_0^\varepsilon + \rho v^\varepsilon) - \rho^{-2}b^\varepsilon(v_0^\varepsilon) - \rho^2 b^\varepsilon(v^\varepsilon).$$

Taking the inferior limit as $\varepsilon \downarrow 0, \varepsilon \in \Lambda$ we get

$$\begin{aligned} \liminf_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} 2b^\varepsilon(v_0^\varepsilon, v^\varepsilon) &\ge b(\rho^{-1}v_0 + \rho v) - \rho^{-2}b(v_0) - \rho^2 S \\ &= 2b(v_0, v) + \rho^2\Big(b(v) - S\Big). \end{aligned}$$

Since $\rho > 0$ is arbitrary, we obtain

$$\liminf_{\varepsilon\downarrow 0,\varepsilon\in\Lambda} b^\varepsilon(v_0^\varepsilon, v^\varepsilon) \ge b(v_0, v);$$

inverting sign to v_0^ε we prove the lemma. ∎

As a corollary, we also obtain the following accessory result.

COROLLARY 3.7. *Let $v^\varepsilon \in \mathcal{H}^\varepsilon$ converge to $v \in \mathcal{H}$ according to Definition* 1.1 *and let us suppose that*

$$\lim_{\varepsilon\downarrow 0} b^\varepsilon(v^\varepsilon) = b(v), \quad \limsup_{\varepsilon\downarrow 0} j^\varepsilon(v^\varepsilon) < +\infty. \tag{3.22}$$

If $\underline{u}^\varepsilon = (u_i^\varepsilon, u_e^\varepsilon)$ is the unique solution of the minimum problem

$$\min\Big\{\underline{a}^\varepsilon(\underline{u}) : \underline{u} \in \underline{\mathcal{V}}_0^\varepsilon, \underline{B}^\varepsilon\underline{u} = v^\varepsilon\Big\}, \tag{3.23}$$

then $\underline{u}^\varepsilon \to \underline{u}$ as $\varepsilon \downarrow 0$ according to Definition 1.1 *and*

$$\underline{a}^\varepsilon(\underline{u}^\varepsilon) = j^\varepsilon(v^\varepsilon) \to \underline{a}(\underline{u}) = j(v), \tag{3.24}$$

with $\underline{u}$ being the unique solution of

$$\min\Big\{\underline{a}(\underline{u}) : \underline{u} \in \underline{\mathcal{V}}_0, \underline{B}\underline{u} = v\Big\}. \tag{3.25}$$

4. Variational formulation of the evolution problems. In this section we collect some basic notation and preliminary results on the variational formulation and the related well posedness of the micro- and macroscopic problems as discussed in [16].

4.1. Nonlinear terms and convex primitives. Recalling (1.10), from now on we set

$$\lambda_F := 1 + \Big(\inf_{x\in\mathbb{R}} F'(x)\Big)^{-}, \quad f(x) := F(x) + \lambda_F\, x, \tag{4.1}$$

$$\varphi(x) := \int_0^x f(\rho)\, d\rho = \frac{1}{2}\lambda_F x^2 + \int_0^x F(\rho)\, d\rho. \tag{4.2}$$

Observe that f is a strictly increasing C^1 function with $f' \geq 1$, so that φ is a strictly convex function with (at least) quadratic growth, thus satisfying

$$\varphi(x) \geq \varphi(0) = 0, \quad \varphi(x) \geq \frac{1}{2}|x|^2 \quad \forall\, x \in \mathbb{R}. \tag{4.3}$$

The conjugate function $\varphi^* : \mathbb{R} \to \mathbb{R}$ defined by

$$\varphi^*(y) = \sup_{x\in\mathbb{R}} \Big(yx - \varphi(x)\Big) \tag{4.4}$$

is still a strictly convex function which satisfies

$$\varphi^*(y) \geq \varphi^*(0) = 0 \quad \forall\, y \in \mathbb{R}, \quad \lim_{|y|\to\infty} \frac{\varphi^*(y)}{|y|} = +\infty, \tag{4.5}$$

$$xy \leq \varphi(x) + \varphi^*(y), \quad xf(x) = \varphi(x) + \varphi^*(f(x)) \quad \forall\, x, y \in \mathbb{R}. \tag{4.6}$$

In particular, we have

$$|f(x)| \leq \varphi^*(f(x)) + \sup_{|z|\leq 1} \varphi(z) \quad \forall\, x \in \mathbb{R} \tag{4.7}$$

and the "subdifferential inequality"

$$y = f(x) \quad \Leftrightarrow \quad y(z - x) \leq \varphi(z) - \varphi(x) \quad \forall\, z \in \mathbb{R}. \tag{4.8}$$

4.2. Vector notation for the complete system. Besides the couples $\underline{u}^\varepsilon = (u_i^\varepsilon, u_e^\varepsilon)$, $\underline{u} = (u_i, u_e)$, the micro- and macroscopic evolution problems involve a third "recovery" variable, w^ε, w. Thus the electric state of the heart will be determined by the three-component vectors

$$^{\varepsilon} := (\underline{u}^\varepsilon, w^\varepsilon), \qquad := (\underline{u}, w). \tag{4.9}$$

As a general rule, boldface letters (as , , $\boldsymbol{\mathcal{V}}, \dots$) will occur when the three-component vectors that determine the electric state of our systems are involved.

We will adopt the usual convention to identify functions $u = u(x,t)$ defined in the space-time cylinders $A \times (0,T)$ (A being some open subset of $\mathbb{R}^d$ endowed with the Lebesgue measure $\mathscr{L}^d$ or some Lipschitz hypersurface endowed with the $(d-1)$-dimensional Hausdorff measure $\mathscr{H}^{d-1}$) with the time-dependent function $u(\cdot,t)$ taking its values in suitable function subspaces of $L^1_{loc}(A)$ (we do not indicate explicitly the dependence on the underlying measure $\mathscr{L}^d$ or $\mathscr{H}^{d-1}$). Vector functions will therefore take their values in suitable product vector spaces, which we are now introducing.

As we said before in section 2.1, the electric potentials $\underline{u}^\varepsilon$ (resp., $\underline{u}$) will take their values in $\underline{\mathcal{V}}_0^\varepsilon$ (resp., $\underline{\mathcal{V}}_0$), whereas the recovery variables w^ε (resp., w) are valued into $\mathcal{H}^\varepsilon$ (resp., $\mathcal{H}$). Therefore, we will also introduce the following product spaces for the vector functions $\ ^\varepsilon, \ $:

$$\boldsymbol{\mathcal{V}}^\varepsilon := \underline{\mathcal{V}}^\varepsilon \times \mathcal{H}^\varepsilon, \quad \boldsymbol{\mathcal{V}}_0^\varepsilon := \underline{\mathcal{V}}_0^\varepsilon \times \mathcal{H}^\varepsilon, \quad \boldsymbol{\mathcal{V}} := \underline{\mathcal{V}} \times \mathcal{H}, \quad \boldsymbol{\mathcal{V}}_0 := \underline{\mathcal{V}}_0 \times \mathcal{H}. \tag{4.10}$$

4.3. The variational formulation of the microscopic problem and its well posedness. As shown in [16], the variational formulation of Problem $\ ^{\boldsymbol{\varepsilon}}$ can be easily obtained by performing the following steps:

1. Choose test functions $\hat{\ } = (\hat{u}_i, \hat{u}_e, \hat{w}) \in \boldsymbol{\mathcal{V}}_0^\varepsilon$;
2. multiply (P_{1a}^ε) by $\hat{u}_{i,e}$, integrate by parts using (P_{1b}^ε) and (P_4^ε), and sum up taking (P_2^ε) into account;
3. take the b^ε-scalar product in $\mathcal{H}^\varepsilon$ of (P_3^ε) with $\hat{w}$;
4. sum up the results of the previous two steps.

In order to write the variational formulation in a compact form, we introduce the bilinear forms which are defined for every $\ , \hat{\ } \in \boldsymbol{\mathcal{V}}^\varepsilon = \underline{\mathcal{V}}^\varepsilon \times \mathcal{H}^\varepsilon$,

$$\ ^\varepsilon(\ , \hat{\ }) := \underline{b}^\varepsilon(\underline{u}, \hat{\underline{u}}) + b^\varepsilon(w, \hat{w}), \qquad \ ^\varepsilon(\ , \hat{\ }) := \underline{a}^\varepsilon(\underline{u}, \hat{\underline{u}}) + \gamma b^\varepsilon(w, \hat{w}), \tag{4.11}$$

$$\ ^\varepsilon(\ , \hat{\ }) := \lambda_F \underline{b}^\varepsilon(\underline{u}, \hat{\underline{u}}) - \Theta b^\varepsilon(w, \underline{B}^\varepsilon \hat{\underline{u}}) + \eta b^\varepsilon(\underline{B}^\varepsilon \underline{u}, \hat{w}), \tag{4.12}$$

together with the related quadratic forms

$$\ ^\varepsilon(\ ) := \ ^\varepsilon(\ , \ ), \qquad \ ^\varepsilon(\ ) := \ ^\varepsilon(\ , \ ), \tag{4.13}$$

the functionals

$$\ ^\varepsilon(\ ) := \varepsilon \int_{\Gamma^\varepsilon} \varphi(u_i - u_e)\, d\mathscr{H}^{d-1}, \qquad \ ^\varepsilon(\ ) := \varepsilon \int_{\Gamma^\varepsilon} \varphi^*(f(u_i - u_e))\, d\mathscr{H}^{d-1}, \tag{4.14}$$

and the nonlinear form

$$\mathcal{F}^\varepsilon(\ , \hat{\ }) = \varepsilon \int_{\Gamma^\varepsilon} f(u_i - u_e)(\hat{u}_i - \hat{u}_e)\, d\mathscr{H}^{d-1} = \varepsilon \int_{\Gamma^\varepsilon} f(v)\hat{v}\, d\mathscr{H}^{d-1}, \tag{4.15}$$

which is well defined by (4.6) if $\ ^\varepsilon(\ ), \ ^\varepsilon(\hat{\ }) < +\infty$. Viewing $u_i^\varepsilon, u_e^\varepsilon, w^\varepsilon$ as time-dependent functions with values in a "space-dependent" functional space, as we discussed in section 4.2, we see that Problem $\ ^{\boldsymbol{\varepsilon}}$ can be solved by looking for the solution $\ ^\varepsilon : (0,T) \to \boldsymbol{\mathcal{V}}_0^\varepsilon$ of the abstract variational equation,

$$\begin{cases} \dfrac{d}{dt}\ ^\varepsilon(\ ^\varepsilon, \hat{\ }) + \ ^\varepsilon(\ ^\varepsilon, \hat{\ }) + \mathcal{F}^\varepsilon(\ ^\varepsilon, \hat{\ }) = \ ^\varepsilon(\ ^\varepsilon, \hat{\ }), \\ \qquad\qquad\qquad\qquad \ ^\varepsilon(\ ^\varepsilon(0), \hat{\ }) = \ ^\varepsilon(\ _0^\varepsilon, \hat{\ }), \end{cases} \tag{4.16}$$

for any $\hat{\ } \in \boldsymbol{\mathcal{V}}_0^\varepsilon$ with $\ ^\varepsilon(\hat{\ }) < +\infty$. Here the initial datum $\ _0^\varepsilon := (u_{0,i}^\varepsilon, u_{0,e}^\varepsilon, w_0^\varepsilon)$ satisfies

$$\ _0^\varepsilon \in \boldsymbol{\mathcal{V}}_0^\varepsilon, \qquad \ ^\varepsilon(\ _0^\varepsilon) < +\infty, \tag{4.17}$$

and it is related to v_0^ε by

$$\underline{B}^\varepsilon \underline{u}_0^\varepsilon = v_0^\varepsilon, \quad \underline{a}^\varepsilon(\underline{u}_0^\varepsilon) = j^\varepsilon(v_0^\varepsilon) = \min\left\{\underline{a}^\varepsilon(\hat{\underline{u}}^\varepsilon) : \underline{B}^\varepsilon \hat{\underline{u}}^\varepsilon = v_0^\varepsilon\right\}. \tag{4.18}$$

The following theorem provides an existence result for the microscopic problem; see [16] for more details.

THEOREM 4.1. *Let us assume that* $\ {}_0^\varepsilon$ *satisfies* (4.17) *and* (4.18)*. Then, there exists a unique solution of the variational formulation* (4.16) *of Problem* $\ ^{\boldsymbol{\varepsilon}}$*,*

$$\ ^\varepsilon = (u_i^\varepsilon, u_e^\varepsilon, w^\varepsilon) \in C^0\big([0,T]; \boldsymbol{\mathcal{V}}_0^\varepsilon\big), \quad v^\varepsilon := \underline{B}^\varepsilon \underline{u}^\varepsilon,$$

with

$$\partial_t\big(\underline{B}^\varepsilon \underline{u}^\varepsilon\big) = \partial_t v^\varepsilon, \qquad \partial_t w^\varepsilon \in L^2(0,T;\mathcal{H}^\varepsilon), \tag{4.19}$$

$$\sup_{t\in(0,T)} \ ^\varepsilon(\ ^\varepsilon) = \sup_{t\in(0,T)} \varepsilon \int_{\Gamma^\varepsilon} \varphi(v^\varepsilon)\, d\mathscr{H}^{d-1} < +\infty, \tag{4.20}$$

$$\int_0^T \ ^\varepsilon(\ ^\varepsilon)\, dt = \varepsilon \int_0^T \int_{\Gamma^\varepsilon} \varphi^*(f(v^\varepsilon))\, d\mathscr{H}^{d-1}\, dt < +\infty. \tag{4.21}$$

Moreover, the solution $\ ^\varepsilon$ *satisfies the a priori estimates*

$$\left.\begin{aligned} &\sup_{t\in[0,T]} \Big(\ ^\varepsilon(\ ^\varepsilon) + \ ^\varepsilon(\ ^\varepsilon) + \ ^\varepsilon(\ ^\varepsilon)\Big) \\ &\int_0^T \Big(b^\varepsilon(\partial_t v^\varepsilon) + b^\varepsilon(\partial_t w^\varepsilon)\Big)\, dt \end{aligned}\right\} \le C\Big(\ ^\varepsilon(\ {}_0^\varepsilon) + \ ^\varepsilon(\ {}_0^\varepsilon) + \ ^\varepsilon(\ {}_0^\varepsilon)\Big) \tag{4.22}$$

for a constant C independent of ε and of the initial datum*; finally, at each time* $t \in [0,T]$, $\underline{u}^\varepsilon(t)$ *is the unique solution of the minimum problem*

$$\underline{B}^\varepsilon \underline{u}^\varepsilon(t) = v^\varepsilon(t), \quad \underline{a}^\varepsilon(\underline{u}^\varepsilon(t)) = j^\varepsilon(v^\varepsilon(t)) = \min\Big\{\underline{a}^\varepsilon(\hat{\underline{u}}^\varepsilon) : \underline{B}^\varepsilon \hat{\underline{u}}^\varepsilon = v^\varepsilon(t)\Big\}. \tag{4.23}$$

Theorem 4.1 and (4.22) rely on two a priori estimates which are interesting by themselves: we will briefly present their formal derivation after a short discussion on the abstract structure of the system (4.16). As we will see in the next section, the main interest of this approach is that the macroscopic problem can be formulated in the same way.

4.4. The variational formulation of the macroscopic problem and its well posedness. The derivation of the variational formulation of this problem is completely analogous to the previous one; as before, we introduce the bilinear forms on $\boldsymbol{\mathcal{V}} \supset \boldsymbol{\mathcal{V}}_0$,

$$(\ ,\hat{\ }) := \underline{b}(\underline{u},\hat{\underline{u}}) + b(w,\hat{w}), \qquad (\ ,\hat{\ }) := \underline{a}(\underline{u},\hat{\underline{u}}) + \gamma b(w,\hat{w}), \tag{4.24}$$

$$(\ ,\hat{\ }) := \lambda_F \underline{b}(\underline{u},\hat{\underline{u}}) - \Theta b(w,\underline{B}\hat{\underline{u}}) + \eta b(\underline{B}\underline{u},\hat{w}), \tag{4.25}$$

together with the related quadratic forms

$$(\ ) := (\ ,\ ), \quad \underline{b}(\ ) := \underline{b}(\ ,\ ), \tag{4.26}$$

the functionals

$$(\ ) := \beta \int_\Omega \varphi(u_i - u_e)\, dx, \qquad (\ ) := \beta \int_\Omega \varphi^*(f(u_i - u_e))\, dx, \tag{4.27}$$

and the nonlinear form

$$\mathcal{F}(\ ,\hat{\ }) = \beta \int_\Omega f(u_i - u_e)(\hat{u}_i - \hat{u}_e)\, dx = \beta \int_\Omega f(v)\hat{v}\, dx, \tag{4.28}$$

which is well defined by (4.6) if $(\ ),\ (\hat{\ }) < +\infty$. The solution $: (0,T) \to \boldsymbol{\mathcal{V}}_0$ of the macroscopic problem thus satisfies the variational evolution equation,

$$\begin{cases} \dfrac{d}{dt}\ (\ ,\hat{\ }) + (\ ,\hat{\ }) + \mathcal{F}(\ ,\hat{\ }) = (\ ,\hat{\ }), \\ \qquad\qquad (\ (0),\hat{\ }) = (\ _0,\hat{\ }), \end{cases} \tag{4.29}$$

for any $\hat{\ } \in \boldsymbol{\mathcal{V}}_0$ with $(\hat{\ }) < +\infty$. Again the initial datum $_0 := (u_{0,i}, u_{0,e}, w_0)$ satisfies

$$_0 \in \boldsymbol{\mathcal{V}}_0, \qquad (\ _0) < +\infty, \tag{4.30}$$

and it is related to v_0 by

$$\underline{B}\underline{u}_0 = v_0, \quad \underline{a}(\underline{u}_0) = j(v_0) = \min\Big\{\underline{a}(\hat{\underline{u}}) : \underline{B}\hat{\underline{u}} = v_0\Big\}. \tag{4.31}$$

THEOREM 4.2. *Let us assume that* $_0$ *satisfies* (4.30), (4.31)*. Then, there exists a unique solution of the variational formulation of Problem* ,

$$= (u_i, u_e, w) \in C^0\big([0,T];\boldsymbol{\mathcal{V}}_0\big), \quad v = \underline{B}\underline{u},$$

with

$$\partial_t\big(\underline{B}\underline{u}\big) = \partial_t v, \qquad \partial_t w \in L^2(0,T;\mathcal{H}), \tag{4.32}$$

$$\sup_{t\in(0,T)}\ (\ ) = \sup_{t\in(0,T)} \int_\Omega \varphi(v)\,dx < +\infty, \tag{4.33}$$

$$\int_0^T (\ )\,dt = \int_0^T\int_\Omega \varphi^*\big(f(v)\big)\,dx\,dt < +\infty. \tag{4.34}$$

Moreover, the solution satisfies the a priori estimates

$$\left.\begin{aligned} &\sup_{t\in[0,T]}\Big(\ (\ ) + (\ ) + (\ )\Big) \\ &\int_0^T \Big(b(\partial_t w) + b(\partial_t v)\Big)\,dt \end{aligned}\right\} \le C\Big(\ (\ _0) + (\ _0) + (\ _0)\Big) \tag{4.35}$$

for a constant C *independent of the initial datum; at each time* $t \in [0,T]$, $\underline{u}(t)$ *is the unique solution of the minimum problem*

$$\underline{B}\underline{u}(t) = v(t), \quad \underline{a}(\underline{u}(t)) = j(v(t)) = \min\Big\{\underline{a}(\hat{\underline{u}}) : \underline{B}\hat{\underline{u}} = v(t)\Big\}. \tag{4.36}$$

4.5. Structural properties and a priori estimates. Now, we point out some distinctive properties of $^\varepsilon,\ ^\varepsilon, \mathcal{F}^\varepsilon$, and $^\varepsilon$ (see [16]): they are also valid for the macroscopic model, which corresponds to $\varepsilon = 0$.

Notation 4.3. In order to avoid tedious repetitions, whenever it is possible we will systematically include the case $\varepsilon = 0$ in our statements simply by making the obvious identifications

$$\underline{u}^0 := \underline{u},\ w^0 := w,\ v^0 := v,\ \ ^0 := \ ,\ \underline{a}^0 := \underline{a},\ b^0 := b,\ \underline{\mathcal{V}}^0 := \underline{\mathcal{V}}, \ldots. \tag{4.37}$$

It can be verified that $^\varepsilon,\ ^\varepsilon,\ \mathcal{F}^\varepsilon$, and $^\varepsilon$ satisfy (see [16]):

(A) $^\varepsilon$ is continuous and *symmetric*; the associated quadratic form (still denoted by $^\varepsilon$) is *nonnegative* but its kernel has infinite dimension, so that (4.16) is a *degenerate evolution equation.*

(B) $\ ^{\varepsilon}$ is a continuous, symmetric, and nonnegative bilinear form, too.

(C) The sum of the quadratic forms $\ ^{\varepsilon} + \ ^{\varepsilon}$ is coercive in $\boldsymbol{\mathcal{V}}_0^{\varepsilon}$, thus providing an equivalent scalar product.

(D) The bilinear form $\ ^{\varepsilon}$ satisfies

$$\exists\, G \geq 0: \quad |\ ^{\varepsilon}(\ ,\hat{\ })|^2 \leq G^2\ ^{\varepsilon}(\ )\ ^{\varepsilon}(\hat{\ }) \quad \forall\ ,\hat{\ } \in \boldsymbol{\mathcal{V}}_0^{\varepsilon}, \tag{4.38}$$

where G is *independent of* ε; in the present case we can choose $G = \lambda_F + \Theta + \eta$.

(E) The nonlinear form $\mathcal{F}$ satisfies the subdifferential inequalities

$$\mathcal{F}^{\varepsilon}(\ ,\ ) = \ ^{\varepsilon}(\ ) + \ ^{\varepsilon}(\ ) \quad \forall \ \in \boldsymbol{\mathcal{V}}_0^{\varepsilon}, \tag{4.39}$$

$$\mathcal{F}^{\varepsilon}(\ ,\hat{\ } - \ ) + \ ^{\varepsilon}(\ ) \leq \ ^{\varepsilon}(\hat{\ }) \quad \forall\ ,\hat{\ } \in \boldsymbol{\mathcal{V}}_0^{\varepsilon} \tag{4.40}$$

for the

$$\text{convex and lower semicontinuous functional } \ ^{\varepsilon}, \quad \text{with } \ ^{\varepsilon}(\ ) \geq \ ^{\varepsilon}(\ ) \quad \forall \ \in \boldsymbol{\mathcal{V}}_0^{\varepsilon}. \tag{4.41}$$

Remark 4.4 (regularity of F). Concerning the nonlinearity of the problems, we note that the regularity assumptions on F can be relaxed, so $F : \mathbb{R} \to \mathbb{R}$ can be a continuous function such that $F(0) = 0$ and

$$\exists\, \lambda_F \geq 0: \quad (F(x) - F(y))(x-y) + (\lambda_F - 1)|x-y|^2 \geq 0 \quad \forall\, x, y \in \mathbb{R}. \tag{4.42}$$

Now we briefly show a formal derivation of the basic a priori estimates for the macroscopic problem (for the microscopic one, one can simply add the superscript ε to each occurrence of $\ ,\ ,\mathcal{F},\ ,\ ,\ ,\dots$); we assume that $\ \in H^1(0,T;\boldsymbol{\mathcal{V}}_0)$ and f has a linear growth, so that $\mathcal{F}$ is a continuous form. The computations below can be made rigorous, e.g., by passing to the limit in the analogous stability estimates for a suitably discretized or regularized system, and will be studied in section 6.

Recalling that

$$\frac{d}{dt}\ (\ (t),\hat{\ }) = \ (\ '(t),\hat{\ }), \quad \frac{1}{2}\frac{d}{dt}\ (\ (t)) = \ (\ '(t),\ ) \quad \forall\, \hat{\ } \in \boldsymbol{\mathcal{V}}_0,$$

choosing $\hat{\ } := \ (t)$ in (4.29) we get

$$\frac{1}{2}\frac{d}{dt}\ (\ (t)) + \ (\ (t)) + \mathcal{F}(\ (t),\ (t)) = \ (\ (t),\ (t)) \leq G\ (\ (t)), \tag{4.43}$$

so that a simple application of the Gronwall lemma and the relation

$$\mathcal{F}(\ ,\ ) = \ (\ ) + \ (\ ) \tag{4.44}$$

yields

$$\max\left[\sup_{t\in[0,T]}\ (\ (t)), \int_0^T \Big(\ (\ ) + \ (\ ) + \ (\ )\Big)\, dt\right] \leq e^{2GT}\ (\ _0). \tag{4.45}$$

Choosing now $\hat{\ } := \ '$ and observing that

$$\frac{d}{dt}\ (\ (t)) = \mathcal{F}(\ (t),\ '(t)), \quad \frac{1}{2}\frac{d}{dt}\ (\ (t)) = \ (\ (t),\ '(t)),$$

the previous estimate and the Cauchy inequality yield

$$
\begin{aligned}
(4.46)\qquad & (\ '(t)) + \frac{d}{dt}\Big(\frac{1}{2}\ (\ (t)) + \ (\ (t))\Big) = \ (\ (t),\ '(t)) \\
& \leq G\ (\ (t))^{1/2}\ (\ '(t))^{1/2} \leq \frac{1}{2}\ (\ '(t)) + \frac{G^2}{2} e^{2Gt}\ (\ _0).
\end{aligned}
$$

Integrating in time we get

$$
(4.47)\qquad \left.\begin{matrix} \sup\limits_{t\in[0,T]} \left(\frac{1}{2}\ (\ ) + \ (\ )\right) \\ \frac{1}{2}\int_0^T\ (\ ')\,dt \end{matrix}\right\} \leq \frac{1}{2}\ (\ _0) + \ (\ _0) + \frac{G}{4} e^{2GT}\ (\ _0).
$$

Combining (4.45) with (4.47) we obtain (4.35) with $C = \max\big(2, (\frac{G}{2}+1)e^{2GT}\big)$.

5. Proof of Theorems 1.3 and 1.4.

5.1. Outline. As we said in the introduction, our approach (inspired by the so-called "minimizing movement" method introduced by De Giorgi [18, 3, 5]) combines general approximation results, yielding uniform error estimates for a semi-implicit Euler time discretization of evolution equations such as (4.16) and (4.29), with homogenization results for the discretized problems.

More precisely, we begin in section 5.2 by considering the approximation of $^{\varepsilon}$ and in the time interval $[0,T]$ by the semi-implicit Euler method of time step $\tau = T/N > 0$: thus we will consider a uniform partition $\mathcal{P}_\tau$ of the time interval $[0,T]$ into N subintervals

$$
(5.1)\qquad \mathcal{P}_\tau := \{0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T\}, \quad t_n := n\tau,
$$

and we will replace the continuous problems by a sequence of discrete microscopic and macroscopic problems $^{\varepsilon,}$, ^{0,} , $n = 1, \ldots, N$, whose solutions $^{\varepsilon,\tau}_{n}$, $^{0,\tau}_{n}$ provide an approximation of $^{\varepsilon}(t)$, (t) for t in the time interval $(t_{n-1}, t_n]$.

Denoting by $^{\varepsilon,\tau}$, $^{0,\tau}$ the piecewise linear interpolant on $\mathcal{P}_\tau$ of the values $^{\varepsilon,\tau}_{n}$, $^{0,\tau}_{n}$, general error estimates for variational evolution problems with the structure discussed in section 4.5 show that the error (measured in the natural energy norms) between the continuous solution $^{\varepsilon}$ of $^{\varepsilon}$ (resp., of) and the discrete solution $^{\varepsilon,\tau}$ (resp., $^{0,\tau}$) vanishes as $\tau \downarrow 0$ *uniformly with respect to* ε. We will devote section 6 to proving these estimates.

Since the nonsymmetric parts $^{\varepsilon}$, are discretized "explicitly" it turns out (see section 5.3) that each step of the discretization scheme (the problems $^{\varepsilon,}$ and ^{0,}) involves the minimization of suitable functionals $\Phi^{\varepsilon,\tau}_n, \Phi^{0,\tau}_n$, which depend only on ε, τ and on the discrete solutions $^{\varepsilon,\tau}_{n-1}$, $^{0,\tau}_{n-1}$ at the previous node of the partition $\mathcal{P}_\tau$; therefore $^{\varepsilon,\tau}_{n}$ and $^{0,\tau}_{n}$ are the unique minima of $\Phi^{\varepsilon,\tau}_n$ and $\Phi^{0,\tau}_n$, respectively.

By adapting the Γ-convergence results of section 3, we shall see (Theorem 5.2) that each discrete value $^{\varepsilon,\tau}_{n}$ converges, as $\varepsilon \downarrow 0$ to $^{0,\tau}_{n}$, whenever τ is fixed. Since for each $\tau > 0$ the discrete interpolant $^{\varepsilon,\tau}$ is determined only by a *finite* number of vectors $^{\varepsilon,\tau}_{n}$, it follows that the discrete solution $^{\varepsilon,\tau}$ converges to the homogenized one $^{0,\tau}$ in the time interval $[0,T]$ as $\varepsilon \downarrow 0$.

Combining this result with the above uniform error estimate between $^{\varepsilon}$, and $^{\varepsilon,\tau}$, $^{0,\tau}$, we will conclude also that the continuous solution $^{\varepsilon}$ converges to at each time $t \in [0,T]$.

Thus, by introducing the auxiliary problems $^{\boldsymbol{\varepsilon},}$, $^{\mathbf{0},}$, the homogenization of time-dependent evolution equations is reduced to the homogenization of a family of functionals depending on the parameters τ, n, which can be tackled by Γ-convergence arguments.

Let us reproduce the above argument in the following scheme:

$$
\begin{array}{ccc}
{}^{\varepsilon} := \begin{cases} \text{Continuous solution of} \\ \text{the evolution problem } \boldsymbol{P}^{\boldsymbol{\varepsilon}} \end{cases} & \underset{\substack{\textit{uniform error} \\ \textit{estimates}}}{\longleftrightarrow} & {}^{\varepsilon,\tau} := \begin{cases} \text{Discrete solution of the} \\ \text{variational problem } \boldsymbol{P}^{\boldsymbol{\varepsilon},\boldsymbol{\tau}} \end{cases} \\
 & & \downarrow \ \Gamma\textit{-convergence} \\
:= \begin{cases} \text{Continuous solution of} \\ \text{the evolution problem } \boldsymbol{P} \end{cases} & \underset{\substack{\textit{uniform error} \\ \textit{estimates}}}{\longleftrightarrow} & {}^{0,\tau} := \begin{cases} \text{Discrete solution of the} \\ \text{variational problem } \boldsymbol{P}^{\mathbf{0},\boldsymbol{\tau}} \end{cases}
\end{array}
$$

Let us now consider in more detail each step of the proof.

5.2. Time discretization. We look for a suitable approximation $\{\ {}^{\varepsilon,\tau}_{n}\}_{n=0}^{N} \subset \boldsymbol{\mathcal{V}}_0^{\varepsilon}$ of the values of $\ ^{\varepsilon}$ on the grid $\mathcal{P}_\tau$ (5.1), which solves the following discrete problem $^{\boldsymbol{\varepsilon},}$:

$$
(\ \boldsymbol{\varepsilon},\ ) \qquad \begin{cases} \text{Given } \ {}^{\varepsilon,\tau}_{0} := \ {}^{\varepsilon}_{0} \quad \text{find } \ {}^{\varepsilon,\tau}_{1}, \ldots, \ {}^{\varepsilon,\tau}_{N} \in \boldsymbol{\mathcal{V}}_0^{\varepsilon} \text{ which recursively solve} \\ {}^{\varepsilon}\left(\dfrac{\ {}^{\varepsilon,\tau}_{n} - \ {}^{\varepsilon,\tau}_{n-1}}{\tau}, \ \hat{}\ \right) + \ {}^{\varepsilon}(\ {}^{\varepsilon,\tau}_{n}, \ \hat{}\ ) + \mathcal{F}^{\varepsilon}(\ {}^{\varepsilon,\tau}_{n}, \ \hat{}\ ) = \ {}^{\varepsilon}(\ {}^{\varepsilon,\tau}_{n-1}, \ \hat{}\ ) \\ \qquad\qquad \text{for every choice of } \ \hat{}\ \in D(\ ) \subset \boldsymbol{\mathcal{V}}_0^{\varepsilon}. \end{cases}
$$

In fact, due to the convexity of and to the coercivity of the quadratic form $\ ^{\varepsilon} + \ ^{\varepsilon}$ on $\boldsymbol{\mathcal{V}}_0^{\varepsilon}$, it is easy to check that each step of ($\boldsymbol{\varepsilon},$) is equivalent to the minimum problem

$$
(\ \boldsymbol{\varepsilon},\ ) \quad \text{find } \ {}^{\varepsilon,\tau}_{n} \in \boldsymbol{\mathcal{V}}_0^{\varepsilon} \text{ which attains the minimum } \quad \min\left\{\Phi_n^{\varepsilon,\tau}(\ ) : \ \in \boldsymbol{\mathcal{V}}_0^{\varepsilon}\right\},
$$

where

$$
\Phi_n^{\varepsilon,\tau}(\ ) := \frac{1}{2\tau}\ {}^{\varepsilon}(\ - \ {}^{\varepsilon,\tau}_{n-1}) + \ {}^{\varepsilon}(\ ) + \ {}^{\varepsilon}(\ ) - \ {}^{\varepsilon}(\ {}^{\varepsilon,\tau}_{n-1}, \ ). \tag{5.2}
$$

We can proceed in a completely analogous way for the macroscopic problem, simply by setting $\varepsilon := 0$ and recalling Notation 4.3. The corresponding discrete solutions $\ ^{\varepsilon,\tau}(t)$ and $\ ^{0,\tau}(t)$ are the piecewise linear interpolants of $\{\ {}^{\varepsilon,\tau}_{n}\}_{n=0}^{N}$ and $\{\ {}^{0,\tau}_{n}\}_{n=0}^{N}$ on the grid $\mathcal{P}_\tau$, i.e.,

$$
\ {}^{\varepsilon,\tau}(t) := (n - t/\tau)\ {}^{\varepsilon,\tau}_{n-1} + (t/\tau - (n-1))\ {}^{\varepsilon,\tau}_{n} \quad \text{if } t \in ((n-1)\tau, n\tau]. \tag{5.3}
$$

The next theorem shows that the discrete solutions of the above schemes converge uniformly to the solutions of $\ ^{\boldsymbol{\varepsilon}}$ and in the intrinsic energy norms induced by the bilinear forms $\ ^{\varepsilon}$, $\ ^{\varepsilon}$. We denote by $\mathsf{E}^{\varepsilon}, \mathsf{E} = \mathsf{E}^0$ (recall Notation 4.3) the (squared) errors

$$
\mathsf{E}^{\varepsilon} = \max_{t \in (0,T)} \ {}^{\varepsilon}(\ {}^{\varepsilon}(t) - \ {}^{\varepsilon,\tau}(t)) + \int_0^T \ {}^{\varepsilon}(\ {}^{\varepsilon}(t) - \ {}^{\varepsilon,\tau}(t))\, dt. \tag{5.4}
$$

THEOREM 5.1 (uniform error estimates). *There exists a constant* $C = C(G, T)$ independent of ε *such that*

$$\mathsf{E}^\varepsilon \le C\tau\Big({}^{\varepsilon}({}^{\varepsilon}_{0}) + {}^{\varepsilon}({}^{\varepsilon}_{0}) + {}^{\varepsilon}({}^{\varepsilon}_{0})\Big). \tag{5.5}$$

The proof of Theorem 5.1 is presented in section 6: it is the extension to a semi-implicit discretization scheme of the arguments developed in [9]; in order to give a self-contained simpler proof, we decided to follow the general scheme introduced by [29] for the derivation of optimal a priori and a posteriori error estimates for evolution variational problems.

5.3. Discrete problems and the role of Γ-convergence. Let us now consider the convergence as $\varepsilon \downarrow 0$ of the solutions ${}^{\varepsilon,\tau}_{n}$ of the discrete problem ${}^{\varepsilon,}$. The following crucial result provides the main induction step.

THEOREM 5.2 (convergence of the discrete approximations). *Let us suppose that the vector* ${}^{\varepsilon,\tau}_{n-1} = (U^{\varepsilon,\tau}_n, W^{\varepsilon,\tau}_n) \in \boldsymbol{\mathcal{V}}^\varepsilon_0$ *converges to* ${}^{0,\tau}_{n-1}$ *according to Definition* 1.1 *as* $\varepsilon \downarrow 0$*, with*

$$\lim_{\varepsilon\downarrow 0} {}^{\varepsilon}({}^{\varepsilon,\tau}_{n-1}) = ({}^{0,\tau}_{n-1}), \quad \limsup_{\varepsilon\downarrow 0} {}^{\varepsilon}({}^{\varepsilon,\tau}_{n-1}) + {}^{\varepsilon}({}^{\varepsilon,\tau}_{n-1}) < +\infty, \tag{5.6}$$

and let us define the functional $\Phi^{\varepsilon,\tau}_n, \Phi^{0,\tau}_n$ *as in* (5.2). *Then the unique minimum* ${}^{\varepsilon,\tau}_{n} \in \boldsymbol{\mathcal{V}}^\varepsilon_0$ *of* $\Phi^{\varepsilon,\tau}_n$ *converges to the unique minimum* ${}^{0,\tau}_{n}$ *of* $\Phi^{0,\tau}_n$ *as* $\varepsilon \downarrow 0$ *according to Definition* 1.1 *and*

$$\lim_{\varepsilon\downarrow 0} {}^{\varepsilon}({}^{\varepsilon,\tau}_{n}) = ({}^{0,\tau}_{n}), \quad \lim_{\varepsilon\downarrow 0} {}^{\varepsilon}({}^{\varepsilon,\tau}_{n}) = ({}^{0,\tau}_{n}), \quad \lim_{\varepsilon\downarrow 0} {}^{\varepsilon}({}^{\varepsilon,\tau}_{n}) = ({}^{0,\tau}_{n}). \tag{5.7}$$

COROLLARY 5.3 (convergence of the discrete solutions). *Let us suppose that the assumption* (1.23) *of Theorem* 1.3 *on the initial data* $(v^\varepsilon_0, w^\varepsilon_0)$ *and* (v_0, w_0) *holds and that* ${}^{\varepsilon}_{0} \in \boldsymbol{\mathcal{V}}^\varepsilon_0$ *is chosen as in* (4.18). *Then each vector* ${}^{\varepsilon,\tau}_{n}$, $n = 1, \ldots, N$, *of the discrete solution* ${}^{\varepsilon,\tau}$ *converges to the corresponding one* ${}^{0,\tau}_{n}$ *of the discrete solution* ${}^{0,\tau}$ *as* $\varepsilon \downarrow 0$ *according to Definition* 1.1.

Thanks to Theorem 3.3, we will deduce Theorem 5.2 from a corresponding coercivity and Γ-convergence result for the functionals $\Phi^{\varepsilon,\tau}_n$. Observe that (5.7) is a consequence of the convergence of the energies $\Phi^{\varepsilon,\tau}_n({}^{\varepsilon,\tau}_{n}) \to \Phi^{0,\tau}_n({}^{0,\tau}_{n})$ given by (3.4) and Remark 3.2, thanks to the separate lower semicontinuity property (3.17).

As in section 3.2 we identify vectors ${}^{\varepsilon} = (u^\varepsilon_i, u^\varepsilon_e, w^\varepsilon) \in \boldsymbol{\mathcal{V}}^\varepsilon_0$, $\;= (u_i, u_e, w) \in \boldsymbol{\mathcal{V}}_0$ with measures $\tilde{\ }^{\varepsilon} = (\tilde u^\varepsilon_i, \tilde u^\varepsilon_e, \tilde u^\varepsilon)$, $\tilde{\ } = (\tilde u_i, \tilde u_e, \tilde u) \in \mathcal{M}^3$ through (1.18) and (1.19), denoting by $\mathsf{m}^\varepsilon : \boldsymbol{\mathcal{V}}^\varepsilon_0 \to \mathcal{M}^3$, $\mathsf{m} : \boldsymbol{\mathcal{V}}_0 \to \mathcal{M}^3$ the corresponding maps. We also extend all the functionals on $\boldsymbol{\mathcal{V}}^\varepsilon_0$ to $\mathcal{M}^3$ as we did in (3.13). We can thus consider the Γ-limit of $\tilde\Phi^{\varepsilon,\tau}_n$ in $\mathcal{M}^3$ as $\varepsilon \downarrow 0$, keeping τ fixed.

THEOREM 5.4 (Γ-convergence). *Let us fix* $\tau > 0$ *and let us suppose that the vectors* ${}^{\varepsilon,\tau}_{n-1} \in \boldsymbol{\mathcal{V}}^\varepsilon_0$ *satisfy the same assumption as in Theorem* 5.2. *If a family* $\tilde{\ }^{\varepsilon} = \mathsf{m}^\varepsilon({}^{\varepsilon})$, ${}^{\varepsilon} \in \boldsymbol{\mathcal{V}}^\varepsilon_0$ *for* $\varepsilon \in \Lambda$*, satisfies*

$$\limsup_{\varepsilon\downarrow 0, \varepsilon\in\Lambda} \tilde\Phi^{\varepsilon,\tau}_n(\tilde{\ }^{\varepsilon}) = \limsup_{\varepsilon\downarrow 0, \varepsilon\in\Lambda} \Phi^{\varepsilon,\tau}_n({}^{\varepsilon}) < +\infty, \tag{5.8}$$

then $\tilde{\ }^{\varepsilon}$ *is relatively compact in* $\mathcal{M}^3$*; it converges to* $\tilde{\ }$ *in* $\mathcal{M}^3$ *if and only if* $\tilde{\ } = \mathsf{m}(\ )$ *for some* $\;\in \boldsymbol{\mathcal{V}}_0$ *and* ${}^{\varepsilon}$ *converges to* $\;$ *according to Definition* 1.1.

The functionals $\tilde{\Phi}_n^{\varepsilon,\tau}$, extensions of (5.2), *satisfy*

$$\Gamma(\mathcal{M}^3)\text{-}\lim_{\varepsilon\downarrow 0}\tilde{\Phi}_n^{\varepsilon,\tau} = \tilde{\Phi}_n^{0,\tau}. \tag{5.9}$$

The proof of Theorem 5.4 is completely analogous to that of Theorem 1.6: if we write

$${}^{\varepsilon,\tau}_{n-1} = (\underline{U}^{\varepsilon,\tau}_{n-1}, W^{\varepsilon,\tau}_{n-1}), \qquad {}^{\varepsilon} = (\underline{U}^{\varepsilon}, W^{\varepsilon}), \quad \underline{U}^{\varepsilon} \in \underline{\mathcal{V}}^{\varepsilon}_0,\ W \in \mathcal{H}^{\varepsilon},$$

we choose $h := \tau^{-1}$ and the term v_0^ε of (1.28) as $\underline{B}^\varepsilon \underline{U}^{\varepsilon,\tau}_{n-1}$, then

$$\Phi_n^{\varepsilon,\tau}(\ {}^{\varepsilon}) = \mathscr{F}^\varepsilon(\underline{U}^\varepsilon) + \ {}^{\varepsilon,\tau}_{n}(\ {}^{\varepsilon}), \tag{5.10}$$

where

$$\begin{aligned} {}^{\varepsilon,\tau}_{n}(\ {}^{\varepsilon}) := {}& \frac{1}{2\tau} b^\varepsilon(W^\varepsilon - W^{\varepsilon,\tau}_{n-1}) + \frac{\gamma}{2} b^\varepsilon(W^\varepsilon) \\ & - \lambda \underline{b}^\varepsilon(\underline{U}^{\varepsilon,\tau}_{n-1}, \underline{U}^\varepsilon) + \Theta b^\varepsilon(W^{\varepsilon,\tau}_{n-1}, \underline{B}^\varepsilon \underline{U}^\varepsilon), -\eta b^\varepsilon(\underline{B}^\varepsilon \underline{U}^{\varepsilon,\tau}_{n-1}, W^\varepsilon). \end{aligned} \tag{5.11}$$

Thus the asymptotic behavior of $\Phi_n^{\varepsilon,\tau}$ can be easily deduced from Theorem 1.6 and Lemma 3.6.

Remark 5.5. Usually, when one considers gradient flows of convex functionals $\mathscr{F}^\varepsilon$ in a *fixed* Hilbert space, their asymptotic behavior is determined by the Γ-convergence and Mosco-convergence of Lyapunov functionals $\mathscr{F}^\varepsilon$ (see, e.g., [6]), and it is not necessary to take into account the dependence of τ.

Here also the underlying Hilbert spaces are changing in a singular way with respect to ε and the major novelty is the explicit presence of the mesh size τ in the minimizing functionals $\Phi_n^{\varepsilon,\tau}$, which reflects the metric of the functional spaces governing the gradient flows. In the present case, this metric is related to the quadratic forms ${}^{\varepsilon}$: therefore it is degenerate and it depends on ε.

5.4. Conclusion of the proof of Theorems 1.3 and 1.4. By Corollary 3.7, the a priori estimates (4.22) and (4.23), and the variational characterizations (4.23), (4.36), we simply have to show that $v^\varepsilon := \underline{B}^\varepsilon \underline{u}^\varepsilon$ and w^ε are converging to $v := \underline{B}\underline{u}$ and w for every $t \in [0,T]$ according to Definition 1.1.

Let us consider the case of v^ε, the argument for w^ε being completely analogous. We set $V^{\varepsilon,\tau}(t) := \underline{B}^\varepsilon \underline{U}^{\varepsilon,\tau}(t)$, $\varepsilon \ge 0$: for every $\zeta \in C^0_c(\Omega)$ and every $t \in [0,T]$ (which will not be indicated explicitly) we have

$$\begin{aligned} \varepsilon \int_{\Gamma^\varepsilon} v^\varepsilon \zeta \, d\mathscr{H}^{d-1} - \beta \int_\Omega v\zeta \, dx = {}& \varepsilon \int_{\Gamma^\varepsilon} \big(v^\varepsilon - V^{\varepsilon,\tau}\big)\zeta \, d\mathscr{H}^{d-1} - \beta \int_\Omega \big(v - V^{0,\tau}\big)\zeta \, dx \\ & + \varepsilon \int_{\Gamma^\varepsilon} V^{\varepsilon,\tau}\zeta \, d\mathscr{H}^{d-1} - \beta \int_\Omega V^{0,\tau}\zeta \, dx \end{aligned}$$

so that, if $Z := \sup_\Omega |\zeta|$ and $S^2 \ge \varepsilon \mathscr{H}^{d-1}(\Gamma^\varepsilon) + \beta \mathscr{L}^d(\Omega)$,

$$\begin{aligned} \left| \varepsilon \int_{\Gamma^\varepsilon} v^\varepsilon \zeta \, d\mathscr{H}^{d-1} - \beta \int_\Omega v\zeta \, dx \right| \le {}& SZ b^\varepsilon\big(v^\varepsilon - V^{\varepsilon,\tau}\big)^{1/2} + SZ b\big(v - V^{0,\tau}\big)^{1/2} \\ & + \left| \varepsilon \int_{\Gamma^\varepsilon} V^{\varepsilon,\tau}\zeta \, d\mathscr{H}^{d-1} - \beta \int_\Omega V^{0,\tau}\zeta \, dx \right| \le C\tau^{1/2} + r_\varepsilon, \end{aligned}$$

where we applied the uniform estimates of Theorem 5.1 and the bounds on the initial data. Passing now to the limit as $\varepsilon \downarrow 0$ keeping τ fixed, we get

$$\limsup_{\varepsilon\downarrow 0} r_\varepsilon = \limsup_{\varepsilon\downarrow 0} \left| \varepsilon \int_{\Gamma^\varepsilon} v^\varepsilon \zeta \, d\mathscr{H}^{d-1} - \beta \int_\Omega v\zeta \, dx \right| \leq C\tau^{1/2}$$

thanks to Corollary 5.3. Finally letting $\tau \downarrow 0$ we get the desired convergence.

The corresponding property for the energy $b^\varepsilon(v^\varepsilon)$ follows by the same argument.

Let us eventually consider Theorem 1.4: if we show the existence of a suitable extension satisfying the uniform bound (1.26), then the thesis follows immediately from Theorem 1.3 and Corollary 2.11. Thanks to the a priori estimates (4.22) and Lemma 2.10, the choice $\breve{u}^\varepsilon_{i,e} := T^\varepsilon_{i,e} u^\varepsilon_{i,e}$ will surely satisfy (1.26); therefore we should find an analogous extension for w^ε. Recalling that w^ε satisfies on Γ^ε the ordinary differential equation

$$\partial_t w^\varepsilon + \gamma w^\varepsilon = \eta v^\varepsilon,$$

so that

$$w^\varepsilon(x,t) = w^\varepsilon_0 e^{-\gamma t} + \eta \int_0^t e^{-\gamma(t-s)} v^\varepsilon(x,s)\, ds \quad \forall\, x \in \Gamma^\varepsilon,\ t \in [0,T],$$

we can use the same formula to extend w^ε to Ω starting from the initial datum $\breve{w}^\varepsilon_0 \in H^1_{loc}(\Omega)$. Therefore, we set $\breve{v}^\varepsilon(x,s) := \breve{u}^\varepsilon_i - \breve{u}^\varepsilon_e \in H^1_{loc}(\Omega)$ and correspondingly

$$\breve{w}^\varepsilon(x,t) = \breve{w}^\varepsilon_0 e^{-\gamma t} + \eta \int_0^t e^{-\gamma(t-s)} \breve{v}^\varepsilon(x,s)\, ds \quad \forall\, x \in \Omega,\ t \in [0,T], \tag{5.12}$$

and it is easy to see by differentiating under the integral sign that (1.26) is satisfied.

6. Uniform error estimates. In this section we prove Theorem 5.1; since all the estimates will depend only on the structural assumptions of section 4.5 and will therefore be independent of ε, for the sake of simplicity we will not indicate the explicit dependence on ε.

Thus, $\{\ {}^{\tau}_{n}\}_{n=0}^{N}$ is a discrete solution of the variational algorithm introduced in section 5.2. We already denoted by $\ {}^{\tau}$ the piecewise linear interpolant of the discrete values, which can be expressed in the form

$$\ {}^{\tau}(t) := (1-\ell(t))U^\tau_{n-1} + \ell(t) U^\tau_n \quad \text{if } t \in ((n-1)\tau, n\tau], \tag{6.1}$$

where ℓ is the piecewise linear (but discontinuous) function associated with the mesh $\mathcal{P}_\tau$ by

$$\ell(t) := \frac{t}{\tau} - (n-1) \quad \text{if } t \in ((n-1)\tau, n\tau].$$

The piecewise constant interpolant $\bar{\ }^{\tau}$ is defined by

$$\bar{\ }^{\tau}(t) := \ {}^{\tau}_{n} \quad \text{if } t \in ((n-1)\tau, n\tau]. \tag{6.2}$$

The basic quantity which will control our estimates is

$$\mathscr{E}(\ ) := \frac{1}{2}\ (\ ) + \ (\ ) \geq \ (\ ). \tag{6.3}$$

We split the proof into several steps, denoting by C different constants which solely depend on G and T.

Discrete variational inequality. The discrete solution solves

$$\text{(6.4)}\qquad \left(\frac{{}^{\tau}_{n} - \ {}^{\tau}_{n-1}}{\tau},\ {}^{\tau}_{n} - \quad \right) + \frac{1}{2}\ (\ {}^{\tau}_{n} - \quad) + \mathscr{E}(\ {}^{\tau}_{n})$$
$$\le \mathscr{E}(\quad) + \ (\ {}^{\tau}_{n-1},\ {}^{\tau}_{n} - \quad) \quad \forall \quad \in D(\phi).$$

This property follows from the well-known (see, e.g., [8]) variational characterization of the minima for a functional, such as Φ^{τ}_{n}, which is the sum of a quadratic continuous form (involving , ,) and a convex functional (in this case).

Stability estimates. There exists a constant $C = C(G,T)$ such that

$$\text{(6.5)}\qquad \sum_{n=1}^{N} \tau\ \left(\frac{{}^{\tau}_{n} - \ {}^{\tau}_{n-1}}{\tau}\right) + \ (\ {}^{\tau}_{n} - \ {}^{\tau}_{n-1}) + \sup_{n=0,\dots,N} \mathscr{E}(\ {}^{\tau}_{n}) \le C\mathscr{E}(\ {}_{0}).$$

We use a "discrete" version of the arguments of the formal a priori estimate of section 4.5. We choose $:= 0$ in (6.4); recalling the identity $2\ (x-y,x) = \ (x) - \ (y) + \ (x-y)$ and multiplying by 2τ we obtain

$$\text{(6.6)}\qquad \begin{aligned} & b(\ {}^{\tau}_{n}) + \ (\ {}^{\tau}_{n} - \ {}^{\tau}_{n-1}) + 2\tau\ (\ {}^{\tau}_{n}) + 2\tau\phi(\ {}^{\tau}_{n}) \\ &\quad \le b(\ {}^{\tau}_{n-1}) + 2\tau\ (\ {}^{\tau}_{n-1},\ {}^{\tau}_{n}) \\ &\quad = b(\ {}^{\tau}_{n-1}) + 2\tau\Big(\ (\ {}^{\tau}_{n-1},\ {}^{\tau}_{n-1}) + \ (\ {}^{\tau}_{n-1},\ {}^{\tau}_{n} - \ {}^{\tau}_{n-1})\Big) \\ &\quad \le b(\ {}^{\tau}_{n-1}) + 2\tau G^2(1+\tau)\ (\ {}^{\tau}_{n-1}) + \ (\ {}^{\tau}_{n} - \ {}^{\tau}_{n-1}). \end{aligned}$$

A simple application of the discrete Gronwall lemma yields

$$\text{(6.7)}\qquad \sup_{0\le m\le N} b(\ {}^{\tau}_{m}) \le C_0\ (\ {}_{0}) \quad \text{for } C_0 := e^{2G^2(1+\tau)T}.$$

Choosing $:= \ {}^{\tau}_{n-1}$ in (6.4) and summing up for $n = 1$ to $m \le N$ we get

$$\text{(6.8)}\qquad \begin{aligned} & \sum_{n=1}^{m}\Big(\tau^{-1}\ \Big(\ {}^{\tau}_{n} - \ {}^{\tau}_{n-1}) + \frac{1}{2}\ (\ {}^{\tau}_{n} - \ {}^{\tau}_{n-1})\Big) + \mathscr{E}(\ {}^{\tau}_{m}) \\ &\quad \le \mathscr{E}(\ {}_{0}) + \sum_{n=1}^{m}\ (\ {}^{\tau}_{n-1},\ {}^{\tau}_{n} - \ {}^{\tau}_{n-1}) \\ &\quad \le \mathscr{E}(\ {}_{0}) + G\sum_{n=1}^{m}\Big(\ (\ {}^{\tau}_{n-1})\ (\ {}^{\tau}_{n} - \ {}^{\tau}_{n-1})\Big)^{1/2} \\ &\quad \le \mathscr{E}(\ {}_{0}) + \frac{G^2\tau}{2}\sum_{n=1}^{m}\ (\ {}^{\tau}_{n-1}) + \frac{\tau^{-1}}{2}\sum_{n=1}^{m}\ (\ {}^{\tau}_{n} - \ {}^{\tau}_{n-1}) \end{aligned}$$

so that choosing $m = N$ we get

$$\sum_{n=1}^{N}\Big(\tau^{-1}\ \Big(\ {}^{\tau}_{n} - \ {}^{\tau}_{n-1}) + \ (\ {}^{\tau}_{n} - \ {}^{\tau}_{n-1})\Big) \le 2\mathscr{E}(\ {}_{0}) + C_0 G^2 T\ (\ {}_{0}),$$
$$\sup_{1\le m\le N} \mathscr{E}(\ {}^{\tau}_{m}) \le \mathscr{E}(\ {}_{0}) + \frac{1}{2}C_0 G^2 T\ (\ {}_{0}).$$

A continuous version of the discrete variational inequalities. The piecewise linear interpolant $\ ^{\tau}$ satisfies

$$
\begin{aligned}
&\left(\frac{d}{dt}\ ^{\tau}(t),\ \ ^{\tau}(t)-\ \ \right)+\frac{1}{2}\ \ (\ ^{-\tau}(t)-\ \ )+\mathscr{E}(\ ^{\tau}(t))\\
&\quad\le \mathscr{E}(\ \ )+\ (\ ^{\tau}(t)-\tau\ell(t)\frac{d}{dt}\ ^{\tau}(t),\ \ ^{\tau}(t)-\ \ )+\mathcal{R}^{\tau}(t),
\end{aligned}
\tag{6.9}
$$

where

$$
\mathcal{R}^{\tau}(t):=(1-\ell)\Big(\mathscr{E}(\ ^{\tau}_{n-1})-\mathscr{E}(\ ^{\tau}_{n})+\ (\ ^{\tau}_{n-1},\ \ ^{\tau}_{n}-\ \ ^{\tau}_{n-1})\Big),
$$

and

$$
\int_0^T\left(\ \left(\frac{d}{dt}\ ^{\tau}(t)\right)+\tau^{-1}\ \ (\ ^{\tau}(t)-\ ^{-\tau}(t))\right)dt\le C\mathscr{E}(\ _0).
\tag{6.10}
$$

We check (6.9) simply by writing the discrete value of all the terms for $t\in(t_{n-1},t_n)$, recalling that

$$
\frac{d}{dt}\ ^{\tau}(t)=\frac{\ ^{\tau}_{n}-\ \ ^{\tau}_{n-1}}{\tau}\quad\text{for } t\in(t_{n-1},t_n).
\tag{6.11}
$$

Thus we have

$$
\begin{aligned}
\left(\frac{d}{dt}\ ^{\tau},\ \ ^{\tau}-\ \ \right)&=\left(\frac{\ ^{\tau}_{n}-\ \ ^{\tau}_{n-1}}{\tau},(1-\ell)\ \ ^{\tau}_{n-1}+\ell\ \ ^{\tau}_{n}-\ \ \right)\\
&=\left(\frac{\ ^{\tau}_{n}-\ \ ^{\tau}_{n-1}}{\tau},\ \ ^{\tau}_{n}-\ \ \right)-(1-\ell)\tau^{-1}\ \ (\ ^{\tau}_{n}-\ \ ^{\tau}_{n-1})
\end{aligned}
\tag{6.12}
$$

$$
\begin{aligned}
\mathscr{E}(\ ^{\tau})&=\mathscr{E}((1-\ell)\ \ ^{\tau}_{n-1}+\ell\ \ ^{\tau}_{n})\le(1-\ell)\mathscr{E}(\ ^{\tau}_{n-1})+\ell\mathscr{E}(\ ^{\tau}_{n})\\
&=\mathscr{E}(\ ^{\tau}_{n})+(1-\ell)\Big(\mathscr{E}(\ ^{\tau}_{n-1})-\mathscr{E}(\ ^{\tau}_{n})\Big)
\end{aligned}
\tag{6.13}
$$

$$
\begin{aligned}
&\left(\ \ ^{\tau}-\tau\ell\frac{d}{dt}\ ^{\tau},\ \ ^{\tau}-\ \ \right)\\
&\quad=\ ((1-\ell)\ \ ^{\tau}_{n-1}+\ell\ \ ^{\tau}_{n}-\ell(\ ^{\tau}_{n}-\ \ ^{\tau}_{n-1}),(1-\ell)\ \ ^{\tau}_{n-1}+\ell\ \ ^{\tau}_{n}-\ \ )\\
&\quad=\ (\ ^{\tau}_{n-1},\ \ ^{\tau}_{n}-\ \ )-(1-\ell)\ (\ ^{\tau}_{n-1},\ \ ^{\tau}_{n}-\ \ ^{\tau}_{n-1}).
\end{aligned}
\tag{6.14}
$$

Equation (6.10) follows directly from (6.5) by (6.11) and

$$
\ ^{-\tau}(t)-\ \ ^{\tau}(t)=(1-\ell)\big(\ ^{\tau}_{n}-\ \ ^{\tau}_{n-1}\big)\quad\text{for } t\in(t_{n-1},t_n).
\tag{6.15}
$$

An estimate for the remainder term.

$$
\int_0^T\left|\mathcal{R}^{\tau}(t)\right|dt\le C_1\mathscr{E}(\ _0)\tau.
\tag{6.16}
$$

First of all, since $\ ^{\tau}_{n}$ minimizes Φ^{τ}_{n}, we easily have

$$
\mathscr{E}(\ ^{\tau}_{n})-\ (\ ^{\tau}_{n-1},\ \ ^{\tau}_{n})\le\mathscr{E}(\ ^{\tau}_{n-1})-\ (\ ^{\tau}_{n-1},\ \ ^{\tau}_{n-1})
$$

so that $\mathcal{R}^{\tau}(t)\ge 0$. Moreover

$$
\begin{aligned}
\int_0^T\mathcal{R}^{\tau}(t)\,dt&=\frac{1}{2}\tau\sum_{n=1}^{N}\Big(\mathscr{E}(\ ^{\tau}_{n-1})-\mathscr{E}(\ ^{\tau}_{n})+\ (\ ^{\tau}_{n-1},\ \ ^{\tau}_{n}-\ \ ^{\tau}_{n-1})\Big)\\
&=\frac{1}{2}\tau\Big(\mathscr{E}(\ _0)+\mathscr{E}(\ _0)+C_0G^2T\ (\ _0)\Big)\le C_1\tau\mathscr{E}(\ _0),
\end{aligned}
$$

where we used, as in (6.8),

$$\sum_{n=1}^{N} \ (\ {}^{\tau}_{n-1},\ {}^{\tau}_{n} - \ {}^{\tau}_{n-1}) \leq \frac{G^2\tau}{2}\sum_{n=1}^{N} \ (\ {}^{\tau}_{n-1}) + \frac{\tau^{-1}}{2}\sum_{n=1}^{N} \ (\ {}^{\tau}_{n} - \ {}^{\tau}_{n-1})$$
$$\leq \frac{C_0G^2T}{2} \ (\ _0) + \mathscr{E}(\ _0) + \frac{C_0G^2T}{2} \ (\ _0).$$

A Gronwall-type estimate for the error. If $\ ^{\eta}$, $\eta > 0$, is the discrete solution associated with the partition $\mathcal{P}_\eta$, we have

$$\sup_{t\in[0,T]} \ (\ ^{\tau}(t) - \ ^{\eta}(t)) + \int_0^T \ (\ ^{\tau}(t) - \ ^{\eta}(t)) \leq C(\tau+\eta)\mathscr{E}(\ _0). \tag{6.17}$$

Let ℓ^τ, ℓ^η be the interpolating functions corresponding to $\mathcal{P}_\tau, \mathcal{P}_\eta$. Choosing $\ :=\ ^{\eta}(t)$ in (6.9) and $\ :=\ ^{\tau}(t)$ in the analogous inequality written for $\ ^{\eta}$, we obtain

$$\frac{d}{dt} \ (\ ^{\tau} - \ ^{\eta}) + \ (\ ^{-\tau} - \ ^{\eta}) + \ (\ ^{\eta} - \ ^{-\tau})$$
$$\leq 2\ (\ ^{\tau} - \ ^{\eta},\ ^{\tau} - \ ^{\eta}) - 2\left(\tau\ell^\tau\frac{d}{dt}\ ^{\tau} - \eta\ell^\eta\frac{d}{dt}\ ^{\eta},\ ^{\tau} - \ ^{\eta}\right) + \mathcal{R}^\tau + \mathcal{R}^\eta$$
$$\leq 3G\ (\ ^{\tau} - \ ^{\eta}) + 2G\tau^2\ \left(\frac{d}{dt}\ ^{\tau}\right) + 2G\eta^2\ \left(\frac{d}{dt}\ ^{\eta}\right) + \mathcal{R}^\tau + \mathcal{R}^\eta.$$

A direct application of the Gronwall lemma, (6.16), and (6.10) yields

$$\sup_{t\in[0,T]} (U^\tau(t) - \ ^{\eta}(t)) \leq \left(\int_0^T 2G\tau^2\ \left(\frac{d}{dt}\ ^{\tau}\right) + 2G\tau^2\ \left(\frac{d}{dt}\ ^{\eta}\right)\right.$$
$$\left. + \mathcal{R}^\tau + \mathcal{R}^\eta\, dt\right)e^{3GT}$$
$$\leq C\tau\mathscr{E}(\ _0).$$

An analogous argument and (6.10) provide the integral bound for $\ (\ ^{\tau} - \ ^{-\tau})$.

If now we pass to the limit as $\eta \downarrow 0$ we obtain the estimates of (5.5).

Appendix. The derivation of the scaled problem $\ ^{\varepsilon}$. For completeness, in this appendix, we present the scaling used to obtain problem $\ ^{\varepsilon}$.

The basic equations modeling the electrical activity of the heart at the cellular level can be obtained as follows. Cardiac tissue consists of interconnected cells surrounded by extracellular fluid. Let Ω_i^ε, Ω_e^ε be the intra- and extracellular ohmic conductive media, Γ^ε be the excitable membrane that separates Ω_i^ε and Ω_e^ε, and let $\nu_i^\varepsilon, \nu_e^\varepsilon$ denote the unit exterior normals to the boundary of Ω_i^ε and Ω_e^ε, respectively, satisfying $\nu_i^\varepsilon = -\nu_e^\varepsilon$ on Γ^ε. The electric behavior of the tissue is described by the intra- and extracellular potentials u_i^ε and u_e^ε and by their driven current densities $\mathbf{j}_{i,e}^\varepsilon = -\Sigma_{i,e}\nabla u_{i,e}^\varepsilon$. Due to the current conservation law, the normal current flux through the membrane Γ^ε is continuous $\nu_i^\varepsilon\cdot\mathbf{j}_i^\varepsilon = \nu_e^\varepsilon\cdot\mathbf{j}_e^\varepsilon$; hence we have

$$\Sigma_i\nabla u_i^\varepsilon\cdot\nu_i^\varepsilon + \Sigma_e\nabla u_e^\varepsilon\cdot\nu_e^\varepsilon = 0 \quad \text{on } \Gamma^\varepsilon\times(0,T),$$

where $\Sigma_{i,e}$ are the cellular conductivity matrices in the intra- and extracellular media.

Since the only active source elements lie on the membrane Γ^ε, each flux equals the membrane current per unit area J_m, i.e.,

$$J_m = \begin{cases} -\Sigma_i \nabla u_i^\varepsilon \cdot \nu_i^\varepsilon \\ \Sigma_e \nabla u_e^\varepsilon \cdot \nu_e^\varepsilon \end{cases} \quad \text{on } \Gamma^\varepsilon \times (0,T). \tag{A.1}$$

The membrane current per unit area J_m consists of a capacitive term and an ionic term (see [23]):

$$J_m := \mathsf{C}_m \partial_t v^\varepsilon + I(v^\varepsilon, w^\varepsilon) \quad \text{on } \Gamma^\varepsilon, \tag{A.2}$$

with C_m the surface capacitance of the membrane.

Moreover, disregarding the presence of applied current terms, we have that currents are conserved in Ω_i^ε and Ω_e^ε; then the intra- and extracellular potentials are solutions of

$$-\operatorname{div}(\Sigma_i \nabla u_i^\varepsilon) = 0 \quad \text{in } \Omega_i^\varepsilon \times (0,T) \qquad -\operatorname{div}(\Sigma_e \nabla u_e^\varepsilon) = 0 \quad \text{in } \Omega_e^\varepsilon \times (0,T) \tag{A.3}$$

with Neumann boundary conditions for u_i^ε, u_e^ε on the remaining part of the boundaries $\Gamma_{i,e}^\varepsilon = \partial\Omega_{i,e}^\varepsilon \setminus \Gamma^\varepsilon$.

We now want to rewrite problem (A.1)–(A.3) in a nondimensional form. To this end we note that we can consider two characteristic length scales: the microscopic scale, related to a typical dimension d_c of the cells (e.g., the cell diameter 15–20 μm or the length of the cell 100 μm), and the macroscopic one determined by a suitable length constant of the tissue denoted by L. The cellular conductivity matrices Σ_i and Σ_e are symmetric positive definite matrices; let $\bar\lambda = \bar\lambda_i + \bar\lambda_e$ with $\bar\lambda_i$, $\bar\lambda_e$ be the average eigenvalues on a cell element and let us consider

$$\sigma_{i,e} = \Sigma_{i,e}/\bar\lambda.$$

Here we assume that $v^\varepsilon = 0$, $w^\varepsilon = 0$ is the equilibrium point for problem (A.1)–(A.3); then we can define the macroscopic space scale along fibers L as

$$\mathsf{L} = \sqrt{\mathsf{d}_c \mathsf{R}_m \bar\lambda} \quad \text{with} \quad \mathsf{R}_m^{-1} = \partial_v I(0,0).$$

Now, we can convert the cellular problem into a nondimensional form by scaling space and time with the macroscopic units of length $\mathsf{L} = \mathsf{d}_c/\varepsilon$ and with respect to the membrane constant $\mathsf{T} = \mathsf{R}_m\ \mathsf{C}_m$; i.e., we perform the space and time scaling

$$\widehat{\ } = \ /\mathsf{L}, \quad \widehat{t} = t/\mathsf{T}.$$

The dimensionless parameter ε is then a small parameter whose order of magnitude is the ratio of the two macro- and microscopic space scales, i.e.,

$$\varepsilon = d_c/\mathsf{L}.$$

We take $\widehat{\ }$ to be the variable of the macroscale behavior and

$$\xi := \widehat{\ }/\varepsilon$$

to be the microscopic space variable measured in a unit cell. For simplicity, in what follows, we omit the hats $\widehat{\ }$ on the dimensionless variables.

Cardiac tissue exhibits a number of significant inhomogeneities, in particular, those related to cell-to-cell communications. The conductivity tensors are considered dependent on both the slow and the fast variables, i.e., $\sigma_{i,e}(x, \frac{x}{\varepsilon})$. The latter dependence of the intracellular conductivity represents an attempt to include the effects of the gap junctions.

We then define the rescaled symmetric conductivity matrices

$$\sigma^{\varepsilon}_{i,e}(x) = \sigma_{i,e}\left(x, \frac{x}{\varepsilon}\right)$$

obtained by the continuous functions $\sigma_{i,e}(x,\xi) : \Omega \times E_{i,e} \to \mathbb{M}^{d\times d}$ satisfying

$$\text{(A.4)} \qquad \begin{array}{c} 0 < \sigma|y|^2 \le \sigma_{i,e}(x,\xi)y\cdot y \le \sigma^{-1}|y^2| \\ \sigma_{i,e}(x,\xi + \ _k) = \sigma_{i,e}(x,\xi) \end{array} \quad \forall\,(x,\xi)\in\Omega\times E_{i,e}, \qquad y \in \mathbb{R}^d.$$

Finally, rescaling (A.1)–(A.3) in the intra- and extracellular potentials we obtain

$$\begin{aligned} -\operatorname{div}\left(\sigma^{\varepsilon}_{i,e}\nabla u^{\varepsilon}_{i,e}\right) &= 0 \quad \text{in } \Omega^{\varepsilon}_{i,e}\times(0,T),\\ -\sigma^{\varepsilon}_i\nabla u^{\varepsilon}_i\cdot\nu^{\varepsilon}_i &= \varepsilon\left(\partial_t v + I(v^{\varepsilon},w^{\varepsilon})\right) \quad \text{on } \Gamma^{\varepsilon}\times(0,T)\\ \sigma^{\varepsilon}_e\nabla u^{\varepsilon}_e\cdot\nu^{\varepsilon}_e &= \varepsilon\left(\partial_t v + I(v^{\varepsilon},w^{\varepsilon})\right) \quad \text{on } \Gamma^{\varepsilon}\times(0,T), \end{aligned}$$

that is, problem $\ ^{\varepsilon}$.

A homogenization process for different mathematical models, describing the response of biological tissues to electromagnetic fields and based on a completely different scaling, can be found in [2].

REFERENCES

[1] E. Acerbi, V. Chiadò Piat, D. Dal Maso, and D. Percivale, *An extension theorem from connected sets, and homogenization in general periodic domain*, Nonlinear Anal., 18 (1992), pp. 481–496.

[2] M. Amar, D. Andreucci, P. Bisegna, and R. Gianni, *Evolution and memory effects in the homogeneization limit for electrical conduction in biological tissue*, Math. Model. Methods Appl. Sci., 14 (2004), pp. 1261–1295.

[3] L. Ambrosio, *Minimizing mouvements*, Rend. Accad. Naz. Sci. XL Mem. Mat. Appl., 5 (1995), pp. 191–246.

[4] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford Mathematical Monographs, Clarendon Press, Oxford University Press, New York, 2000.

[5] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Lectures in Mathematics ETH Zürich, Birkhäuser, Basel, 2005.

[6] H. Attouch, *Variational Convergence for Functions and Operators*, Applicable Math. Ser., Pitman (Advanced Publishing Program), Boston, MA, 1984.

[7] C. Baiocchi, *Discretization of evolution variational inequalities*, in Partial Differential Equations and the Calculus of Variations, Vol. I, F. Colombini, A. Marino, L. Modica, and S. Spagnolo, eds., Birkhäuser Boston, Boston, MA, 1989, pp. 59–92.

[8] C. Baiocchi and A. Capelo, *Variational and Quasivariational Inequalities*, John Wiley and Sons, Belfast, 1984.

[9] F. Bassetti, *Variable time-step discretization of degenerate evolution equations in Banach spaces*, Numer. Funct. Anal. Optim., 24 (2003), pp. 391–426.

[10] A. Braides, *Omogeneizzazione di integrali non coercivi*, Ricerche Mat., 32 (1984), pp. 347–368.

[11] A. Braides, *Approximation of Free-Discontinuity Problems*, Lecture Notes in Math. 1694, Springer-Verlag, Berlin, 1998.

[12] A. Braides and A. Defranceschi, *Homogenization of Multiple Integrals*, Clarendon Press, Oxford, 1998.

[13] R. G. Casten, H. Cohen, and P. A. Lagerstrom, *Perturbation analysis of an approximation to the Hodgkin–Huxley theory*, Quart. Appl. Math., 32 (1974/75), pp. 365–402.

[14] P. Colli Franzone, L. Guerri, and S. Rovida, *Wavefront propagation in an activation model of the anisotropic cardiac tissue: Asymptotic analysis and numerical simulations*, J. Math. Biol., 28 (1990), pp. 121–176.

[15] P. Colli Franzone, L. Guerri, and S. Tentoni, *Mathematical modeling of the excitation process in the myocardial tissue: Influence of the fiber rotation on the wavefront propagation and the potential field*, Math. Biosci., 101 (1990), pp. 155–235.

[16] P. Colli Franzone and G. Savaré, *Degenerate evolution systems modeling the cardiac electric field at micro and macroscopic level*, in Evolution Equations, Semigroups and Functional Analysis, Progr. Nonlinear Differential Equations Appl. 50, Birkhäuser, Basel, 2002, pp. 49–78.

[17] G. Dal Maso, *An Introduction to Γ-Convergence*, Birkäuser Boston, Boston, MA, 1993.

[18] E. De Giorgi, *New problems on minimizing movements*, in Boundary Value Problems for PDE and Applications, C. Baiocchi and J. L. Lions, eds., Masson, 1993, pp. 81–98.

[19] E. De Giorgi and T. Franzoni, *Su un tipo di convergenza variazionale*, Rend. Semin. Mat. Brescia, 3 (1979), pp. 63–101.

[20] R. FitzHugh, *Mathematical models of excitation and propagation in nerve*, in Biological Engineering, H. P. Schwan, ed., McGraw-Hill, New York, 1969.

[21] C. S. Henriquez, *Simulating the electrical behavior of cardiac tissue using the bidomain model*, Crit. Rev. Biomed. Engrg., 21 (1993), pp. 1–77.

[22] R. H. Hoyt, M. L. Cohen, and J. E. Saffitz, *Distribution and three-dimensional structure of intercellular junctions in canine myocardium*, Circ. Res., 64 (1989), pp. 563–574.

[23] J. Jack, D. Noble, and R. W. Tsien, *Electric Current Flow in Excitable Cells*, Clarendon Press, Oxford, 1983.

[24] J. Keener and J. Sneyd, *Mathematical Physiology*, Springer-Verlag, New York, 1998.

[25] J. P. Keener and A. V. Panfilov, *A biophysical model for defibrillation of cardiac tissue*, Biophys. J., 71 (1996), pp. 1335–1345.

[26] J. P. Keener, *The effect of gap junctional distribution on defibrillation*, Chaos, 1 (1998), pp. 175–187.

[27] U. Mosco, *Convergence of convex sets and of solutions of variational inequalities*, Adv. Math., 3 (1969), pp. 510–585.

[28] J. S. Neu and W. Krassowska, *Homogenization of syncitial tissues*, Crit. Rev. Biomed. Engrg., 21 (1993), pp. 137–199.

[29] R. H. Nochetto, G. Savaré, and C. Verdi, *A posteriori error estimates for variable time-step discretizations of nonlinear evolution equations*, Comm. Pure Appl. Math., 53 (2000), pp. 525–589.

[30] M. L. Pressler, P. N. Munster, and X. Huang, *Gap junction distribution in the heart: Functional relevance*, in Cardiac Electrophysiology from Cell to Bedside, 2nd ed., D. P. Zipes and J. Jalife, eds., W. B. Saunders, Philadelphia, 1995, pp. 144–151.

[31] B. J. Roth, *How the anisotropy of the intracellular and extracellular conductivities influences stimulation of cardiac muscle*, J. Math. Biol., 30 (1992), pp. 633–646.

[32] J. Rulla, *Error analysis for implicit approximations to solutions to Cauchy problems*, SIAM J. Numer. Anal., 33 (1996), pp. 68–87.

[33] G. Savaré, *Weak solutions and maximal regularity for abstract evolution inequalities*, Adv. Math. Sci. Appl., 6 (1996), pp. 377–418.

[34] G. Savaré and G. Schimperna, *Domain perturbations and estimates for the solutions of second order elliptic equations*, J. Math. Pures Appl. (9), 81 (2002), pp. 1071–1112.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1371–1395

LAGRANGIAN SOLUTIONS OF SEMIGEOSTROPHIC EQUATIONS IN PHYSICAL SPACE*

MICHAEL CULLEN† AND MIKHAIL FELDMAN‡

Abstract. The semigeostrophic equations are a simple model of large-scale atmosphere/ocean flows. Previous work by J.-D. Benamou and Y. Brenier, M. Cullen and W. Gangbo, and M. Cullen and H. Maroofi proves that the semigeostrophic equations can be solved in the cases, respectively, of 3-dimensional (3-d) incompressible flow between rigid boundaries, vertically averaged 3-d incompressible flow with a free surface, and fully compressible flow. However, all these results prove only the existence of weak solutions in "dual" variables, where the dual variables result from a change of variables introduced by Hoskins. This makes it difficult to relate the solutions to the full Euler or Navier–Stokes equations, or to those of other simple atmosphere/ocean models. We therefore seek to extend these results to prove existence of a solution in physical variables. We do this using the Lagrangian form of the equations in physical space. The proof is based on the recent results of L. Ambrosio on transport equations and ODE for BV vector fields.

Key words. semigeostrophic equations, Lagrangian solutions, transport equations, BV vector fields

AMS subject classifications. 35A05, 35Q35

DOI. 10.1137/040615444

1. Introduction. The semigeostrophic equations are a simple model of large-scale atmosphere/ocean flows, where "large-scale" is defined to mean that the flow is rotation-dominated [4]. They are also accurate in the case where one horizontal scale becomes small, allowing them to describe weather fronts and jet streams. Previous work by Benamou and Brenier [2], Cullen and Gangbo [5], and Cullen and Maroofi [6] proves that the semigeostrophic equations can be solved in the cases, respectively, of 3-dimensional (3-d) incompressible flow between rigid boundaries, vertically averaged 3-d incompressible flow with a free surface, and fully compressible flow. However, all these results only prove the existence of weak solutions in "dual" variables, where the dual variables result from the change of variables introduced by [11]. This makes it difficult to relate the solutions to the full Euler or Navier–Stokes equations, or to those of other simple atmosphere/ocean models. We therefore seek to extend the results of [2] and [5] to prove existence of a solution in physical variables. We do this using the Lagrangian form of the equations in physical space. The proof is based on the recent results of L. Ambrosio [1] on transport equations and ODE for BV vector fields. It is not clear whether it is possible to prove existence of a weak solution to the Eulerian form of the equations.

The transport theory of [1] gives uniqueness. However, the analysis of [2], [5], and [6] does not give uniqueness, because of the dependence of the transport velocity on the transported quantity. This question remains open.

*Received by the editors September 21, 2004; accepted for publication (in revised form) June 1, 2005; published electronically January 10, 2006. This work was supported by National Science Foundation grants DMS-0074037 and DMS-0354729.

http://www.siam.org/journals/sima/37-5/61544.html

†Met Office, Fitzroy Road, Exeter Devon EX1 3PB, United Kingdom (mike.cullen@metoffice.com).

‡Department of Mathematics, University of Wisconsin, Madison, WI 53706 (feldman@math.wisc.edu). This author's research was supported by National Science Foundation grant DMS-0200644.

The remaining part of the paper is organized as follows: in section 2 we consider a 3-d incompressible semigeostrophic system in a domain with rigid boundary, and in section 3 we consider the semigeostrophic shallow water model.

2. Lagrangian solutions of 3-d incompressible semigeostrophic system in a domain with rigid boundary in physical space.

2.1. Background. Let $\Omega \subset \mathbf{R}^3$ be an open bounded set. We study the following semigeostrophic system:

$$
\begin{aligned}
&D_t(v_1^g, v_2^g) + (\partial_1 p, \partial_2 p) = (u_2, -u_1), \quad (v_1^g, v_2^g) = (-\partial_2 p, \partial_1 p),\\
&D_t \rho = 0, \quad \operatorname{div} u = 0, \quad \partial_3 p + \rho = 0,\\
&D_t = \partial_t + u \cdot \nabla, \quad \nabla = (\partial_{x_1}, \partial_{x_2}, \partial_{x_3}),
\end{aligned}
\tag{2.1}
$$

where $u = (u_1, u_2, u_3)$ is the velocity, p is the pressure, and ρ is the density. All these quantities are functions of $(t, x) \in (0, T) \times \Omega$. The initial and boundary data are

$$
\begin{aligned}
&u \cdot \nu = 0 \quad \text{on } [0, T) \times \partial\Omega,\\
&p(0, x) = p_0(x) \quad \text{in } \Omega,
\end{aligned}
\tag{2.2}
$$

ν is the outward normal to $\partial\Omega$, and $p_0(x)$ is a given function.

Introducing the function

$$P(t, x) = p(t, x) + \frac{1}{2}(x_1^2 + x_2^2),$$

we rewrite (2.1) as the following system of equations for P, $u = (u_1, u_2, u_3)$ depending on $(t, x) \in (0, T) \times \Omega$:

$$
\begin{aligned}
&D_t X = J(X - x),\\
&\operatorname{div} u = 0,\\
&X = \nabla P,\\
&u \cdot \nu = 0 \quad \text{on } [0, T) \times \partial\Omega,\\
&P(0, x) = P_0(x) \quad \text{in } \Omega,
\end{aligned}
\tag{2.3}
$$

where $P_0(x) = p_0(x) + \frac{1}{2}(x_1^2 + x_2^2)$. Here

$$J = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Note that a solution of (2.3) determines a solution of the original system (2.1), (2.2). Indeed, given a solution (P, u) of (2.3), we have $X = \nabla P$, we set the function u in (2.1) to be equal to the function u in (2.3), and

$$
\begin{aligned}
&p(t, x) = P(t, x) - \frac{1}{2}(x_1^2 + x_2^2),\\
&\rho(t, x) = -\partial_3 p(t, x).
\end{aligned}
$$

Then (2.1), (2.2) holds, in particular the equation $D_t \rho = 0$ follows from the equation $D_t X_3 = 0$ of (2.3). The form (2.3) is more convenient than the original system because it allows the system to be stated in the Lagrangian sense.

The solutions of [2] and [5] were obtained by interchanging dependent and independent variables in (2.3) to give a set of equations in the dual coordinates (t, X), where $X = (X_1, X_2, X_3)$. We now regard x as a function of (t, X) which is shown in [2] to be given by defining

$$P^*(t,X) = \sup_{x\in\Omega}(x \cdot X - P(t,x)) \quad \text{for } (t,X) \in [0,T) \times \mathbf{R}^3 \tag{2.4}$$

and setting $x = \nabla P^*(t, X)$. The semigeostrophic system then takes the form

$$\partial_t \alpha + \nabla \cdot (U\alpha) = 0 \quad \text{in } [0,T) \times \mathbf{R}^3, \tag{2.5}$$

$$\nabla P(t,\cdot)\#\chi_\Omega = \alpha(t,\cdot) \quad \text{for any } t \in [0,T); \tag{2.6}$$

$$U(t,X) = J(X - \nabla P^*(t,X)), \tag{2.7}$$

$$\alpha(0,X) = \alpha_0(X) \quad \text{for a.e. } X \in \mathbf{R}^3. \tag{2.8}$$

The initial data $\alpha_0(X)$ are chosen by applying (2.6) with $P(0,\cdot) = P_0(x)$. Here (2.6) means that the map $\nabla P(t,\cdot)$ pushes χ_Ω forward to $\alpha(t,\cdot)$; see Definition A.1 in Appendix A.

The proofs in [2] and [5] depended on the use of the convexity principle introduced by Cullen and Purser [7], which requires that the function $P(t,\cdot)$ is convex. Shutts and Cullen [13] relate this condition to a physical stability condition required for the semigeostrophic approximation to be appropriate. Equation (2.4) shows that $P^*(t,\cdot)$ also has to be convex, as the Legendre–Fenchel transform of P.

In this paper we are interested in the solution of the problem in the original ("physical") coordinates, i.e., in the solution of problem (2.3). The unknown functions in this problem are the (modified) pressure P and the velocity u. The Eulerian form of the first equation of (2.3) is

$$\partial_t X + u \cdot \nabla X = J(X - x). \tag{2.9}$$

Using $\operatorname{div} u = 0$, this expression can be written in the divergence form, which yields the following definition of a weak (Eulerian) solution of (2.3).

Definition 2.1. *Let $u : [0,T) \times \Omega \to \mathbf{R}^3$ and $P : [0,T) \times \Omega \to \mathbf{R}^1$ satisfy $u \in L^1([0,T) \times \Omega, \mathbf{R}^3)$, $\nabla P \in L^\infty([0,T) \times \Omega) \cap C([0,T); L^1(\Omega))$; and $P(t,\cdot)$ is convex in Ω for every $t \in [0,\infty)$. The pair (u,P) is a weak Eulerian solution of* (2.3) *if*

$$\begin{aligned}\int_{(0,T)\times\Omega} &\{\nabla P(t,x) \cdot [\partial_t \phi(t,x) + (u(t,x)\cdot\nabla)\phi(t,x)] \\ &+ J[\nabla P(t,x) - x]\cdot\phi(t,x)\}dtdx + \int_\Omega \nabla P_0(x)\cdot\phi(0,x)dx = 0\end{aligned} \tag{2.10}$$

for any $\phi \in C^1_c([0,T) \times \Omega; \mathbf{R}^3)$, and

$$\int_{[0,T)\times\Omega} u(t,x)\cdot\nabla\psi(t,x)dtdx = 0 \tag{2.11}$$

for any $\psi \in C^1_c([0,T) \times \overline{\Omega})$.

Remark 2.2. Equality (2.11) is a weak form of the equation $\operatorname{div} u = 0$ in $(0,T)\times\Omega$ with the boundary condition $u_\nu = 0$ on $(0,T) \times \partial\Omega$ in (2.3).

In order to obtain a solution (u,P), we can find P and $x(t,X) = \nabla P^*$ by solving the problem (2.5)–(2.8) in the dual coordinates, using the results of [2], [5]. Thus the problem is to find u such that (2.10), (2.11) hold. Since u in physical coordinates is by definition equal to $D_t x(t,X)$, (2.7) and (2.4) yield the following expression for u in terms of P and P^*:

$$u(t,x) = \partial_t \nabla P^*(t,X) + U\cdot\nabla(\nabla P^*(t,X)) \\ = \partial_t\nabla P^*(t,\nabla P(t,x)) + D^2P^*(t,\nabla P(t,x))[J(\nabla P(t,x)-x)]. \tag{2.12}$$

Formally, if (P,α) satisfy (2.5)–(2.8), P^* is defined by (2.4), and u is defined by (2.12), then (P,u) satisfy (2.3). However, because of the low regularity of $P(t,x)$ obtained as a weak solution of the problem (2.5)–(2.8), it is not clear how to make these calculations rigorous. There is a further difficulty. The product terms $u_i\partial_j P$ that appear in (2.10) are not well defined given that $\nabla P\in L^\infty$ and that u defined by (2.12) is a measure, which is all the regularity of P currently available.

In this paper we circumvent these difficulties by defining Lagrangian solutions of the problem (2.3). We prove existence of such Lagrangian solutions for initial data P_0 satisfying some mild strict convexity conditions. The proof is based on the recent results of L. Ambrosio [1] on transport equations and ODE for BV vector fields.

In the conclusion of this introduction, we show that the time-stepping procedure and estimates of [5] can be applied to the model considered in [2] to improve some estimates of [2]. This is required to allow the solutions in the dual variables to be transferred to real space.

Let $q>1$. Denote by $C([0,T);L^q_w(\mathbf{R}^3))$ the set of all measurable functions $v(t,x)$ on $[0,T)\times\mathbf{R}^3$, such that $v_t(\cdot)=v(t,\cdot)\in L^q(\mathbf{R}^3)$ for any $t\in[0,T)$, and for any $\{t_k\}_{k=1}^\infty, t^*\in[0,T)$ satisfying $\lim_{k\to\infty}t_k=t_*$ there holds $v_{t_k}\rightharpoonup v_{t^*}$ weakly in $L^q(\mathbf{R}^3)$ (weakly-* if $q=\infty$).

THEOREM 2.3 (see [2], [5]). *Let $\Omega\subset\mathbf{R}^3$ be an open bounded set and $\overline{\Omega}\subset B$, where B is an open ball $B(0,S)$. Let $P_0(x)$ be a convex bounded function in B satisfying*

$$\alpha_0 := DP_0\#\chi_\Omega \in L^q(\mathbf{R}^3) \tag{2.13}$$

for some $q>1$. Then for any $T>0$ there exist functions $\alpha(t,X), P(t,X)$ on $[0,T)\times\mathbf{R}^3$ such that

i. *α, P satisfy*

$$\alpha\in L^\infty([0,T);L^q(\mathbf{R}^3))\cap C([0,T);L^q_w(\mathbf{R}^3)),$$
$$P\in L^\infty([0,T);W^{1,\infty}(\Omega))\cap C([0,T);W^{1,r}(\Omega)),$$

where r is any number in $[1,\infty)$;

ii. *let $R_0=S(1+T)$, then*

$$\operatorname{supp}(\alpha(t,\cdot))\subset B(0,R_0) \quad \textit{for all } t\in[0,T); \tag{2.14}$$

iii. *$P(t,\cdot)$ is convex in Ω;*

iv. *P^* defined by (2.4) satisfies*

$$P^*(t,\cdot) \textit{ is convex in } \mathbf{R}^3 \textit{ for any } t\in[0,T),$$
$$P^*\in L^\infty_{loc}([0,T)\times\mathbf{R}^3),$$
$$\nabla P^*\in L^\infty([0,T)\times\mathbf{R}^3;\mathbf{R}^3)\cap C([0,T);L^r(B(0,R);\mathbf{R}^3))$$

for any $R>0$ and any $r\in[1,\infty)$. Moreover,

$$\|\nabla P^*(t,\cdot)\|_{L^\infty(\mathbf{R}^3)}\le S \quad \textit{for every } t\in[0,T); \tag{2.15}$$

v. (α, P, P^*) *satisfy* (2.5)–(2.8)*, where the evolution equation* (2.5) *and the initial condition* (2.8) *are understood in the weak sense: for any* $\phi \in C_c^1([0,T)\times \mathbf{R}^3)$

$$\begin{aligned}\int_{(0,T)\times\mathbf{R}^3}&[\partial_t\phi(t,X)+U(t,X)\cdot\nabla\phi(t,X)]\alpha(t,X)dtdX\\ &+\int_{\mathbf{R}^3}\alpha_0(X)\phi(0,X)dX=0.\end{aligned}\tag{2.16}$$

Proof. We sketch the proof.

Since P_0 is convex and bounded in B, then

$$\|\nabla P_0\|_{L^\infty(\Omega)} \le C(B,\Omega,\|P_0\|_{L^\infty(B)}).\tag{2.17}$$

Thus α_0 defined by (2.13) has compact support.

As in [2], [5], we construct solutions using a time-stepping procedure. Fix $x^* \in \Omega$. Let $h > 0$ be small, chosen so that T/h is an integer. Let $\eta_h(\cdot) = \frac{1}{h^3}\eta(\frac{\cdot}{h})$ be a standard mollifier. Define

$$\alpha_h^0 = (\alpha_0) * \eta_h,$$

and then inductively define the following quantities for $k = 0, 1, \ldots, T/h$.

Suppose $\alpha_h^k \in L^q(\mathbf{R}^3)$, with compact support, is defined. Then let P_h^k be the unique convex function satisfying

$$\nabla P_h^k \# \chi_\Omega = \alpha_h^k, \qquad P_h^k(x^*) = P_0(x^*).$$

Existence and uniqueness of such P_h^k follows from Brenier [3].

$$\begin{aligned}&(P_h^k)^*(X) = \sup_{x\in\Omega}(x\cdot X - P_h^k(x)) \quad \text{for } X \in \mathbf{R}^3,\\ &Q_h^k = \eta_h * (P_h^k)^*,\\ &U_h^k(X) = J\left[X - \nabla Q_h^k(X)\right].\end{aligned}\tag{2.18}$$

Since $\Omega \subset B(0,S)$, it follows that

$$\|\nabla(P_h^k)^*,\ \nabla Q_h^k\|_{L^\infty(\mathbf{R}^3)} \le S.\tag{2.19}$$

From (2.18), (2.19)

$$\|(P_h^k)^* - Q_h^k\|_{L^\infty(\mathbf{R}^3)} \le hS.$$

To define α_h^{k+1} we solve

$$\frac{\partial\alpha_h}{\partial t} + \operatorname{div}(\alpha_h U_h^k) = 0 \qquad \text{in } \mathbf{R}^3\times[kh,(k+1)h],\tag{2.20}$$

$$\alpha_h(kh,X) = \alpha_h^k(X),\tag{2.21}$$

and set

$$\alpha_h^{k+1}(X) = \alpha_h((k+1)h,X).\tag{2.22}$$

Similar to [5, Lemma 4.1] we show that $\alpha_h \in L^\infty([kh,(k+1)h], L^q(\mathbf{R}^3))$ and that

$$\|\alpha_h(t,\cdot)\|_{L^q(\mathbf{R}^3)} = \|\alpha_0\|_{L^q(\mathbf{R}^3)} \qquad \text{for every } t \in [kh,(k+1)h],\tag{2.23}$$

$$\operatorname{supp}(\alpha_h(t,\cdot)) \subset B(0,S(1+(k+1)h)) \qquad \text{for every } t \in [kh,(k+1)h],\tag{2.24}$$

$$W_1(\alpha_h(t_1,\cdot),\alpha_h(t_2,\cdot)) \le C(T)|t_1-t_2| \qquad \text{for all } t_1,t_2 \in [kh,(k+1)h],\tag{2.25}$$

where q is from (2.13), and $W_1(\alpha_h(t_1,\cdot),\alpha_h(t_2,\cdot))$ is the 1-Wasserstein distance between $\alpha_h(t_1,\cdot)$ and $\alpha_h(t_2,\cdot)$ (see, e.g., [5, Appendix A] for the definition of W_1).

We repeat this procedure for $k = 1,\dots,T/h$. In particular, we thus define a function

$$\alpha_h \in L^\infty([0,T],L^q(\mathbf{R}^3)). \tag{2.26}$$

For every $t \in (0,T)$ define $P_h(t,\cdot)$ to be the unique convex function satisfying

$$\nabla P_h(t,\cdot)\#\chi_\Omega = \alpha_h(t,\cdot), \qquad P_h(t,x^*) = P_0(x^*),$$

and then define $Q_h(t,\cdot)$, $U_h(t,\cdot)$ by using the $P_h(t,\cdot)$ in (2.18) instead of P_h^k.

We also define functions $\overline{P}_h$ on $[0,T]\times\Omega$ and $\overline{\alpha}_h$, $\overline{Q}_h$, $\overline{U}_h$ on $[0,T]\times\mathbf{R}^3$ by setting them equal to $P_h^k,\alpha_h^k,Q_h^k,U_h^k$, respectively, on the time interval $t\in[kh,(k+1)h)$.

We note the following fact.

Lemma 2.4 (see [3]). *Let $p \geq 1$. Let $\Omega\subset\mathbf{R}^n$ be an open bounded set, $x^*\in\Omega$, $a\in\mathbf{R}$. Let $\rho_j,\rho\in L^p(\mathbf{R}^n)$ for $j=1,\dots$ be such that $\rho_j\rightharpoonup\rho$ weakly in $L^p(\mathbf{R}^n)$ and $\operatorname{supp}(\rho_j,\rho)\subset B(0,R)$. Let ϕ_j, ϕ be the convex functions on $\mathbf{R}^n$ satisfying $\nabla\phi_j\#\chi_\Omega = \rho_j$, $\nabla\phi\#\chi_\Omega=\rho$, $\phi(x^*),\phi_j(x^*)=a$ for $j=1,\dots$. Then $\phi_j\to\phi$ in $W^{1,r}(\Omega)$ for any $r\in[1,\infty)$.*

From (2.23)–(2.25), repeating the argument of [5, pp. 263–268], with the use of Lemma 2.4 we obtain a sequence $h_j\to 0+$ and a function $\alpha\in L^q([0,T]\times\mathbf{R}^3)$, where q is from (2.13), such that

$$\operatorname{supp}(\alpha_{h_j}(t,\cdot))\subset B(0,S(1+T)) \quad \text{for each } t\in[0,T],\ \ j=1,2,\dots, \tag{2.27}$$

$$\alpha_{h_j}\rightharpoonup\alpha \quad \text{weakly in } L^q([0,T]\times\mathbf{R}^3), \tag{2.28}$$

$$\alpha_{h_j}(t,\cdot)\rightharpoonup\alpha(t,\cdot) \quad \text{weakly in } L^q(\mathbf{R}^3) \quad \text{for each } t\in[0,T], \tag{2.29}$$

and such that, denoting by $P(t,\cdot)$, for every $t\in(0,T)$, the unique convex function satisfying

$$\nabla P(t,\cdot)\#\chi_\Omega = \alpha(t,\cdot), \qquad P(t,x^*) = P_0(x^*),$$

and denoting by $P^*(t,\cdot)$ the convex dual of $P(t,\cdot)$ defined by (2.4), we have

$$\overline{P}^*_{h_j}(t,\cdot)\to P^*(t,\cdot) \quad \text{in } C(\overline{B(0,R)}) \quad \text{for each } R>0,\ t\in[0,T], \tag{2.30}$$

$$\overline{Q}_{h_j}(t,\cdot)\to P^*(t,\cdot) \quad \text{in } C(\overline{B(0,R)}) \quad \text{for each } R>0,\ t\in[0,T], \tag{2.31}$$

$$\alpha_{h_j}\overline{U}_{h_j}\rightharpoonup\alpha J(id-\nabla P^*) \quad \text{weakly in } L^q([0,T]\times\mathbf{R}^3;\mathbf{R}^3). \tag{2.32}$$

Then the proof of Theorem 2.3 is completed as in [5, pp. 268–269]. □

2.2. Statement of results. In this paper we study the system in the "physical" space (t,x) and define its weak Lagrangian solutions. In order to do that, we first rewrite system (2.3) in terms of F,P, where $F:[0,T]\times\Omega\to\Omega$ is the (formal) Lagrangian flow corresponding to the full wind velocity $u=(u_1,u_2,u_3)$, and then we define the corresponding weak solution F,P. This gives the following.

Definition 2.5. *Let $\Omega\subset\mathbf{R}^3$ be an open bounded set, and let $T>0$. Let $P_0(x)\in W^{1,\infty}(\Omega))$ be convex. Let $r\in[1,\infty)$. Let $P:[0,T)\times\Omega\to\mathbf{R}^1$ satisfy*

$$P\in L^\infty([0,T);W^{1,\infty}(\Omega))\cap C([0,T);W^{1,r}(\Omega)), \tag{2.33}$$

$$P(t,\cdot) \quad \textit{is convex in } \ \Omega \ \textit{ for each } \ t\in[0,T). \tag{2.34}$$

Let $F : [0,T) \times \Omega \to \Omega$ be a Borel map satisfying

$$F \in C([0,T); L^r(\Omega; \mathbf{R}^3)). \tag{2.35}$$

Then the pair (P, F) is called a weak Lagrangian solution of (2.3) in $[0,T) \times \Omega$ if (P, F) has the following properties:

i. *$F(0,x) = x$, $P(0,x) = P_0(x)$ for a.e. $x \in \Omega$.*
ii. *For any $t > 0$ the mapping $F_t = F(t,\cdot) : \Omega \to \Omega$ is Lebesgue measure preserving, in the sense that $F_t \# \chi_\Omega = \chi_\Omega$.*
iii. *There exists a Borel map $F^* : [0,T) \times \Omega \to \Omega$ such that for every $t \in (0,T)$ the map $F_t^* = F^*(t,\cdot) : \Omega \to \Omega$ is Lebesgue measure preserving: $F_t^* \# \chi_\Omega = \chi_\Omega$ and satisfies $F_t^* \circ F_t(x) = x$, and $F_t \circ F_t^*(x) = x$ for a.e. $x \in \Omega$.*
iv. *The function*

$$Z(t,x) = \nabla P(t, F_t(x)) \tag{2.36}$$

is a weak solution of

$$\begin{aligned} &\partial_t Z(t,x) = J[Z(t,x) - F(t,x)] \qquad \text{in } [0,T) \times \Omega, \\ &Z(0,x) = \nabla P_0(x) \qquad \text{in } \Omega \end{aligned} \tag{2.37}$$

in the following sense: for any $\varphi \in C_c^1([0,T) \times \Omega; \mathbf{R}^3)$

$$\begin{aligned} \int_{(0,T)\times\Omega} &[Z(t,x) \cdot \partial_t \varphi(t,x) + J(Z(t,x) - F(t,x)) \cdot \varphi(t,x)] dt dx \\ &+ \int_\Omega \nabla P_0(x) \cdot \varphi(0,x) dx = 0. \end{aligned} \tag{2.38}$$

Remark 2.6. We comment on Definition 2.5:

- Continuity in time of P, F, considered as maps on $[0,T)$ with values in $W^{1,r}(\Omega)$ and $L^r(\Omega)$, respectively, required in (2.33), (2.35), combined with initial conditions in Definition 2.5(i), imply that

$$\lim_{t\to 0+} \|F_t - Id\|_{L^r(\Omega)} = 0, \qquad \lim_{t \to 0+} \|P_t - P_0\|_{W^{1,r}(\Omega)} = 0, \tag{2.39}$$

  where $Id : \Omega \to \Omega$ is the identity mapping. Furthermore, the continuity property of F in (2.35) can be interpreted as "generic continuity" of particle paths in physical space.
- The property Definition 2.5(ii) of the flow F is the Lagrangian form of the equation $\operatorname{div} u = 0$ in $(0,T) \times \Omega$ with the boundary condition $u_\nu = 0$ on $(0,T) \times \partial\Omega$ in (2.3).
- (2.38) is a weak Lagrangian form of the first equation of (2.3) with an initial condition for P.

Remark 2.7 (semigroup property). The weak Lagrangian solution in Definition 2.5 satisfies the following semigroup property. For every $t_1, t_2 \geq 0$ define $F_{(t_1,t_2)} = F_{t_1} \circ F_{t_2}^*$, where F_t^* is defined as in Definition 2.5(iii). Then for any $t_1, t_2, t_3 \geq 0$

$$F_{(t_1,t_2)} \circ F_{(t_2,t_3)} = F_{(t_1,t_3)} \qquad \text{for a.e. } x \in \Omega. \tag{2.40}$$

This follows from the property $F_t^* \circ F_t(x) = x$ for a.e. $x \in \Omega$ in Definition 2.5(iii), and from $\mathcal{L}^3$-measure-preserving properties of maps F_t, F_t^*. Indeed, it follows that

$$F_{t_1} \circ F_{t_3}^*(x) = F_{t_1} \circ F_{t_2}^* \circ F_{t_2} \circ F_{t_3}^*(x) = F_{t_1} \circ F_{t_2}^* \circ F_{t_2} \circ F_{t_3}^* \circ F_{t_3} \circ F_{t_3}^*(x)$$

for a.e. $x \in \Omega$, which is (2.40).

To justify Definition 2.5 we now show that a weak Lagrangian solution (F, P) with the additional regularity property $\partial_t F \in L^\infty([0,T) \times \Omega)$ determines a weak Eulerian solution of (2.3), and that a smooth Lagrangian solution determines a classical solution of (2.3).

Lemma 2.8 (consistency of weak Lagrangian solutions). *Let $\Omega \subset \mathbf{R}^3$ be an open bounded set, and let $T > 0$. Let (F, P) be a weak Lagrangian solution of* (2.3) *in* $[0,T) \times \Omega$.

i. *If $\partial_t F \in L^\infty([0,T) \times \Omega; \mathbf{R}^3)$, then the function*

$$u(t,x) := (\partial_t F)(t, F_t^*(x)) \tag{2.41}$$

satisfies $u \in L^\infty([0,T) \times \Omega; \mathbf{R}^3)$, and (u, P) is a weak Eulerian solution of (2.3) *in* $[0,T) \times \Omega$ *in the sense of Definition* 2.1*;*

ii. *If $(F, F^*, P) \in C^2([0,T] \times \overline{\Omega})$, then the function* (2.41) *satisfies $u \in C^1([0,T] \times \overline{\Omega}; \mathbf{R}^3)$, and (u, P) is a classical solution of* (2.3) *in* $[0,T) \times \Omega$.

Proof. We first prove (i). Since F^* is a Borel map, and $\partial_t F \in L^\infty([0,T) \times \Omega)$, the right-hand side of (2.41) is a bounded measurable function, and thus $u \in L^\infty([0,T) \times \Omega)$. It remains to prove that (P, u) is a weak Eulerian solution. We prove first that (2.11) holds. Let $\psi \in C_c^1([0,T) \times \overline{\Omega})$. Fix $t \in (0,T)$. Since $F_t \# \chi_\Omega = \chi_\Omega$, then

$$\int_\Omega (\partial_t \psi)(t, F_t(x)) dx = \int_\Omega \partial_t \psi(t,x) dx.$$

Integrating with respect to t and using $\partial_t F \in L^\infty([0,T) \times \Omega)$, we get

$$\int_{[0,T)\times\Omega} \{\partial_t[\psi(t, F_t(x))] - \partial_t F_t(x) \cdot (\nabla\psi)(t, F_t(x))\} dt dx = \int_{[0,T)\times\Omega} \partial_t \psi(t,x) dt dx.$$

Using $\psi(T, \cdot) \equiv 0$ and (2.39), we get

$$\int_{[0,T]\times\Omega} \partial_t F_t(x) \cdot (\nabla\psi)(t, F_t(x)) dt dx = 0.$$

Making the change of variables $y = F_t(x)$ and using properties (ii) and (iii) of Definition 2.5, we get

$$\int_{[0,T]\times\Omega} (\partial_t F_t)(F_t^* y) \cdot \nabla\psi(t,y) dt dy = 0.$$

Since (2.41) defines u, then (2.11) follows.

Now we prove (2.10). From properties P and F in Definition 2.5 it follows that $Z(t,x)$ defined by (2.36) satisfies $Z \in L^\infty([0,T] \times \overline{\Omega})$. Then, since Ω is a bounded set and $F_t \# \chi_\Omega = \chi_\Omega$ for all $t \in [0,T)$, (A.2) of Corollary A.3 allows us to make the change of variables $y = F_t(x)$ in the first integral of (2.38). Then $x = F_t^*(y)$ for a.e. $(t,x) \in [0,T) \times \Omega$ by (iii) in Definition 2.5, and from (2.38) we get for any $\varphi \in C_c^1([0,T) \times \Omega)$

$$\begin{aligned}\int_{(0,T)\times\Omega} [\nabla P(t,y) \cdot (\partial_t \varphi)(t, F_t^*(y)) + J(\nabla P(t,y) - y)\varphi(t, F_t^*(y))] dt dy \\ + \int_\Omega \nabla P_0(x) \varphi(0,x) dx = 0.\end{aligned} \tag{2.42}$$

Next we show that (2.42) holds for all $\varphi \in L^\infty([0,T)\times\Omega)$ satisfying $\partial_t\varphi \in L^\infty([0,T)\times\Omega)$ and $\mathrm{supp}(\varphi)\subset[0,T-\varepsilon]\times\overline{\Omega}$ for some $\varepsilon>0$. Indeed, for such φ we construct an approximating sequence $\varphi_j \in C^1_c([0,T)\times\Omega)$ as follows. Extend φ to $[0,\infty)\times\Omega$ by defining $\varphi(t,\cdot)\equiv 0$ for $t\ge T$, and further extend φ to $(-\infty,\infty)\times\Omega$ by defining $\varphi(t,x)=\varphi(-t,x)$ for $t<0$, $x\in\Omega$. Let $h>0$ and $\Omega_h=\{x\in\Omega \mid \mathrm{dist}(x,\partial\Omega)>h\}$. Now $\varphi\chi_{\Omega_h}$ is defined on $\mathbf{R}^1\times\mathbf{R}^3$. Let $\eta_h(t,x)=\frac{1}{h^4}\eta(\frac{|(t,x)|}{h})$, where $\eta(\cdot)$ is a standard mollifier. Let $j>\frac{1}{\varepsilon}$ be integer. Then functions $\varphi_j=(\varphi\chi_{\Omega_{4h}})*\eta_h$ with $h=\frac{1}{j}<\varepsilon$ satisfy $\varphi_j\in C^1_c([0,T)\times\Omega)$ with $\|\varphi_j,\partial_t\varphi_j\|_{L^\infty([0,T)\times\Omega)}\le C$, where C does not depend on j, and $(\varphi_j,\partial_t\varphi_j)\to(\varphi,\partial_t\varphi)$ a.e. on $[0,T)\times\Omega$ as $j\to\infty$. Since $F^*_t\#\chi_\Omega=\chi_\Omega$ for all t, it follows that $(\varphi_j,\partial_t\varphi_j)(t,F^*_t(y))\to(\varphi,\partial_t\varphi)(t,F^*_t(y))$ for a.e. $(t,y)\in[0,T)\times\Omega$. With this, since Ω is bounded, $\nabla P\in L^\infty([0,T)\times\Omega)$, and (2.42) holds for each φ_j; the bounded convergence theorem implies (2.42) for φ.

Let $\varphi(t,x)=\eta(t,F_t(x))$, where $\eta\in C^1_c([0,T)\times\Omega)$. Then $\varphi\in L^\infty([0,T)\times\Omega)$, and $\mathrm{supp}(\varphi)\subset[0,T-\varepsilon]\times\overline{\Omega}$ since $\mathrm{supp}(\eta)\subset[0,T-\varepsilon]\times\overline{\Omega}$ for some $\varepsilon>0$. Moreover, since $\partial_tF\in L^\infty([0,T)\times\Omega)$, it follows that $\partial_t\varphi\in L^\infty([0,T)\times\Omega)$ with

$$\partial_t\varphi(t,x)=(\partial_t\eta)(t,F_t(x))+\sum_{i=1}^{3}\partial_tF^j_t(x)(\partial_{x_j}\eta)(t,F_t(x)) \qquad \text{a.e. in} \quad [0,T)\times\Omega,$$

where we used notation $F(t,x)=(F^1,F^2,F^3)(t,x)$.

Thus (2.42) holds for φ. By properties (ii) and (iii) in Definition 2.5

$$\varphi(t,F^*_t(y))=\eta(t,y), \qquad \partial_t\varphi(t,F^*_t(y))=\partial_t\eta(t,y)+[(\partial_tF)(t,F^*_t(y))]\cdot\nabla\eta(t,y)$$

for a.e. $(t,y)\in[0,T)\times\Omega$. Thus, inserting $\varphi(t,x)=\eta(t,F_t(x))$ into (2.42), we obtain for every $\eta\in C^1_c([0,T)\times\Omega)$

$$\int_{(0,T)\times\Omega}[\nabla P(t,y)\cdot(\partial_t\eta(t,y)+[(\partial_tF)(t,F^*_t(y))]\cdot\nabla\eta(t,y))$$
$$+J(\nabla P(t,y)-y)\eta(t,y)]dtdy+\int_\Omega\nabla P_0(x)\eta(0,F_0(x))dx=0.$$

Using property (i) in Definition 2.5 and (2.41), and changing notations y to x and η to φ, we obtain (2.10). Assertion (i) of Lemma 2.8 is proved.

Now assertion (ii) follows directly from (i). ◻

Our main result is the following.

THEOREM 2.9. *Let $\Omega\subset\mathbf{R}^3$ be an open bounded set and $\overline{\Omega}\subset B$ where B is an open ball $B(0,S)$. Let $P_0(x)$ be a convex bounded function in B. Assume that P_0 satisfies*

$$DP_0\#\chi_\Omega\in L^q(\mathbf{R}^3) \tag{2.43}$$

for some $q>1$. Then for any $T>0$ there exists a weak Lagrangian solution (P,F) of (2.3) *in $[0,T)\times\Omega$, where* (2.33), (2.35) *are satisfied for any $r\in[1,\infty)$. Moreover, the function $Z(t,x)$ defined by* (2.36) *satisfies $Z(\cdot,x)\in W^{1,\infty}([0,T);\mathbf{R}^3)$ for a.e. $x\in\Omega$, and* (2.37) *is satisfied, in addition to the weak form* (2.38)*, in the following sense:*

$$\begin{aligned}&\partial_tZ(t,x)=J(Z(t,x)-F(t,x)) \qquad \textit{for } \mathcal{L}^4\textit{-a.e. in } (t,x)\in(0,T)\times\Omega,\\ &Z(0,x)=\nabla P_0(x) \qquad \textit{for } \mathcal{L}^3\textit{-a.e. in } x\in\Omega.\end{aligned} \tag{2.44}$$

Remark 2.10. The condition (2.43) is equivalent to

$$\det D^2P^*_0\in L^q(\nabla P_0(\Omega)),$$

where P_0^* is the Legendre–Fenchel transform of P_0, i.e., $P_0^*(X) = \sup_{x\in\Omega}[x\cdot X - P_0(x)]$ for $X \in \mathbf{R}^3$. The condition (2.43) is a certain strict convexity condition for P_0. In particular, if P_0 is uniformly strictly convex in B, in the sense that there exists $\varepsilon > 0$ such that $P_0(x) - \varepsilon|x|^2$ is a convex function in B, then $\det D^2 P_0^* \in L^\infty(\nabla P_0(\Omega))$, and $\nabla P_0(\Omega)$ is a bounded set by (2.17). Thus (2.43) is satisfied.

We prove Theorem 2.9 in sections 2.3 and 2.4.

2.3. Lagrangian flow in the dual space. Let Ω, T, P_0 be as in Theorem 2.9.

Let $\alpha_0 = \nabla P_0 \# \chi_\Omega$. By Theorem 2.3, there exists a solution (α, P, P^*) of the problem (2.5)–(2.7) with initial data α_0 satisfying all assertions of Theorem 2.3.

Note that the vector field U defined by (2.7) is divergence free. Thus the evolution equation (2.5) and its weak form (2.16) can be seen as the transport equation

$$\partial_t \alpha + U \cdot \nabla \alpha = 0. \tag{2.45}$$

Since $\nabla P^* \in L^\infty([0,T) \times \mathbf{R}^3))$, and $P_t^*(\cdot)$ is convex in $\mathbf{R}^3$ for all $t \in [0,T)$, it follows from (2.7) that

$$U \in L^\infty_{loc}([0,T) \times \mathbf{R}^3) \text{ and } U \in L^\infty([0,T); BV(B(0,R))) \quad \text{for all } R > 0.$$

By (2.14) α has compact support in $[0,T] \times \mathbf{R}^3$. Thus we can modify U away from $B(0,R_1)$, where large R_1 will be chosen below so that, in particular, $\text{supp}(\alpha) \subset [0,T] \times \overline{B(0,R_1)}$ and the modified function $\tilde{U}$ satisfies

$$\begin{aligned} &\tilde{U} \in L^\infty([0,T) \times \mathbf{R}^3), \qquad \tilde{U} \in L^\infty([0,T); BV(B(0,R))) \quad \text{for all } R > 0; \\ &\operatorname{div} \tilde{U}(t,\cdot) = 0 \quad \text{in } \mathbf{R}^3 \quad \text{for all } t \in [0,T). \end{aligned} \tag{2.46}$$

To construct $\tilde{U}$ satisfying (2.46), we choose $\zeta \in C^\infty(\mathbf{R}^1)$ such that

$$\zeta \equiv 1 \quad \text{on } \{|s| < R_1\}, \quad \zeta \equiv 0 \quad \text{on } \{|s| > R_1 + 1\}, \quad 0 \le \zeta \le 1 \text{ on } \mathbf{R}^1, \tag{2.47}$$

and define, for $X \in \mathbf{R}^3$,

$$H(X) = (\zeta(|X_1|)X_1, \zeta(|X_2|)X_2, \zeta(|X_3|)X_3). \tag{2.48}$$

Define $\tilde{U}$ by

$$\tilde{U}(t,X) = J\,[H(X) - \nabla P^*(t,X)]. \tag{2.49}$$

Then $\tilde{U}$ satisfies (2.46), and from (2.47)–(2.48), (2.15)

$$U = \tilde{U} \quad \text{in } B(0,R_1), \tag{2.50}$$

$$\|U(t,\cdot)\|_{L^\infty(\mathbf{R}^3)} \le S + R_1 + 1 \quad \text{for all } t \in [0,T). \tag{2.51}$$

Now the theory developed by Ambrosio [1] applies to (2.45) with U replaced by $\tilde{U}$.

LEMMA 2.11. *Let $\tilde{U}$ be defined by* (2.47)–(2.49). *There exists a unique locally bounded Borel measurable map $\Phi : [0,T) \times \mathbf{R}^3 \to \mathbf{R}^3$ satisfying*

i. $\Phi(\cdot, X) \in W^{1,\infty}([0,T); \mathbf{R}^3)$ *for a.e.* $X \in \mathbf{R}^3$*;*
ii. $\Phi(0,X) = X$ *for* $\mathcal{L}^3$*-a.e.* $X \in \mathbf{R}^3$*;*
iii. *for a.e.* $(t,X) \in \mathbf{R}^3 \times (0,T)$

$$\partial_t \Phi(t,X) = \tilde{U}(t, \Phi(t,X)); \tag{2.52}$$

iv. $\Phi(t,\cdot) : \mathbf{R}^3 \to \mathbf{R}^3$ *is a* $\mathcal{L}^3$*-measure-preserving map for every* $t \in [0,T)$.

Proof. Since $\tilde{U}$ satisfies (2.46), by [1, section 6] there exists a unique regular flow relative to $\tilde{U}$. This flow determines a locally bounded Borel measurable map $\Phi : [0,T) \times \mathbf{R}^3 \to \mathbf{R}^3$ satisfying (i)–(iii) and the following property.

Let $\tilde{U}_\varepsilon : [0,T) \times \mathbf{R}^3 \to \mathbf{R}^3$ be a family of approximations to $\tilde{U}$ satisfying the following:

$$\begin{aligned}
&\tilde{U}_\varepsilon \in C([0,T) \times \mathbf{R}^3; \mathbf{R}^3), \qquad \sup_\varepsilon \|\tilde{U}_\varepsilon\|_{L^\infty([0,T)\times\mathbf{R}^3;\mathbf{R}^3)} < \infty,\\
&\tilde{U}_\varepsilon \to \tilde{U} \quad \text{in } \ L^1_{loc}((0,T) \times \mathbf{R}^3; \mathbf{R}^3),\\
&\|\nabla \tilde{U}_\varepsilon\|_{L^\infty([0,T)\times B_R;\mathbf{R}^3)} \le C(\varepsilon, R) < \infty \quad \text{for any } \ \varepsilon, R > 0,\\
&\operatorname{div} \tilde{U}_\varepsilon = 0.
\end{aligned} \tag{2.53}$$

Let $\Phi_\varepsilon(t,X)$ be the unique solution in $[0,T]$ of the ODE $\frac{d}{dt}\Phi_\varepsilon(t,X) = \tilde{U}(t, \Phi_\varepsilon(t,X))$ with the initial condition $\Phi_\varepsilon(0,X) = X$. Then

$$\lim_{\varepsilon\to 0} \int_{B_R} \sup_{t\in[0,T]} |\Phi(t,X) - \Phi_\varepsilon(t,X)|\, dX = 0 \quad \text{for any } \ R > 0. \tag{2.54}$$

Note that a family $\tilde{U}_\varepsilon$ satisfying the conditions stated above exists. Indeed, let $\eta_\varepsilon : \mathbf{R}^3 \times \mathbf{R}^1 \to \mathbf{R}^1$ be a family of mollifiers. Extend $P^*(t,X)$ to the time interval $(-\infty,\infty)$ by setting $P^*(t,X) = P^*(0,X)$ for $t < 0$ and $P^*(t,X) = P^*(T,X)$ for $t > 0$, and let

$$\tilde{U}_\varepsilon(t,X) = J\left[H(X) - (\nabla P^* * \eta_\varepsilon)(t,X)\right],$$

where the convolution is with respect to (t,X)-variables. Then all properties in (2.53) are satisfied.

Also, since $\tilde{U}_\varepsilon$ is smooth and $\operatorname{div} \tilde{U}_\varepsilon = 0$, it follows that each $\Phi_\varepsilon(t,\cdot)$ is a measure-preserving diffeomorphism. Thus (2.54) implies (iv).

Conversely, a map Φ satisfying (i)–(iv) determines a regular Lagrangian flow relative to $\tilde{U}$. This implies uniqueness of Φ by Theorem 6.4 of [1]. ∎

LEMMA 2.12. *Let* $\tilde{U}$ *be defined by* (2.47)–(2.49)*, and let* Φ *be the map defined in Lemma* 2.11. *Then we have the following conditions:*

i. *If* R_1 *in the definition of* $\tilde{U}$ *is chosen sufficiently large depending only on* Ω*,* T *and* $\|\nabla P_0\|_{L^\infty(\Omega)}$*, then*

$$\Phi(t,X) \subset B(0,R_1) \quad \text{for a.e. } \ (t,X) \in [0,T) \times \nabla P_0(\Omega). \tag{2.55}$$

In particular,

$$\partial_t \Phi(t,X) = U(t, \Phi(t,X)) \quad \text{for a.e. } \ (t,X) \in [0,T) \times \nabla P_0(\Omega). \tag{2.56}$$

ii. *There exists a Borel map* $\Phi^* : [0,T) \times \mathbf{R}^3 \to \mathbf{R}^3$ *such that for every* $t \in (0,T)$ *the map* $\Phi_t^* : \mathbf{R}^3 \to \mathbf{R}^3$ *is* $\mathcal{L}^3$*-measure preserving, and such that* $\Phi_t^* \circ \Phi_t(x) = x$ *and* $\Phi_t \circ \Phi_t^*(x) = x$ *for a.e.* $x \in \mathbf{R}^3$.

Proof. We prove (i). For the family $\tilde{U}_\varepsilon$ satisfying (2.53) constructed in the proof of Lemma 2.11, we have for $0 < \varepsilon < 1/2$, using (2.15),

$$|\tilde{U}_\varepsilon(t,X)| \le |X| + 1 + S. \tag{2.57}$$

Thus

$$\frac{1}{2}\frac{d}{dt}|\Phi_\varepsilon(t,X)|^2 = \tilde{U}(t, \Phi_\varepsilon(t,X)) \cdot \Phi_\varepsilon(t,X) \le \frac{3}{2}|\Phi_\varepsilon(t,X)|^2 + \frac{1}{2}(1+S)^2.$$

Thus, by the Gronwall inequality,

$$|\Phi_\varepsilon(t,X)| \le e^{3T/2}\left(|X|^2+(1+S)^2T^2\right)^{\frac{1}{2}} \qquad \text{for all } (t,X)\in[0,T]\times\mathbf{R}^3. \tag{2.58}$$

The same is true for $|\Phi(t,X)|$ for a.e. $(t,X)\in[0,T]\times\mathbf{R}^3$. Now choosing

$$R_1 = 2e^{3T/2}\left(\|\nabla P_0\|^2_{L^\infty(\Omega)}+(1+S)^2T^2\right)^{1/2}+1, \tag{2.59}$$

we obtain (2.55), which implies (2.56).

It remains to prove (ii). This is in fact a general property of regular Lagrangian flows constructed by Ambrosio [1, section 6]. It can be seen as follows. For $t_1,t_2\in[0,T]$ with $t_1\ge t_2$ and $x\in\mathbf{R}^3$, denote by $\hat{\Phi}(t_1,t_2,x)$ the regular Lagrangian flow relative to $\tilde{U}$ and starting at time t_1; i.e., $\hat{\Phi}(t_1,\cdot,\cdot)$ satisfies properties (i), (iii), (iv) of Lemma 2.11 on time interval $[t_1,T]$ and $\hat{\Phi}(t_1,t_1,X)=X$ for $\mathcal{L}^3$-a.e. $X\in\mathbf{R}^3$. In the case $t_2\le t_1$, $\hat{\Phi}(t_1,t_2,x)$ denotes the regular flow Lagrangian flow relative to $\tilde{U}$ backwards in time; i.e., $\hat{\Phi}(t_1,\cdot,\cdot)$ satisfies properties (i), (iii), (iv) of Lemma 2.11 on time interval $[0,t_1]$ and $\hat{\Phi}(t_1,t_1,X)=X$ for $\mathcal{L}^3$-a.e. $X\in\mathbf{R}^3$. Repeating the argument of Remark 6.7 of [1], we see that the following semigroup property holds: for every $t_1,t_2,t_3\in[0,T]$

$$\hat{\Phi}(t_1,t_2,\hat{\Phi}(t_2,t_3,x))=\hat{\Phi}(t_1,t_3,x) \qquad \text{for a.e. } x\in\mathbf{R}^n.$$

Also, from the proof of Theorem 6.6 of [1], one can see that if $\tilde{U}_\varepsilon$ is a family of approximations to $\tilde{U}$ satisfying (2.53), and if $\hat{\Phi}^\varepsilon(t_1,t_2,x)$ is the regular Lagrangian flow of $\tilde{U}_\varepsilon$ starting at time t_1, then $\hat{\Phi}^\varepsilon$ converges to $\hat{\Phi}$ in $L^1_{loc}([0,T]\times[0,T]\times\mathbf{R}^3)$. Thus, possibly after a modification on a negligible set, $\hat{\Phi}$ is a Borel map since it is an a.e. limit of continuous maps. Since $\Phi=\hat{\Phi}(0,\cdot,\cdot)$, the map $\Phi^*(t,x)=\hat{\Phi}(t,0,x)$ satisfies all properties asserted in (ii). $\square$

Now we prove the following property of transport equations.

LEMMA 2.13. *Let* $\tilde{U}\in L^\infty((0,T)\times\mathbf{R}^3)$ *with* $\tilde{U}(t,\cdot)\in BV_{loc}(\mathbf{R}^3)$ *and* $\operatorname{div}\tilde{U}(t,\cdot)=0$ *in* $\mathbf{R}^3$ *in the sense of distributions for every* $t\in(0,T)$. *Let a locally bounded Borel measurable map* $\Phi:\mathbf{R}^3\times[0,T)\to\mathbf{R}^3$ *satisfy properties* (i)–(iv) *of Lemma* 2.11. *Let* $q\in(1,\infty]$, $v_0\in L^q(\mathbf{R}^3)$, *and* $\operatorname{supp}(v_0)\subset B(0,R)$. *Let* $v(t,x):=(\Phi_t\#v_0)(x)$. *Then*

$$v\in L^\infty((0,T);L^q(\mathbf{R}^3))\cap C([0,T];L^q_w(\mathbf{R}^3)), \quad \textit{supp}\ v\subset B(0,R_T)\times[0,T],$$

where $R_T=R+T\|\tilde{U}\|_{L^\infty}$, *and* v *is a weak solution of*

$$\partial_t v+\operatorname{div}(v\tilde{U})=0 \tag{2.60}$$

on $(0,T)\times\mathbf{R}^3$, *with initial data* $v_{|t=0}=v_0$, *in the sense that for any* $\varphi\in C^1_c([0,T)\times\mathbf{R}^3)$

$$\int_{[0,T)\times\mathbf{R}^3} v\left(\partial_t\varphi+\tilde{U}\cdot\nabla\varphi\right)dtdx+\int_{\mathbf{R}^3}v_0(x)\varphi(0,x)dx=0.$$

Proof. Let p be such that $\frac{1}{p}+\frac{1}{q}=1$ (where $p=1$ if $q=\infty$). Fix $t\in[0,T)$. From the definition of v and using that Φ_t is $\mathcal{L}^3$-measure preserving and locally bounded, we get for all $\varphi\in C_c(\mathbf{R}^3)$

$$\begin{aligned}\int_{\mathbf{R}^3}\varphi(x)v_t(x)dx &= \int_{\mathbf{R}^3}\varphi(\Phi_t(x))v_0(x)dx \le \|v_0\|_{L^q}\left[\int_{\mathbf{R}^3}\left|\varphi(\Phi_t(x))\right|^p dx\right]^{\frac{1}{p}}\\ &= \|v_0\|_{L^q}\left[\int_{\mathbf{R}^3}|\varphi(x)|^p dx\right]^{\frac{1}{p}} = \|v_0\|_{L^q}\|\varphi\|_{L^p}.\end{aligned}$$

Thus $v_t \in L^q(\mathbf{R}^3)$ with $\|v_t\|_{L^q} \le \|v_0\|_{L^q}$ for any t.

Also, weak continuity of v with respect to t also follows: if $\{t_k\}_{k=1}^\infty, t_* \in [0,T)$, and $\lim_{k\to\infty} t_k = t_*$, then we get for every $\varphi \in C_c(\mathbf{R}^3)$ using (i), (iv) of Lemma 2.11 and the dominated convergence theorem

$$\int_{\mathbf{R}^3} v_{t_k}(x)\varphi(x)dx = \int_{\mathbf{R}^3} v_0(x)\varphi(\Phi_{t_k}(x))dx \to \int_{\mathbf{R}^3} v_0(x)\varphi(\Phi_{t_*}(x))dx = \int_{\mathbf{R}^3} v_{t_*}(x)\varphi(x)dx.$$

The bounds on the support of v_t follow from $|\Phi_t(X)| \le |X| + T\|\tilde U\|_{L^\infty}$, which can be obtained by approximating $\tilde U$ by smooth vector fields as in the proof of Lemmas 2.11 and 2.12.

Now we prove that v is a weak solution of (2.60) with initial data $v_{|t=0} = v_0$. Let $\varphi \in C_c^\infty(\mathbf{R}^3 \times [0,T))$. Then using the definition $v_t = \Phi_t \# v_0$ of v, the properties of Φ stated in (i)–(iv) of Lemma 2.11, we get

$$\begin{aligned}
\int_{\mathbf{R}^3\times[0,T)} v\partial_t\varphi dtdx &= \int_{\mathbf{R}^3\times[0,T)} v_0(x)(\partial_t\varphi)(t,\Phi_t(x))dtdx \\
&= \int_{\mathbf{R}^3} v_0(x)\int_0^T \frac{d}{dt}\left[\varphi(t,\Phi_t(x))\right]dtdx \\
&\quad - \int_{\mathbf{R}^3\times[0,T)} v_0(x)\nabla\varphi(t,\Phi_t(x))\cdot\partial_t\Phi_t(x)dtdx \\
&= -\int_{\mathbf{R}^3} v_0(x)\varphi(0,x)dx - \int_{\mathbf{R}^3\times[0,T)} v_0(x)\nabla\varphi(t,\Phi_t(x))\cdot\tilde U(t,\Phi_t(x))dx \\
&= -\int_{\mathbf{R}^3} v_0(x)\varphi(0,x)dx - \int_{\mathbf{R}^3\times[0,T)} v_t(x)\nabla\varphi(t,x)\cdot\tilde U(t,x)dx.
\end{aligned}$$

Lemma 2.13 is proved. ∎

Now we prove that the solution (α, P) of (2.5)–(2.8) satisfies the property that α is a Lagrangian solution of the transport equation (2.5), in the sense of (2.61).

PROPOSITION 2.14. *Let Ω, T, q, P_0 be as in Theorem* 2.3. *Let (α, P) be the weak solution of* (2.5)–(2.8) *constructed in Theorem* 2.3. *Let $\tilde U$ be defined by* (2.47)–(2.49), *and let Φ be the regular Lagrangian flow of $\tilde U$ defined in Lemma* 2.11. *If R_1 in the definition of $\tilde U$ is chosen sufficiently large depending only on Ω, T, and $\|\nabla P_0\|_{L^\infty(\Omega)}$, then for every $t \in [0,T]$*

$$\alpha_t = \Phi_t \# \alpha_0. \tag{2.61}$$

Moreover, for every $t \in [0,T]$

$$\alpha_t(x) = \alpha_0(\Phi_t^*(x)) \quad \textit{for a.e. } x \in \mathbf{R}^3, \tag{2.62}$$

where the map Φ_t^* is defined in Lemma 2.12(ii).

Proof. We choose R_1 to be defined by (2.59). Then, in particular, $R_1 \ge R_0 + 1$ where $R_0 = S(1+T)$ is the number in (2.14).

We use notations introduced in the proof of Theorem 2.3, and for h, k considered there define

$$\tilde U_h^k(X) = J\left[H(X) - \nabla Q_h^k(X)\right], \tag{2.63}$$

where Q_h^k is defined by (2.18), and $H(X)$ is defined by (2.47)–(2.48) with R_1 given by (2.59). Define functions $\tilde U_h$ on $[0,T]\times\Omega$ by setting them equal to $\tilde U_h^k$ on the time interval $t \in [kh, (k+1)h)$. Then, from (2.19) and (2.47)–(2.48), we have $\tilde{U}_h \in L^\infty((0,T) \times \mathbf{R}^3)$ with

$$\|\tilde{U}_h(t,\cdot)\|_{L^\infty(\mathbf{R}^3)} \le S + R_1 + 1 \quad \text{for all } t \in [0,T),\ h \in (0,1). \tag{2.64}$$

Let $h_j \to 0$ be a sequence for which (2.27)–(2.32) hold. From (2.31), using convexity of $Q_{h_j}(t,\cdot)$ we conclude that

$$\nabla \overline{Q}_{h_j}(t,\cdot) \to \nabla P^*(t,\cdot) \quad \text{a.e. in } \mathbf{R}^3 \quad \text{for each } t \in [0,T),$$

and from this, noting that $U_{h_j}(X,t) = J[H(X) - \nabla \overline{Q}_{h_j}(X,t)]$ on $[0,T) \times \mathbf{R}^3$, we get

$$\tilde{U}_{h_j}(t,\cdot) \to \tilde{U}(t,\cdot) \quad \text{a.e. in } \mathbf{R}^3 \quad \text{for each } t \in [0,T), \tag{2.65}$$

where $\tilde{U}$ is defined by (2.47)–(2.49). By (2.64), (2.65) and the dominated convergence theorem

$$\tilde{U}_{h_j} \to \tilde{U} \quad \text{in } L^r([0,T) \times B(0,R); \mathbf{R}^3) \quad \text{for all } \ R > 0,\ r \in [1,\infty). \tag{2.66}$$

Our choice $R_1 \ge R_0 + 1$ implies that $\overline{U}_h(t,X) = \tilde{U}_h(t,X)$ in $[0,T) \times B(0,R_0+1)$. Then from (2.24) it follows that α_h satisfy (2.20) with U_h^k replaced by $\tilde{U}_h^k$. Thus each α_h is a weak solution of

$$\begin{aligned} &\frac{\partial \alpha_h}{\partial t} + \operatorname{div}(\alpha_h \tilde{U}_h) = 0 \qquad \text{in } \ \mathbf{R}^3 \times (0,T), \\ &\alpha_h(0,X) = \alpha_h^0(X). \end{aligned} \tag{2.67}$$

For each $h \in (0,1)$, from (2.19), (2.47)–(2.48), and (2.63)

$$\|\nabla \tilde{U}_h\|_{L^\infty((0,T)\times\mathbf{R}^3)} \le C\left(\frac{S}{h} + 1\right), \tag{2.68}$$

where C depends only on functions η, ζ. Also, from its definition, $\tilde{U}_h$ is a divergence-free vector field. Then there exists a unique Lagrangian flow $\Phi_h : \mathbf{R}^3 \times \mathbf{R}^1 \to \mathbf{R}^3$ induced by $\tilde{U}_h$, and for each t the map $(\Phi_h)_t : \mathbf{R}^3 \to \mathbf{R}^3$ is $\mathcal{L}^3$ measure-preserving (see, e.g., [8, Theorem III.2]). From this, for the sequence $\{h_j\}$ considered above, using also the properties (2.64), (2.68) of $\tilde{U}_{h_j}$, the properties (2.46) of $\tilde{U}$, and the convergence in (2.66), we can apply [1, Theorem 6.5] to conclude that for each $t \in [0,T]$

$$(\Phi_{h_j})_t \to \Phi_t \quad \text{in } L^1_{loc}(\mathbf{R}^3) \quad \text{as } j \to \infty. \tag{2.69}$$

Let $q > 1$ be as in (2.13). Then $\|\alpha_h^0\|_{L^q(\mathbf{R}^3)} \le \|\alpha_0\|_{L^q(\mathbf{R}^3)}$. From Lemma 2.13, it follows that the function $v(t,\cdot) = (\Phi_h)_t \# \alpha_h^0$ satisfies $v \in L^\infty([0,T], L^q(\mathbf{R}^3))$ and v is a weak solution of (2.67). Since $\tilde{U}_h \in L^\infty((0,T), W^{1,\infty}(\mathbf{R}^3))$ by (2.64) and (2.68)), it follows that (2.67) has at most one weak solution in $v \in L^\infty([0,T], L^q(\mathbf{R}^3))$; see, e.g., [8, Theorem II.2]. Using (2.26), we conclude that

$$\alpha_h(t,\cdot) = (\Phi_h)_t \# \alpha_h^0. \tag{2.70}$$

From (2.70) and (2.27), we get for any $t > 0$, $j = 1, \dots$ and any $\varphi \in C_c(\mathbf{R}^3)$

$$\int_{\mathbf{R}^3} \varphi(\Phi_{h_j}(t,X)) \alpha_{h_j}^0(X) dX = \int_{\mathbf{R}^3} \varphi(y) \alpha_{h_j}(t,y) dy.$$

Passing to the limit $j \to \infty$ in the last equality, by using (2.69), the fact that $\alpha^0_{h_j} \to \alpha_0$ in $L^q(\mathbf{R}^3)$, the dominated convergence theorem in the left-hand side, and (2.29) in the right-hand side we obtain

$$\int_{\mathbf{R}^3} \varphi(\Phi(t,X))\alpha_0(X)dX = \int_{\mathbf{R}^3} \varphi(y)\alpha(t,y)dy \tag{2.71}$$

for any $\varphi \in C_c(\mathbf{R}^3)$. This implies (2.61).

Since Φ_t is a measure-preserving map, we use Lemma 2.12(ii) to conclude that the left-hand side of (2.71) is equal to $\int_{\mathbf{R}^3} \varphi(y)\alpha_0(\Phi_t^*(y))dy$, and now (2.71) implies (2.62). ∎

2.4. Lagrangian flow in the physical space. Throughout this section we assume that Ω, T, P_0, α, P, U, $\tilde{U}$, Φ are as in Proposition 2.14. Moreover, we fix R_1 in the definition of $\tilde{U}$ sufficiently large so that the conclusions of Proposition 2.14 hold.

Below we use the following notation: for a function $g(t,x)$ we denote $g_t(x) := g(t,x)$.

We intend to define a Lagrangian flow in the physical space $F : [0,T) \times \Omega \to \Omega$ by defining $F_t : \Omega \to \Omega$ for $t \in [0,T)$ by

$$F_t := \nabla P_t^* \circ \Phi_t \circ \nabla P_0, \tag{2.72}$$

where P_t^* is the convex dual of P_t, and $\Phi_t : \mathbf{R}^3 \to \mathbf{R}^3$ is the Lagrangian flow in the dual space constructed in Lemma 2.11. For that, we need to prove first the following.

LEMMA 2.15. *The right-hand side of* (2.72) *is defined $\mathcal{L}^4$-a.e. in* $[0,T) \times \Omega$. *Moreover, for any $t \in [0,T)$ the right-hand side of* (2.72) *is defined $\mathcal{L}^3$-a.e. in Ω. The map $F : [0,T) \times \Omega \to \Omega$ defined by* (2.72) *is Borel.*

Proof. Since P_0 is a bounded convex function on B, then ∇P_0 exists on $\Omega \setminus N_0^1$, where N_0^1 is a Borel subset of Ω with $\mathcal{L}^3(N_0^1) = 0$. Also, since P^* is a bounded Borel function on $[0,T] \times \mathbf{R}^3$ and for every t the function $P^*(t,\cdot)$ is convex in $\mathbf{R}^3$, then ∇P^* exists on $([0,T] \times \mathbf{R}^3) \setminus N^2$, where N^2 is a Borel subset of $[0,T] \times \mathbf{R}^3$ with $\mathcal{L}^4(N^2) = 0$; moreover, denoting $N_t^2 := N^2 \cap (\{t\} \times \mathbf{R}^3)$, we have $\mathcal{L}^3(N_t^2) = 0$ for every $t \in [0,T]$.

Then the right-hand side of (2.72) is defined for all

$$(t,x) \in ([0,T] \times \Omega) \setminus (([0,T] \times N_0^1) \cup M),$$
$$\text{where } \; M = \left\{(s,y) \in [0,T] \times (\Omega \setminus N_0^1) \mid (s, \Phi(s, \nabla P_0(y))) \in N^2\right\}.$$

From its definition, M is a Borel set.

It remains to prove that $\mathcal{L}^4(M) = 0$ and $\mathcal{L}^3(M_t) = 0$ for every $t \in [0,T]$, where $M_t := M \cap (\{t\} \times \mathbf{R}^3)$. By Fubini's theorem, it is sufficient to prove that $\mathcal{L}^3(M_t) = 0$ for every $t \in [0,T]$.

Fix $t \in [0,T)$. Then, since $\Phi_t : \mathbf{R}^3 \to \mathbf{R}^3$ is $\mathcal{L}^3$-measure preserving, we have $\mathcal{L}^3\left(\Phi_t^{-1}(N_t^2)\right) = 0$, and using that $\nabla P_0 \# \chi_\Omega = \alpha_0$ and thus $\nabla P_0 \# \chi_{\Omega \setminus N_0^1} = \alpha_0$, we compute

$$|M_t| = \left|\left\{x \in \Omega \setminus N_0^1 \mid \nabla P_0(x) \in \Phi_t^{-1}(N_t^2)\right\}\right| = \int_{\Phi_t^{-1}(N_t^2)} \alpha_0(z)dz = 0$$

since $\alpha_0 \in L^1_{loc}(\mathbf{R}^3)$.

Thus we can define $F : [0,T) \times \Omega \to \Omega$ by (2.72). Then by Lemma 2.11, F is a Borel mapping. ∎

It remains to prove that if F is defined by (2.72), then (F, P) is a weak Lagrangian solution in the sense of Definition 2.5.

We first show that the initial condition for the flow is satisfied.

PROPOSITION 2.16. *$F(0, x) = x$ for a.e. $x \in \Omega$.*

Proof. By (2.72), $F_0(x) = \nabla P_0^* \circ \Phi_0 \circ \nabla P_0(x)$ for all $x \in \Omega \setminus N_0$, where N_0 is a Borel set with $\mathcal{L}^3[N_0] = 0$.

From convexity of P_0 in B and (2.4), there exist Borel sets $N_1 \subset \Omega$, $N_2 \subset \mathbf{R}^3$ with $\mathcal{L}^3(N_1) = \mathcal{L}^3(N_2) = 0$ such that P_0 (resp. P_0^*) is differentiable on $\Omega \setminus N_1$ (resp. $\mathbf{R}^3 \setminus N_2$). Moreover, if $x \in \Omega \setminus [N_1 \cup (\nabla P_0)^{-1}(N_2)]$, then $\nabla P_0^* \circ \nabla P_0(x) = x$.

By Lemma 2.11(ii), $\Phi_0(x) = x$ in $\mathbf{R}^3 \setminus N_3$, where N_3 is a Borel set with $\mathcal{L}^3[N_3] = 0$.

Thus $F_0(x) = x$ for all $x \in \Omega \setminus [N_0 \cup N_1 \cup (\nabla P_0)^{-1}(N_2 \cup N_3)]$. In order to complete the proof of the proposition, we need to show that $\mathcal{L}^3[(\nabla P_0)^{-1}(N_2 \cup N_3) \cap \Omega] = 0$.

Denote $\mu = \alpha_0 dx$ in $\mathbf{R}^3$. Since $\alpha_0 \in L^q(\mathbf{R}^3)$, then $\mu[N_2 \cup N_3] = 0$, and since $\nabla P_0 \# \chi_\Omega = \alpha_0$, we get $\mathcal{L}^3[(\nabla P_0)^{-1}(N_2 \cup N_3) \cap \Omega] = \mu[N_2 \cup N_3] = 0$. ∎

The next step is to prove that the property stated in Definition 2.5(ii) is satisfied.

PROPOSITION 2.17. *For every $t > 0$ the map $F_t : \Omega \to \Omega$ is $\mathcal{L}^3$-measure preserving.*

Proof. In order to complete the proof, we need to justify the following calculation: for any $\varphi \in C(\mathbf{R}^3)$,

$$\int_\Omega \varphi(F_t(x))dx = \int_\Omega \varphi \circ \nabla P_t^* \circ \Phi_t \circ \nabla P_0(x)dx \tag{2.73}$$

$$= \int_{\mathbf{R}^3} \varphi \circ \nabla P_t^* \circ \Phi_t(y)\alpha_0(y)dy \tag{2.74}$$

$$= \int_{\mathbf{R}^3} \varphi \circ \nabla P_t^*(z)\alpha_t(z)dz \tag{2.75}$$

$$= \int_\Omega \varphi(x)dx. \tag{2.76}$$

Then the proposition follows from Lemma A.2.

Now we prove these equalities.

Equality (2.73) follows from the definition of F_t.

Since $\nabla P_0 \# \chi_\Omega = \alpha_0$ and $\alpha_0 \in L^q(\mathbf{R}^3)$ with compact support, Corollary A.3 implies

$$\int_\Omega \psi(\nabla P_0(x))dx = \int_{\mathbf{R}^3} \psi(y)\alpha_0(y)dy \qquad \text{for all} \quad \psi \in L^\infty(\mathbf{R}^3).$$

Choosing $\psi = \varphi \circ \nabla P_t^* \circ \Phi_t \in L^\infty(\mathbf{R}^3)$, we conclude that the right-hand side of (2.73) is equal to (2.74).

Equality of expressions (2.75) and (2.74) follows from (2.61) and Corollary A.3, since $\varphi \circ \nabla P_t^* \in L^\infty(\mathbf{R}^3)$ and $\alpha_0 \in L^q(\mathbf{R}^3)$ with compact support.

Equality of expressions (2.75) and (2.76) follows from $\nabla P_t^* \# \alpha_t = \chi_\Omega$. ∎

Now we prove that (2.35) holds for all $r \in [1, \infty)$.

PROPOSITION 2.18. *For any $t_0 \in [0, T)$, any $r \in [1, \infty)$ and*

$$\lim_{t \to t_0,\, t \in [0,T)} \int_\Omega |F_t(x) - F_{t_0}(x)|^r dx = 0.$$

Proof. By Lemma 2.15, for any $t \in [0,T)$ (2.72) holds $\mathcal{L}^3$-a.e. in Ω. Also, $\nabla P_0 \# \chi_\Omega = \alpha_0$. Thus we get for any $t, t_0 \in [0,T)$

$$\begin{aligned}\int_\Omega |F_t(x) - F_{t_0}(x)|^r dx &= \int_\Omega |\nabla P_t^* \circ \Phi_t \circ \nabla P_0(x) - \nabla P_{t_0}^* \circ \Phi_{t_0} \circ \nabla P_0(x)|^r dx \\ &= \int_{\mathbf{R}^3} |\nabla P_t^* \circ \Phi_t(y) - \nabla P_{t_0}^* \circ \Phi_{t_0}(y)|^r \alpha_0(y) dy \\ &\le C \int_{\mathbf{R}^3} |\nabla P_t^* \circ \Phi_t(y) - \nabla P_{t_0}^* \circ \Phi_t(y)|^r \alpha_0(y) dy \\ &\quad + C \int_{\mathbf{R}^3} |\nabla P_{t_0}^* \circ \Phi_t(y) - \nabla P_{t_0}^* \circ \Phi_{t_0}(y)|^r \alpha_0(y) dy \\ &=: C(I_1 + I_2).\end{aligned}$$

We show first that $I_1 \to 0$ as $t \to t_0$. Note that $\|\alpha_t\|_{L^q(\mathbf{R}^3)} = \|\alpha_0\|_{L^q(\mathbf{R}^3)}$ for $t \in (0,T)$; this follows from (2.62) since Φ_t^* is a $\mathcal{L}^3$-measure-preserving map. Now we use (2.61) to estimate

$$\begin{aligned} I_1 &= \int_{\mathbf{R}^3} |\nabla P_t^* \circ \Phi_t(y) - \nabla P_{t_0}^* \circ \Phi_t(y)|^r \alpha_0(y) dy \\ &= \int_{\mathbf{R}^3} |\nabla P_t^*(y) - \nabla P_{t_0}^*(y)|^r \alpha_t(y) dy \\ &\le \|\nabla P_t^* - \nabla P_{t_0}^*\|_{rq'}^r \|\alpha_t\|_q \\ &= \|\nabla P_t^* - \nabla P_{t_0}^*\|_{rq'}^r \|\alpha_0\|_q \to 0 \quad \text{as } t \to t_0 \end{aligned}$$

by Theorem 2.3.

Now we show that $I_2 \to 0$ as $t \to t_0$. Since $\nabla P_t^*(y) \in B$ for a.e. y for each t, and $\alpha_0 \in L^1(\mathbf{R}^3)$ (since α_0 is in L^q and has compact support), then by the dominated convergence theorem it remains to prove that, for every t_0,

$$\nabla P_{t_0}^* \circ \Phi_t(y) - \nabla P_{t_0}^* \circ \Phi_{t_0}(y) \to 0 \quad \text{as } t \to t_0 \tag{2.77}$$

for a.e. $y \in \mathbf{R}^3$. First we note that, since Φ_t is measure preserving, then it follows from Lemma 2.11(i) that

$$\Phi_t(y) \to \Phi_{t_0}(y) \text{ as } t \to t_0 \text{ in } [0,T] \text{ for a.e. } y \in \mathbf{R}^3.$$

If y is such a point, and if in addition $\Phi_{t_0}(y)$ is a point of continuity for $\nabla P_{t_0}^*$, then convergence in (2.77) holds at y. As $P_{t_0}^*$ is convex, $\nabla P_{t_0}^*$ is differentiable a.e. Since Φ_{t_0} is measure preserving, it follows that $\Phi_{t_0}(y)$ is a point of continuity for $\nabla P_{t_0}^*$ for a.e. y. Thus (2.77) holds for a.e. $y \in \mathbf{R}^3$. The proposition is proved. ∎

It remains to show that properties (iii) and (iv) in Definition 2.5 are satisfied. The first step is the following.

LEMMA 2.19. *$\nabla P_t \circ F_t(x) = \Phi_t \circ \nabla P_0(x)$ for a.e. $(t,x) \in (0,T) \times \Omega$.*

Proof. To prove the lemma, we need to justify the formal computation

$$\nabla P_t \circ F_t = \nabla P_t \circ \nabla P_t^* \circ \Phi_t \circ \nabla P_0 = \Phi_t \circ \nabla P_0,$$

since $\nabla P_t \circ \nabla P_t^*$ is the identity on the support of α_t.

Now we make this argument rigorous.

Since $P(t,x)$ is a bounded Borel function in $[0,T] \times B$ and $P(t,\cdot)$ is a convex function in B for every $t \in [0,T]$, then ∇P exists in $([0,T] \times \Omega) \setminus N^1$, where N^1 is a

Borel subset of $[0,T]\times\Omega$ with $\mathcal{L}^4(N^1)=0$; moreover, denoting $N^1_t := N^1\cap(\{t\}\times\mathbf{R}^3)$, we have $\mathcal{L}^3(N^1_t)=0$ for every $t\in[0,T]$.

Then $F^{-1}(N^1)$ is a Borel subset of $[0,T]\times\Omega$ satisfying

$$\mathcal{L}^4(F^{-1}(N^1))=0.$$

Indeed, it is enough to show that $\mathcal{L}^3(F_t^{-1}(N^1_t))=0$ for every $t\in[0,T]$, and this follows from $\mathcal{L}^3(N^1_t)=0$ and Proposition 2.17.

Now we can conclude, using Lemma 2.15, that

$$Z_t(x)=\nabla P_t\circ\nabla P_t^*\circ\Phi_t\circ\nabla P_0(x)\quad\text{ for }(t,x)\in([0,T]\times\Omega)\setminus\tilde{N},\quad\text{ where }\mathcal{L}^4(\tilde{N})=0.$$

Let $\tilde{M}=\{(t,x)\in[0,T]\times\mathbf{R}^3\mid\nabla P(\nabla P^*(t,x),t)\neq x\}$, where we include (t,x) such that either $\nabla P^*(t,x)$ or $\nabla P(\nabla P^*(t,x),t)$ do not exist. Then $\tilde{M}$ is a Borel set.

The proof of the lemma will be completed if we show that

$$\mathcal{L}^4\left[\{(t,x)\in([0,T]\times\Omega)\setminus\tilde{N}\ \mid\ (\Phi_t\circ\nabla P_0(x),t)\in\tilde{M}\}\right]=0. \tag{2.78}$$

Since $\nabla P_t\#\chi_\Omega=\alpha_t$ and $\nabla P_t^*\#\alpha_t=\chi_\Omega$, it follows that for any t

$$\nabla P_t\circ\nabla P_t^*(x)=x\quad\text{ for }\alpha_t\text{-a.e. }\ x\in\mathbf{R}^3.$$

Then, denoting $\tilde{M}_t=\tilde{M}\cap(\mathbf{R}^3\times\{t\})$, we have

$$\int_{\tilde{M}_t}\alpha_t(x)dx=0\quad\text{ for any }t\in[0,T].$$

Thus for any $t\in[0,T]$ we get, using that $\mathcal{L}^3(\tilde{N}_t)=0$ and thus $\nabla P_0\#\chi_{\Omega\setminus\tilde{N}_t}=\alpha_0$, and also using Lemma 2.11(iv), Lemma 2.12(ii), and (2.62), the following:

$$\begin{aligned}\mathcal{L}^3\left[\{x\in\Omega\setminus\tilde{N}_t\ \mid\ \Phi_t\circ\nabla P_0(x)\in\tilde{M}_t\}\right]&=\mathcal{L}^3\left[\{x\in\Omega\setminus\tilde{N}_t\ \mid\ \nabla P_0(x)\in\Phi_t^*(\tilde{M}_t)\}\right]\\&=\int_{\Phi_t^*(\tilde{M}_t)}\alpha_0(x)dx=\int_{\tilde{M}_t}\alpha_0(\Phi_t^*(x))dx\\&=\int_{\tilde{M}_t}\alpha_t(x)dx=0.\end{aligned}$$

Now (2.78) follows from Fubini's theorem. The lemma is proved. □

Now we show existence of the map F^* from Definition 2.5(iii).

PROPOSITION 2.20. *The map F satisfies property* (iii) *in Definition* 2.5.

Proof. Similar to Lemma 2.15, we can show that for every $t\in[0,T]$ the expression $\nabla P_0^*\circ\Phi_t^*\circ\nabla P_t(x)$ is defined for a.e. $x\in\Omega$, and that the map $F_t^*:\Omega\to\Omega$ defined by $F_t^*=\nabla P_0^*\circ\Phi_t^*\circ\nabla P_t$ is Borel.

Since F_t is measure preserving, then $F_t^*\circ F_t(x)=\nabla P_0^*\circ\Phi_t^*\circ\nabla P_t\circ F_t(x)$ for a.e. $x\in\Omega$. Using Lemma 2.19, get $F_t^*\circ F_t(x)=\nabla P_0^*\circ\Phi_t^*\circ\Phi_t\circ\nabla P_0(x)$ a.e. in Ω. Since $\Phi_t^*\circ\Phi_t(y)=y$ for a.e. y and thus for α_0-a.e. $y\in\mathbf{R}^3$, and $\nabla P_0\#1_{|\Omega}=\alpha_0$, then $\Phi_t^*\circ\Phi_t\circ\nabla P_0(x)=\nabla P_0(x)$ for a.e. $x\in\Omega$. Thus $F_t^*\circ F_t(x)=\nabla P_0^*\circ\nabla P_0(x)=x$ for a.e. $x\in\Omega$.

By a similar argument, $F_t\circ F_t^*(x)=x$ for a.e. $x\in\Omega$. □

Finally we show property (iv) of Definition 2.5.

PROPOSITION 2.21. *Equality* (2.38) *holds for any* $\phi \in C_c^1((0,T)\times\Omega, \mathbf{R}^3)$. *Moreover, possibly after modifying* $Z(t,x)$ *on a negligible subset of* $(0,T)\times\Omega$, *we have* $Z(\cdot,x) \in W^{1,\infty}([0,T);\mathbf{R}^3)$ *for a.e.* $x\in\Omega$, *and* (2.44) *holds.*

Proof. From the definition of the Lagrangian flow Φ, i.e., properties (i)–(iii) of Lemma 2.11, for a.e. $X \in \mathbf{R}^3$ and every $t\in[0,T)$

$$\Phi(t,X) = X + \int_0^t \tilde{U}(s,\Phi_s(X))ds.$$

Thus the above equality holds for all $X\in\mathbf{R}^3\backslash N$, where $|N|=0$. Since $\nabla P_0 \# \chi_\Omega = \alpha_0$ and $\alpha_0 \in L^q(\mathbf{R}^3)$, it follows that $\left|(\nabla P_0)^{-1}(N)\cap\Omega\right| = \int_N \alpha_0(z)dz = 0$. Thus for a.e. $x\in\Omega$ and every $t\in[0,T)$

$$\Phi(t,\nabla P_0(x)) = \nabla P_0(x) + \int_0^t U(s,\Phi_s(\nabla P_0(x)))ds, \tag{2.79}$$

where we replaced $\tilde{U}(s,\Phi_s(\nabla P_0(x)))$ by $U(s,\Phi_s(\nabla P_0(x)))$ based on (2.55), (2.56). Multiplying the last equality by $\partial_t\eta(t,x)$, where $\eta\in C_c^1([0,T)\times\mathbf{R}^3)$, and integrating we get

$$\begin{aligned}\int_{[0,T)\times\Omega}\partial_t\eta(t,x)\Phi(t,\nabla P_0(x))dtdx &= \int_{[0,T)\times\Omega}\partial_t\eta(t,x)\nabla P_0(x)dtdx\\ &\quad+\int_{[0,T)\times\Omega}\partial_t\eta(t,x)\int_0^t U(s,\Phi_s(\nabla P_0(x)))dsdtdx.\end{aligned}$$

Note that $\eta(T,x)\equiv 0$. In the right-hand side, we perform the integration with respect to t in the first integral, and integrate by parts with respect to t in the second integral, to get

$$\begin{aligned}\int_{[0,T)\times\Omega}\partial_t\eta(t,x)\Phi(t,\nabla P_0(x))dtdx &= -\int_\Omega \eta(0,x)\nabla P_0(x)dx\\ &\quad-\int_{[0,T)\times\Omega}\eta(t,x)U(t,\Phi_t(\nabla P_0(x)))dtdx.\end{aligned}\tag{2.80}$$

Now we compute using (2.7), (2.72), and Lemma 2.19

$$U(t,\Phi_t(\nabla P_0(x))) = J[\Phi_t(\nabla P_0(x)) - \nabla P_t^*(\Phi_t(\nabla P_0(x)))] = J[Z(t,x)-F(t,x)]$$

for a.e. (t,x). Substituting this into the right-hand side of (2.80) and using Lemma 2.19 to replace $\Phi(t,\nabla P_0(x))$ by $Z(t,x)$ in the left-hand side of (2.80), we obtain (2.38).

Finally, $Z(\cdot,x)\in W^{1,\infty}([0,T);\mathbf{R}^3)$ for a.e. $x\in\Omega$ (possibly after modifying $Z(t,x)$ on a negligible subset of $(0,T)\times\Omega$) follows from Lemmas 2.19 and 2.11(i). Then (2.79) and Lemma 2.19 imply (2.44). ∎

Now the properties of P in Theorem 2.3, and the properties of (F,P) proved in Propositions 2.16, 2.17, 2.18, 2.20, and 2.21 imply Theorem 2.9.

3. Lagrangian solutions of 2-dimensional semigeostrophic shallow water equations in physical space.

3.1. Model formulation and background. In this section we extend the approach of the previous section to the model considered by Cullen and Gangbo in [5]. They study a shallow water approximation to the following free boundary problem for the system (2.1) in an evolving region $\mathcal{D}(t)$: the free boundary condition is

$$u \cdot \nu = \text{normal speed of the boundary} \quad \text{on } \partial\left[\cup_{t\in[0,T)}\{t\}\times\mathcal{D}(t)\right], \tag{3.1}$$

and the initial conditions are

$$\begin{aligned} &\mathcal{D}(0) = \mathcal{D}_0, \\ &p(0,x) = p_0(x) \quad \text{in } \Omega, \end{aligned} \tag{3.2}$$

where $\mathcal{D}_0, p_0(x)$ are a given set and function, respectively. In [5], problem (2.1), (3.1), (3.2) is considered in the following shallow water approximation.

The fluid is contained within a region Ω of (x_1, x_2)-plane but the height $h(t, x_1, x_2)$ above the reference level is unknown and can evolve in time:

$$\mathcal{D}(t) = \{(x_1, x_2, x_3) \in \mathbf{R}^3 \quad | \quad (x_1, x_2) \in \Omega, \quad 0 \le x_3 \le h(t, x_1, x_2)\}. \tag{3.3}$$

The pressure on the top boundary of the fluid is a given constant p_0, and

$$p(t, x_1, x_2, x_3) = [h(t, x_1, x_2) - x_3] + p_0. \tag{3.4}$$

The horizontal components of velocity are independent of x_3.

Then the problem (2.1), (3.1), (3.2) in an evolving 3-d domain can be rewritten as a problem in the 2-d domain Ω, for the unknown height function $h(t,x)$ defined on $[0,T)\times\Omega$, and horizontal components of velocity $v = (v_1, v_2)$. As noted by [5], it is possible for h to become zero on part of Ω. We set $v = 0$ in such regions. Then the problem (2.1), (3.1), (3.2) in the shallow water approximation (3.3)–(3.4) can be written as the following problem for $h(t,x)$ and $v = (v_1, v_2)$ defined on $[0,T)\times\Omega$ and $v = (v_1, v_2)$ defined in Ω:

$$D_t X = J(X - x) \qquad \text{in } [0,T)\times\Omega, \tag{3.5}$$

$$\partial_t h + \operatorname{div}(hv) = 0 \qquad \text{in } [0,T)\times\Omega, \tag{3.6}$$

$$X = \nabla P, \quad \text{where } P(t, x_1, x_2) = h(t, x_1, x_2) + \frac{1}{2}(x_1^2 + x_2^2), \tag{3.7}$$

$$D_t = \partial_t + v\cdot\nabla, \quad \nabla = (\partial_{x_1}, \partial_{x_2}), \tag{3.8}$$

$$v \cdot \nu = 0 \quad \text{on } [0,T)\times\partial\Omega, \tag{3.9}$$

$$h(0,x) = h_0(x) \quad \text{in } \Omega. \tag{3.10}$$

Here $J \equiv J_2 = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. Note that (3.6) corresponds to the condition (3.1) on the surface $\{(x_1, x_2, x_3) \mid h(x_1, x_2) > 0,\ x_3 = h(x_1, x_2)\}$, which is the top of the fluid.

The Cullen–Purser stability condition implies that the function $x \to P(t,x)$ is convex in Ω for any t, where $P(t,x)$ is defined in (3.7). Thus the initial height function $h_0(x)$ must satisfy

$$h_0 + \frac{1}{2}|x|^2 \text{ is convex in } \Omega.$$

Cullen and Gangbo in [5] rewrite the system (3.5)–(3.10) in dual variables (t, X):

$$\partial_t \alpha + \nabla\cdot(U\alpha) = 0 \quad \text{in } [0,T)\times\mathbf{R}^2, \tag{3.11}$$

$$\nabla P_t \# h_t = \alpha_t \quad \text{for any } t \in [0,T), \tag{3.12}$$

$$U(t,X) = J_2(X - \nabla P^*(t,X)) \quad \text{in } [0,T) \times \mathbf{R}^2, \tag{3.13}$$

$$P^*(t,X) = \sup_{x \in \Omega}(x \cdot X - P(t,x)) \quad \text{for } (t,X) \in [0,T) \times \mathbf{R}^2, \tag{3.14}$$

$$\alpha_0(X) \quad \text{is prescribed on } \mathbf{R}^2. \tag{3.15}$$

Cullen and Gangbo in [5] prove existence of a weak solution of system (3.11)–(3.15). The precise statement of their result is Theorem 2.3 with the following changes: $\mathbf{R}^3$ replaced by $\mathbf{R}^2$; $DP_0\#\chi_\Omega$ replaced by $DP_0\#h_0$ in (2.13); and (2.5)–(2.8) replaced by (3.11)–(3.15) in the assertion Theorem 2.3(v), where $h(t,x) = P(t,x) - \frac{1}{2}|x|^2$.

In this section we study the shallow water semigeostrophic model (3.5)–(3.10) in physical space. The unknown functions in this problem are the (modified) pressure P and the velocity v. The Eulerian formulation of weak solutions is obtained from the following argument. The Eulerian form of (3.5) is

$$\partial_t X + v \cdot \nabla X = J(X - x). \tag{3.16}$$

Multiplying (3.16) by h and adding (3.5) multiplied by X, we get

$$\partial_t(hX) + \sum_{j=1}^{2} \partial_{x_j}(hv_j X) = hJ(X - x).$$

Thus we have the following.

Definition 3.1. *Let $v : [0,T) \times \Omega \to \mathbf{R}^2$ and $P : [0,T) \times \Omega \to \mathbf{R}^1$ satisfy $v \in L^1([0,T) \times \Omega, \mathbf{R}^2)$, $\nabla P \in L^\infty([0,T) \times \Omega) \cap C([0,T); L^1(\Omega))$, and $P(t,\cdot)$ is convex in Ω for every $t \in [0,\infty)$. Let $h(t,x) = P(t,x) - \frac{1}{2}|x|^2$. The pair (v,P) is a weak Eulerian solution of* (3.5)–(3.10) *if*

$$\begin{aligned} &\int_{(0,T)\times\Omega} \{\nabla P(t,x) \cdot [\partial_t \phi(t,x) + (v(t,x) \cdot \nabla)\phi(t,x)] \\ &\qquad + J[\nabla P(t,x) - x] \cdot \phi(t,x)\} h(t,x)\,dt\,dx \\ &\qquad + \int_\Omega [\nabla h_0(x) + x] \cdot \phi(0,x) h_0(x)\,dx = 0 \end{aligned} \tag{3.17}$$

for any $\phi \in C_c^1([0,T) \times \Omega; \mathbf{R}^2)$ and

$$\int_{[0,T)\times\Omega} [\partial_t \psi(t,x) + v(t,x) \cdot \nabla \psi(t,x)] h\,dt\,dx + \int_\Omega \psi(0,x) h_0(x)\,dx = 0 \tag{3.18}$$

for any $\psi \in C_c^1([0,T) \times \overline{\Omega})$.

Remark 3.2.

- Equality (3.17) is a weak form of (3.5). In the second integral in the left-hand side of (3.17), we used the fact that $P_0(x) = h_0(x) + \frac{1}{2}|x|^2$.
- Equality (3.18) is a weak form of (3.6) with the boundary condition (3.9) and initial condition (3.10).

Remark 3.3. Equality (3.17) of Definition 3.1 essentially requires the evolution equation (3.5) to hold only in the fluid region $\mathcal{D} = \{(t,x) \in [0,T) \times \Omega \mid h(t,x) > 0\}$. This is natural since, physically, the evolution is defined only in $\mathcal{D}$.

As in the case of the incompressible model (2.3), we can find P by solving the problem (3.11)–(3.15) in the dual coordinates, and v then is expressed by (2.12), but regularity of this solution is not sufficient to show that such v, P satisfy (3.5) in the Eulerian sense. Thus we define a Lagrangian solution.

3.2. Lagrangian solutions for shallow water model in physical space. Similar to section 2.2, we rewrite (3.5)–(3.10) in terms of (F, P), where $F : [0, T] \times \Omega \to \Omega$ is a (formal) Lagrangian flow corresponding to the full wind velocity $v = (v_1, v_2)$, and then we define the corresponding weak solution F, P. The difference is now that the vector field v is not divergence-free, but instead the transport equation $\partial_t h + \mathrm{div}\,(hv) = 0$ holds. Since F is a Lagrangian flow of v, solutions h of this transport equation satisfy $F_t \# h_0 = h_t$ for any $t \in (0, T)$. This property in the present replaces the Lebesgue-measure-preserving property of F in section 2.2.

Then we come to the following.

Definition 3.4. *Let $\Omega \subset \mathbf{R}^2$ be an open set, and let $T > 0$. Let $P_0(x)$ be a convex bounded function in Ω such that $h_0(x) = P_0(x) - \frac{1}{2}|x|^2 \geq 0$ in Ω. Let $r \in [1, \infty)$. Let $P : [0, T) \times \Omega \to \mathbf{R}^1$ satisfy*

$$P \in L^\infty([0, T); W^{1,\infty}(\Omega)) \cap C([0, T); W^{1,r}(\Omega)), \tag{3.19}$$

$$P(t, \cdot) \text{ is convex in } \Omega \text{ for each } t \in [0, T). \tag{3.20}$$

Let $h(t, x) = P(t, x) - \frac{1}{2}|x|^2$. Let $F : [0, T) \times \Omega \to \Omega$ be a Borel map satisfying

$$F \in C([0, T); L^r(\Omega, h_0 dx)). \tag{3.21}$$

Then the pair (P, F) is called a weak Lagrangian solution of (3.5)–(3.10) *in* $[0, T) \times \Omega$ *if (P, F) has the following properties:*

i. *$F(0, x) = x$ for $h_0\mathcal{L}^2$-a.e. $x \in \Omega$, and $P(0, x) = P_0(x)$ for a.e. $x \in \Omega$.*

ii. *For any $t \in [0, T]$ the mapping $F_t = F(t, \cdot) : \Omega \to \Omega$ satisfies $F_t \# h_0 = h_t$.*

iii. *There exists a Borel map $F^* : [0, T) \times \Omega \to \Omega$ such that for every $t \in (0, T)$ the map $F^*_t = F^*(t, \cdot) : \Omega \to \Omega$ satisfies $F^*_t \# h_t = h_0$, $F^*_t \circ F_t(x) = x$ for $h_0\mathcal{L}^2$-a.e. $x \in \Omega$, and $F_t \circ F^*_t(x) = x$ for $h_t\mathcal{L}^2$-a.e. $x \in \Omega$.*

iv. *The function*

$$Z(t, x) = \nabla P(t, F(t, x)) \tag{3.22}$$

is a weak solution of

$$\begin{aligned} &\partial_t Z(t, x) = J[Z(t, x) - F(t, x)] \qquad \text{on supp } h_0 \text{ in } (0, T) \times \Omega, \\ &Z(0, x) = \nabla P_0(x) \qquad \text{on supp } h_0 \text{ in } \Omega \end{aligned} \tag{3.23}$$

in the following sense: for any $\phi \in C^1_c([0, T) \times \Omega; \mathbf{R}^2)$

$$\begin{aligned} \int_{(0,T)\times\Omega} &[Z(t, x) \cdot \partial_t \phi(t, x) \\ &+ J(Z(t, x) - F(t, x))\phi(t, x)] h_0(x) dt dx \\ &+ \int_\Omega \nabla P_0(x) \cdot \phi(0, x) h_0(x) dx = 0. \end{aligned} \tag{3.24}$$

Remark 3.5. Remark 2.6, with obvious modifications, applies to the present case. In particular, continuity in time of P, F in (3.19), (3.21), combined with initial conditions in Definition 3.4(i), imply that

$$\begin{aligned} &\lim_{t\to 0+} \int_\Omega |F_t(x) - x|^r h_0(x) dx = 0, \\ &\lim_{t\to 0+} \|P_t - P_0\|_{W^{1,r}(\Omega)} = 0, \qquad \lim_{t\to 0+} \|h_t - h_0\|_{W^{1,r}(\Omega)} = 0, \end{aligned} \tag{3.25}$$

where $r \in [1, \infty)$ is from (3.19), (3.21).

Remark 3.6 (semigroup property). Similar to Remark 2.7, weak Lagrangian solution in Definition 3.4 satisfies the semigroup property: for any $t_1, t_2, t_3 \geq 0$ (2.40) holds for $h_{t_3}dx$ a.e. $x \in \Omega$, where $F_{(t_i,t_j)} = F_{t_i} \circ F^*_{t_j}$. The proof is based on properties of F_t, F^*_t in Definition 3.4(ii) and (iii), and the argument is similar to the one in Remark 2.7.

Remark 3.7. In Definition 3.4 we essentially consider the flow mapping F_t and its inverse F^*_t only in the fluid region $\mathcal{D} = \{(t,x) \in [0,T) \times \Omega \mid h(t,x) > 0\}$. That is, for any $t \in [0,T)$, Definition 3.4 does not contain any conditions for the map F_t away from $\mathcal{D}(0) = \{x \in \Omega \mid h_0(x) > 0\}$ or for the map F^*_t away from $\mathcal{D}(t) = \{x \in \Omega \mid h_t(x) > 0\}$. This is natural since, physically, the evolution is defined only in $\mathcal{D}$. A similar feature of weak Eulerian solutions is discussed in Remark 3.3. In particular, we can define F and F^* arbitrarily outside the domains $[0,T) \times \mathcal{D}_0$ and $\mathcal{D}$, respectively. Note also that the set $\Omega \setminus \mathcal{D}(t)$ is $h_t\mathcal{L}^2$-negligible for each t; thus Definition 3.4 determines the maps F_t and F^*_t almost everywhere with respect to the measures $h_0\mathcal{L}^2$ and $h_t\mathcal{L}^2$, respectively.

We now justify Definition 3.4 by showing that a smooth Lagrangian solution (F,P) with additional regularity property $\partial_t F \in L^\infty([0,T) \times \Omega)$ determines a weak Eulerian solution of (3.5)–(3.10).

LEMMA 3.8 (consistency of weak solutions for shallow water model). *Let $\Omega \subset \mathbf{R}^2$ be an open bounded set and $T > 0$. Let (F,P) be a weak Lagrangian solution of* (3.5)–(3.10) *in $[0,T) \times \Omega$ satisfying $\partial_t F \in L^\infty([0,T) \times \Omega; \mathbf{R}^2)$. Then the function*

$$v(t,x) := \frac{\partial F}{\partial t}(t, F^*_t(x)) \quad \text{for} \quad (t,x) \in [0,T) \times \Omega \tag{3.26}$$

satisfies $v \in L^\infty([0,T) \times \Omega; \mathbf{R}^2)$, and (v,P) is a weak Eulerian solution of (3.5)–(3.10) *in the sense of Definition* 3.1.

Proof. The proof is similar to the proof of Lemma 2.8(i). Thus we only sketch the argument and emphasize the structural differences between the models (2.3) and (3.5)–(3.10).

We show (3.18). Fix $t \in [0,T)$. Since $F_t \# h_0 = h_t$, then for each $\psi \in C^1_c([0,T) \times \overline{\Omega})$

$$\int_\Omega (\partial_t \psi)(t, F_t(x)) h_0(x) dx = \int_\Omega \partial_t \psi(t,x) h(t,x) dx.$$

Integrating with respect to t and using $\partial_t F \in L^\infty([0,T) \times \Omega)$, we get

$$\begin{aligned} &\int_{[0,T)\times\Omega} \{\partial_t[\psi(t, F_t(x))h_0(x)] - \partial_t F_t(x) \cdot (\nabla\psi)(t, F_t(x))h_0(x)\} \, dtdx \\ &\quad = \int_{[0,T)\times\Omega} \partial_t \psi(t,x) h(t,x) dtdx. \end{aligned}$$

Since $\psi(T,x) \equiv 0$ and $F(0,x) = x$ for all $x \in \Omega$ and using that $h \in C([0,T); W^{1,r}(\Omega))$ by (3.25), we get

$$\begin{aligned} &-\int_\Omega \partial_t \psi(0,x) h_0(x) dx - \int_{[0,T]\times\Omega} \partial_t F(t,x) \cdot (\nabla\psi)(t, F_t(x)) h_0(x) dtdx \\ &\quad = \int_{[0,T)\times\Omega} \partial_t \psi(t,x) h(t,x) dtdx. \end{aligned}$$

Now, changing variables $y = F_t(x)$ in the second integral in the left-hand side using $F_t\#h_0 = h_t$, we rewrite this integral as $\int_{[0,T]\times\Omega}(\partial_t F)(t, F_t^*(x)) \cdot \nabla\psi(t,x)h(t,x)dtdx$. Recalling the definition (3.26) of v, we get (3.18).

To show (3.17), we change variables $y = F_t(x)$ in (3.24), and using $F_t\#h_0 = h_t$ get

$$\begin{aligned}\int_{(0,T)\times\Omega}[\nabla P(t,y)\cdot\partial_t\phi(t,F_t^*(y)) \quad &+ \quad J(\nabla P(t,y)-y)\phi(t,F_t^*(y))]h(t,y)dtdy\\ &+ \textstyle\int_\Omega \nabla P_0(x)\cdot\phi(0,x)h_0(x)dx = 0,\end{aligned}$$

for all $\phi \in C_c^1([0,T)\times\Omega;\mathbf{R}^2)$. Now, similar to the proof of Lemma 2.8(i), we show that this equality holds for all ϕ of the form $\varphi(t,x) = \eta(t, F_t(x))$, where $\eta \in C_c^1([0,T)\times\Omega)$, and this, with use of the property in Definition 3.4(iii), implies (3.17). ∎

The main result of this section is the following.

THEOREM 3.9. *Let $\Omega \subset \mathbf{R}^2$ be an open bounded set and $\overline{\Omega} \subset B$, where B is an open ball $B(0,S)$. Let $h_0(x) \geq 0$ be such that $P_0(x) = h_0(x) + \frac{1}{2}|x|^2$ is a convex bounded function in B and that*

$$DP_0\#h_0 \in L^q(\nabla P_0(\Omega)) \text{ for some } q > 1. \tag{3.27}$$

Then for any $T > 0$ there exists a weak Lagrangian solution (P,F) of (3.5)–(3.10) *in $[0,T)\times\Omega$, where the properties stated in* (3.19) *are satisfied for any $r \in [1,\infty)$. Moreover, the function $Z(t,x)$ defined by* (3.22) *satisfies $Z(\cdot,x) \in W^{1,\infty}([0,T);\mathbf{R}^2)$ for $h_0\mathcal{L}^2$-a.e. $x \in \Omega$, and* (3.23) *is satisfied, in addition to the weak form* (3.24)*, in the following sense:*

$$\begin{aligned}&\partial_t Z(t,x) = J(Z(t,x) - F(t,x)) \qquad \text{for } h_0\mathcal{L}^2\times\mathcal{L}^1\text{-a.e. in } (t,x)\in,\\ &Z(0,x) = \nabla P_0(x) \qquad \text{for } h_0\mathcal{L}^2\text{-a.e. in } x\in\Omega.\end{aligned} \tag{3.28}$$

The proof of this theorem follows closely the proof of Theorem 2.9. Indeed, since in the present case we have $U \in BV(\mathbf{R}^2)$ and $\operatorname{div} U(t,\cdot) = 0$ in the sense of distributions by (3.13), we can repeat the argument of section 2.3 to define Lagrangian flow $\Phi_t : \mathbf{R}^2 \to \mathbf{R}^2$ in the dual space, with the same properties as Φ_t in section 2.3. Then we define $F_t : \Omega \to \Omega$ by (2.72) and repeat the argument of section 2.4 with some obvious changes which come from the fact that the property $\nabla P_t\#\chi_\Omega = \alpha_t$ is now replaced by $\nabla P_t\#h_t = \alpha_t$.

Appendix A. Some properties of measure-preserving maps.

DEFINITION A.1. *Let $\Psi : \mathbf{R}^m \to \mathbf{R}^n$ be a Borel map.*

i. *Let μ, ν be Radon measures on $\mathbf{R}^m$ and $\mathbf{R}^n$, respectively. The map Ψ pushes μ forward to ν, denoted $\Psi\#\mu = \nu$, if $\mu[\Psi^{-1}(A)] = \nu[A]$ for all Borel $A \subset \mathbf{R}^m$.*

ii. *Let $f \in L^1(\mathbf{R}^m)$ and $g \in L^1(\mathbf{R}^n)$. The map Ψ pushes f forward to g, denoted $\Psi\#f = g$, if Ψ pushes the measure $f\mathcal{L}^m$ forward to $g\mathcal{L}^n$.*

LEMMA A.2. *Let $\Psi : \mathbf{R}^m \to \mathbf{R}^n$ be a Borel map. Let μ, ν be Radon measures on $\mathbf{R}^m$ and $\mathbf{R}^n$, respectively. Then $\nu = \Psi\#\mu$ if and only if*

$$\int_{\mathbf{R}^m}\phi(\Psi(x))d\mu(x) = \int_{\mathbf{R}^n}\phi(y)d\nu(y)$$

for all ν-integrable functions ϕ on $\mathbf{R}^n$.

Proof. Since every ν-measurable function ϕ on $\mathbf{R}^n$ can be represented as

$$f = \sum_{k=1}^\infty \frac{1}{k}\chi_{A_k}$$

for some ν-measurable sets A_k in $\mathbf{R}^n$ [9, sect. 1.1.2, Thm. 7], and since μ and ν are Radon measures, Lemma A.2 follows directly from Definition A.1. $\square$

COROLLARY A.3. *Let Ψ, μ, ν be as in Lemma* A.2. *Assume also that $\nu = g\mathcal{L}^n$, where $g \in L^1(\mathbf{R}^n)$. If $\nu = \Psi\#\mu$, then*

$$\int_{\mathbf{R}^m} \phi(\Psi(x))d\mu(x) = \int_{\mathbf{R}^n} \phi(y)g(y)dy \qquad \textit{for all} \quad \phi \in L^\infty(\mathbf{R}^n). \tag{A.1}$$

In particular, if $f \in L^1(\mathbf{R}^m)$, $g \in L^1(\mathbf{R}^n)$, and $\Psi\#f = g$, then

$$\int_{\mathbf{R}^m} \phi(\Psi(x))f(x)dx = \int_{\mathbf{R}^n} \phi(y)g(y)dy \qquad \textit{for all} \quad \phi \in L^\infty(\mathbf{R}^n). \tag{A.2}$$

Proof. If $\nu = g\mathcal{L}^n$, where $g \in L^1(\mathbf{R}^n)$, then every $\phi \in L^\infty(\mathbf{R}^n)$ is ν-integrable. Now we apply Lemma A.2 to get (A.1). Finally, (A.2) follows from (A.1). $\square$

Note added in proof. After this paper was accepted for publication, we were informed by Helena Nussenzveig Lopes that, by combining the techniques of this paper with the ones of Lopes Filho and Nussenzveig Lopes [12], the results of Theorems 2.9 and 3.9 can be extended to the case $q = 1$.

REFERENCES

[1] L. AMBROSIO, *Transport equation and Cauchy problem for BV vector fields*, Invent. Math., 158 (2004), pp. 227–260.

[2] J.-D. BENAMOU AND Y. BRENIER, *Weak existence for the semigeostrophic equations formulated as a coupled Monge–Ampere/transport problem*, SIAM J. Appl. Math., 58 (1998), pp. 1450–1461.

[3] Y. BRENIER, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.

[4] M. CULLEN, *On the accuracy of the semi-geostrophic approximation*, Quart. J. Roy. Meteorol. Soc., 126 (2000), pp. 1099–1115.

[5] M. CULLEN AND W. GANGBO, *A variational approach for the 2-dimensional semi-geostrophic shallow water equations*, Arch. Ration. Mech. Anal., 156 (2001), pp. 241–273.

[6] M. CULLEN AND H. MAROOFI, *The fully compressible semi-geostrophic system from meteorology*, Arch. Ration. Mech. Anal., 167 (2003), pp. 309–336.

[7] M. CULLEN AND R. J. PURSER, *An extended Lagrangian theory of semi-geostrophic frontogenesis*, J. Atmospheric Sci., 41 (1984), pp. 1477–1497.

[8] R. J. DIPERNA AND P.-L. LIONS, *Ordinary differential equations, transport theory and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[9] L. C. EVANS AND R. F. GARIEPY, *Measure Theory and Fine Properties of Functions*, Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992.

[10] H. FEDERER, *Geometric Measure Theory*, Springer-Verlag, New York, 1969.

[11] B. J. HOSKINS, *The geostrophic momentum approximation and the semi-geostrophic equations*, J. Atmospheric Sci., 32 (1975), pp. 233–242.

[12] M. C. LOPES FILHO AND H. J. NUSSENZVEIG LOPES, *Existence of a weak solution for the semigeostrophic equation with integrable initial data*, Proc. Roy. Soc. Edinburgh Sect. A, 132 (2002), pp. 329–339.

[13] G. J. SHUTTS AND M. CULLEN, *Parcel stability and its relation to semi-geostrophic theory*, J. Atmospheric Sci., 44 (1987), pp. 1318–1330.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1396–1416

NEAR-PERIODIC LOCAL MINIMIZERS OF SINGULARLY PERTURBED FUNCTIONALS WITH NONLOCAL TERM*

NIKO TZOUKMANIS†

Abstract. We consider singular perturbed functionals with a nonlocal term which are a one-dimensional model for an austenite/twinned martensite interface. The parameter range is chosen in a way that a limit functional can be derived. By examining the limit problem, we show the existence of highly oscillating, near-periodic minimizers of the original functionals which are close to periodic sawtooth functions.

Key words. nonlocal variational problems, Γ-convergence, singular perturbations, solid-solid phase transitions

AMS subject classifications. 49K40, 74N15, 74G65

DOI. 10.1137/04061708X

1. Introduction.

Mathematical models for phase transitions often lead to variational problems involving nonlocal terms. Examples are given by phase separation in two-phase fluids (cf. [2], [3]) and the formation of diblock copolymer (cf. [20], [19], [21], [22], [23]). Nonlocal terms also appear in models for microstructure in solid-to-solid phase transitions.

In martensitic transformations, the phenomenon of twinning can occur. Hereby, a homogeneous phase called austenite is separated from a family of thin lamellae called twinned martensite. A nonlinear theory based on energy minimization has been established by Ball and James (cf. [5], [6]). In [12], Kohn and Müller have introduced a two-dimensional nonlocal model for an austenite/twinned martensite interface and discussed mathematical properties in [13]. Restriction of the model to the common boundary of the phases delivers a one-dimensional problem (cf. [4]), which is given by the functionals

$$J^\varepsilon(u) = \int_{-1}^{1} \varepsilon^2 (u'')^2 + W(u')dy + \sigma \|u\|^2_{H^{1/2}} \tag{1.1}$$

with parameters $\varepsilon > 0$ and $\sigma > 0$, where

$$u \in H^{2,0}_{per}(-1,1) = H^2(-1,1) \cap \{u(-1) = u(1), u'(-1) = u'(1)\} \cap \left\{ \int_{-1}^{1} u = 0 \right\}.$$

Hereby, W is a smooth double-well potential with $W(v) \geq 0$ and $W(v) = 0$ if and only if $v = -1$ or $v = 1$. We assume that W has at least quadratic growth at infinity; i.e., there exist $k > 0$, $v_0 > 0$ with $W(v) \geq k|v|^2$ for every $v \geq v_0$. The $H^{1/2}$-norm is given by

$$\|u\|^2_{H^{1/2}} = 2\pi \sum_{k=-\infty}^{\infty} |k||u_k|^2 = \int_{-1}^{1}\int_{-1}^{1} g(x-y)(u(x)-u(y))^2 dxdy,$$

*Received by the editors October 14, 2004; accepted for publication (in revised form) June 27, 2005; published electronically January 10, 2006. This article provides an overview of the author's doctoral thesis [24].

http://www.siam.org/journals/sima/37-5/61708.html

†Institut für Mathematik, Universität Augsburg, Germany. Current address: EurotaxGlass's International AG, Wilhelm-Röntgen-Straße 7, 63477 Maintal, Germany (niko.tzoukmanis@eurotaxglass.com, niko.tzoukmanis@t-online.de).

where u_k are the Fourier coefficients of u, and $g(t) = \frac{\pi}{4(1-\cos(\pi t))}$ (cf. [4], [13]). The W-term is the elastic energy in the martensite, while $\|u\|^2_{H^{1/2}}$, which is a nonlocal term, is derived from the energy in the austenite in the two-dimensional model by restriction to the interface. The variable u models the deformation of the material at the interface. The second-order singular perturbation term describes the surface energy and mathematically has the meaning of selecting certain "regular" minimizing sequences for the unperturbed functional J^0. It is easy to see that J^0 has no minimizer in $H^1(-1,1)$ and that any sequence of functions (u_n) with $u_n' = \pm 1$ on $(-1,1)$ and $u_n \to 0$ is a minimizing sequence. The perturbed functionals, however, do have minimizers due to convexity in the highest derivative (cf. [7, Chapter 3, Theorem 4.1]) and one is interested in the asymptotic behavior of the minimizers for $\varepsilon \to 0$.

Functionals similar to (1.1) have been studied in the literature but with a simpler right-hand term. In the parameter range $\sigma \sim 1$, and with $\|\cdot\|^2_{H^{1/2}}$ replaced by $\|\cdot\|^2_{L^2}$, Müller has shown in [18] that for small ε, the minimizers u_ε of J^ε are periodic with period $P^\varepsilon \sim \varepsilon^{1/3}$ and that intervals where $u_\varepsilon' \approx 1$ and $u_\varepsilon' \approx -1$ are separated by transition layers of size $\sim \varepsilon$ where u_ε' switches between these two values. In [4], where a new rescaling technique involving the Young measure on micropatterns has been introduced to characterize the minimizers of this problem, it has been conjectured that a similar result holds for the problem with the $H^{1/2}$-norm. More precisely, it is expected that the minimizers of (1.1) are close to periodic functions with period $P^\varepsilon \sim \varepsilon^{1/2}$ if ε is sufficiently small. However, the characterization of global minimizers turns out to be very difficult due to the nonlocal structure of $\|u\|^2_{H^{1/2}}$.

In this work, we examine a simpler but general nonlocal term whose kernel g is bounded and has no singularities but possesses the same symmetry and periodicity properties as the function g defined above. Following Ren and Wei (cf. [21]), we consider the parameter range $\sigma = \varepsilon$ rather than $\sigma = 1$. Physically, this has an interpretation related to the size of the sample (see [23]). Mathematically, it has the advantage that we can easily derive a limit functional via the Modica–Mortola theorem. Otherwise, the rescaling method introduced by Alberti and Müller [4] becomes necessary. The aim is to show a result similar to [21]. In that work, Ren and Wei have identified near-periodic local minimizers of functionals which in the special case where $m = 0$ (the parameter m being the average of u) can be written in the form (1.1) with $\sigma = \varepsilon$ but with $\|\cdot\|_{H^{1/2}}$ replaced by $\|\cdot\|_{L^2}$. We consider

$$I^\varepsilon(u) = \int_{-1}^{1} \varepsilon(u'')^2 + \frac{1}{\varepsilon}W(u')dy + E(u) \tag{1.2}$$

for $\varepsilon > 0$, where $u \in H^{2,0}_{per}(-1,1)$ and

$$E(u) = \int_{-1}^{1}\int_{-1}^{1} g(x-y)(u(x)-u(y))^2 dxdy$$

with a kernel $g : \mathbb{R} \to \mathbb{R}$ that satisfies the following conditions:

(g_1) $g(x+2) = g(x)$ for every $x \in \mathbb{R}$.
(g_2) $g(x) = g(-x)$ for every $x \in \mathbb{R}$.
(g_3) $0 < g_0 \le g(x) \le g_1$ for every $x \in \mathbb{R}$.
(g_4) $g\big|_{[0,1]} \in C^3[0,1]$.
(g_5) $g'(1) = 0$.

In particular, g is Lipschitz continuous on $\mathbb{R}$. For examples, see Figure 1.1. W is a potential as described above, the most common example being given by $W(v) = (v^2-1)^2$. The main result of the paper is the following.

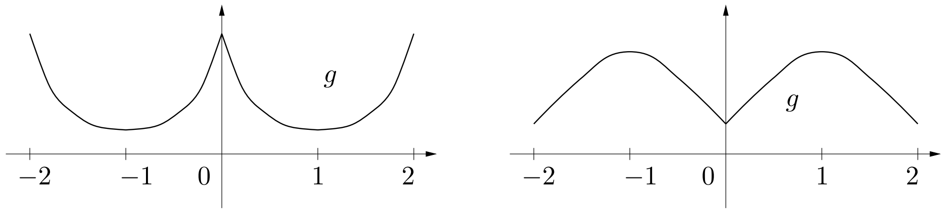


FIG. 1.1. *Examples of the kernel g.*

THEOREM 1.1. *For $\varepsilon > 0$, let I^ε be as defined by (1.2). Then there exists $N_0 \in \mathbb{N}$ such that for every even $N \geq N_0$, we find $\varepsilon_0 > 0$ and $\delta > 0$ satisfying the following condition:*

For every $\varepsilon \in (0, \varepsilon_0)$, there is a local minimizer $u_\varepsilon \in H^{2,0}_{per}(-1,1)$ of I^ε with respect to the $H^1(-1,1)$-norm such that

$$I^\varepsilon(u_\varepsilon) \leq I^\varepsilon(u) \quad \textit{for every} \quad u \in H^{2,0}_{per}(-1,1) \quad \textit{with} \quad \|u - \bar{u}_N\|_{H^1(-1,1)} < \delta$$

and

$$\lim_{\varepsilon \to 0} \|u_\varepsilon - \bar{u}_N\|_{H^1(-1,1)} = 0,$$

where $\bar{u}_N$ is a sawtooth function with N equidistant corners. More precisely, $\bar{u}_N$ is the unique function in $H^1(-1,1)$ satisfying $\int_{-1}^1 \bar{u}_N(x)dx = 0$ and

$$\bar{u}_N'(x) = (-1)^{i-1} \quad \textit{for} \quad x \in \left(-1 + \frac{2}{N}(i-1), -1 + \frac{2}{N}i\right), \quad i = 1, \ldots, N.$$

Remarks.

(a) The theorem states that if $\varepsilon > 0$ is chosen sufficiently small, the functionals I^ε have strongly oscillating local minimizers close to a sawtooth function with equidistant corners (see Figure 1.2). Near the corners of $\bar{u}_N$, the local minimizers u_ε have transition layers where u_ε' switches between -1 and 1 (see Figure 1.3).

(b) Since we consider periodic functions, only the case where N is even makes sense.

(c) To some extent, this theorem generalizes Theorem 1.1 in [21], since the functionals considered there can, in the case where $m = 0$, be written in the form (1.2) with $g = 1$. However, Ren and Wei found necessary conditions for critical points and thus identified local minimizers with an arbitrary number N

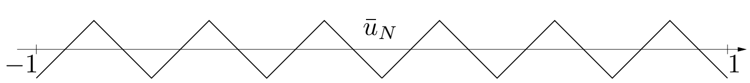


FIG. 1.2. *The function $\bar{u}_N$ for $N = 12$.*

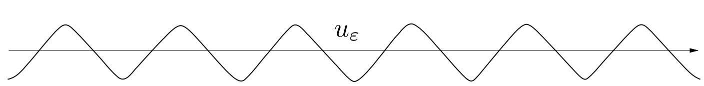


FIG. 1.3. *Near-periodic local minimizer u_ε resembling $\bar{u}_N$.*

of transition layers, which we cannot do here due to the very general nonlocal term E in (1.2). One has to let $N \to \infty$ to gain information, so that we are only able to characterize local minimizers with a sufficiently large number of oscillations.

Unfortunately, it was not possible to include all steps of the proof in this paper, since some of them are very long and technical. The complete proof is found in [24], which we cite whenever it is necessary. In this paper, we often make use of generic constants C whose value may change during estimations. Also, we sometimes make use of the 2-periodicity of functions and identify them with their periodical extensions to $\mathbb{R}$, so the integration domain may change from $(-1,1)$ to $(0,2)$, for example. $a \vee b$ denotes the maximum of two real numbers a and b, while $a \wedge b$ is their minimum. $C_0^\infty(a,b)$ is the space of C^∞-functions with compact support in (a,b). For a set $I \subset \mathbb{R}$, χ_I denotes the characteristic function of I.

2. The Γ-limit of $\;^{\varepsilon}$. The Γ-limit of the functionals I^ε is easily obtained via the Modica–Mortola theorem, which is found in [15], [16], [17] in its original form, while Alberti [1] gives an alternative proof. Although we consider periodic functions and use the L^2-topology rather than L^1, the proof can be easily transferred to the functionals I^ε, and only minor modifications are required (for details, see [24, Chapter 3], where Alberti's proof is transferred to this case). For a general theory on Γ-convergence, we refer the reader to [8]. We define the Hilbert space

$$Y = H^{1,0}_{per}(-1,1) = \left\{ u \in H^1(-1,1) \,|\, u(-1) = u(1),\ \int_{-1}^1 u(x)dx = 0 \right\},$$

equipped with the usual H^1-norm. We define the set of sawtooth functions with slope -1 and 1 only by

$$\mathcal{S}(-1,1) = \{u : [-1,1] \to \mathbb{R} \,|\, u' \in BV((-1,1),\{-1,1\})\} \tag{2.1}$$

and the set of periodic sawtooth functions with average zero by

$$\mathcal{S}^0_{per}(-1,1) = \left\{ u \in \mathcal{S}(-1,1) \,|\, u(-1) = u(1),\ \int_{-1}^1 u(x)dx = 0 \right\}. \tag{2.2}$$

We extend the functionals I^ε to Y by

$$I^\varepsilon(u) = \begin{cases} \displaystyle\int_{-1}^1 \varepsilon(u'')^2 + \frac{1}{\varepsilon}W(u')dx + E(u) & \text{for } u \in H^{2,0}_{per}(-1,1), \\ +\infty & \text{for } u \in Y \setminus H^{2,0}_{per}(-1,1) \end{cases} \tag{2.3}$$

and define $I : Y \to [0,\infty]$ by

$$I(u) = \begin{cases} A_0\#(Su' \cap [-1,1)) + E(u) & \text{for } u \in \mathcal{S}^0_{per}(-1,1), \\ +\infty & \text{for } u \in Y \setminus \mathcal{S}^0_{per}(-1,1), \end{cases} \tag{2.4}$$

where

$$A_0 = 2\int_{-1}^1 \sqrt{W(v)}dv$$

and the term $\#(Su' \cap [-1,1))$ counts the corners of u (i.e., the discontinuity points of u') on $[-1,1)$, which means that if the periodic extension of u has a corner at -1 (and thus at 1), this corner is counted once. The Modica–Mortola theorem yields that the integral in (2.3) Γ-converges to the counting term in (2.4). Since the nonlocal energy E is L^2-continuous (see Lemma 4.2 in [24]), this implies (cf. [8] or Proposition 2.11 (iv) in [4]) the following theorem.

THEOREM 2.1.

(i) *Let* $(u_\varepsilon)_{\varepsilon>0} \subset H^{2,0}_{per}(-1,1)$, $u \in Y$, *such that* $\|u_\varepsilon - u\|_{H^1(-1,1)} \to 0$ *for* $\varepsilon \to 0$. *Then*

$$I(u) \leq \liminf_{\varepsilon\to 0} I^\varepsilon(u_\varepsilon).$$

(ii) *For every* $u \in \mathcal{S}^0_{per}(-1,1)$, *we find a sequence* $(u_\varepsilon)_{\varepsilon>0} \subset H^{2,0}_{per}(-1,1)$ *such that* $\|u_\varepsilon - u\|_{H^1(-1,1)} \to 0$ *for* $\varepsilon \to 0$ *and*

$$\limsup_{\varepsilon\to 0} I^\varepsilon(u_\varepsilon) \leq I(u).$$

In particular, $I^\varepsilon \xrightarrow{\Gamma} I$ *in* Y.

The connection between local minimizers of singularly perturbed functionals and those of the Γ-limit has been studied by De Giorgi (cf. section 3 in [10]), Kohn and Sternberg (cf. Theorem 4.1 in [14]), and Ren and Wei (cf. Proposition 2.3 in [21]). However, these results assume that the Γ-limit I has an isolated local minimizer u, which cannot hold in this case due to the translation invariance of I. By considering orbits of functions rather than the functions themselves, one can transfer the proof of Proposition 2.3 in [21] to this case to obtain the theorem below (for details see [24, Chapter 4]). For $u \in Y$, we also denote by u the periodic extension of u to $\mathbb{R}$. Then, obviously, $u \in H^1_{loc}(\mathbb{R})$. For $\tau \in \mathbb{R}$, we define the τ-translation $T_\tau u : \mathbb{R} \to \mathbb{R}$ by

$$T_\tau u(x) = u(x-\tau) \quad \text{for every} \quad x \in \mathbb{R},$$

and the orbit of u is defined as the set

$$\mathcal{O}(u) = \{T_\tau u \,|\, \tau \in \mathbb{R}\}$$

of all translations of u. Then the following holds.

THEOREM 2.2. *Let* $\bar{u} \in \mathcal{S}^0_{per}(-1,1)$. *Assume that there exists a* $\delta > 0$ *such that for every* $u \in \mathcal{S}^0_{per}(-1,1)$ *with* $\|u - \bar{u}\|_{H^1(-1,1)} < \delta$ *and* $u \notin \mathcal{O}(\bar{u})$ *we have*

$$I(u) > I(\bar{u}).$$

Then there exists $\varepsilon_0 > 0$ *such that for every* $\varepsilon \in (0, \varepsilon_0)$ *we find* $u_\varepsilon \in H^{2,0}_{per}(-1,1)$ *with* $\|u_\varepsilon - \bar{u}\|_{H^1(-1,1)} < \frac{\delta}{2}$ *and*

$$I^\varepsilon(u_\varepsilon) \leq I^\varepsilon(u) \quad \textit{for every} \quad u \in H^{2,0}_{per}(-1,1) \quad \textit{satisfying} \quad \|u - \bar{u}\|_{H^1(-1,1)} < \frac{\delta}{2}.$$

Furthermore,

$$\lim_{\varepsilon\to 0} \|u_\varepsilon - \bar{u}\|_{H^1(-1,1)} = 0.$$

In particular, for every $\varepsilon \in (0, \varepsilon_0)$, u_ε *is an* H^1*-local minimizer of* I^ε *in* $H^{2,0}_{per}(-1,1)$.

Thus, Theorem 1.1 follows if we can show that for large N, the regular sawtooth function $\bar{u}_N$ is a strict local minimizer of I on $\mathcal{S}^0_{per}(-1,1)$ "up to translation" in the sense of the condition in Theorem 2.2. Since I is a sum of a counting term and the nonlocal energy E, it is sufficient to identify $\bar{u}_N$ (N large) as a strict H^1-local minimizer of E (up to translation) in the set of all $u \in \mathcal{S}^0_{per}(-1,1)$ with N corners on $[-1,1)$, since—roughly speaking—a sufficiently small H^1-ball around $\bar{u}_N$ can only contain such sawtooth functions which have at least N corners. A larger number of corners, however, would increase I by an integer, so that functions with more than N corners need not be considered.

3. Perturbation of periodic sawtooth functions. In order to show that for large N the sawtooth function $\bar{u}_N$ with N equidistant corners is a local minimizer of the nonlocal energy E on the set of all $u \in \mathcal{S}^0_{per}(-1,1)$ with $\#(Su' \cap [-1,1)) = N$, we consider all possible perturbations of $\bar{u}_N$ in this set. The periodicity will be dropped for the moment, and we consider E as a functional on $\mathcal{S}(-1,1)$. The corners of $\bar{u}_N$ are given by

$$t_i = -1 + \frac{2}{N}i \quad \text{for} \quad i = 0, \dots, N, \tag{3.1}$$

and we have

$$\bar{u}_N'(x) = (-1)^{i-1} \quad \text{for} \quad x \in (t_{i-1}, t_i)$$

for $i = 1, \dots, N$. Let $r = (r_1, \dots, r_{N-1}) \in \mathbb{R}^{N-1}$ be an arbitrary vector, and set $r_0 = r_N = 0$. For $\delta \in (-\delta_0, \delta_0)$, we define the shifted corners by

$$t_i^{(\delta)} = t_i + r_i\delta \quad \text{for} \quad i = 0, \dots, N, \tag{3.2}$$

where $\delta_0 > 0$ is chosen such that for every $\delta \in (-\delta_0, \delta_0)$, the monotonicity

$$-1 = t_0^{(\delta)} < t_1^{(\delta)} < \cdots < t_{N-1}^{(\delta)} < t_N^{(\delta)} = 1 \tag{3.3}$$

still holds. For every $\delta \in (-\delta_0, \delta_0)$, we define $u_\delta \in \mathcal{S}(-1,1)$ as the unique function satisfying

$$\begin{cases} u_\delta(-1) = \bar{u}_N(-1) = -\frac{1}{N}, \\ u_\delta'(x) = (-1)^{i-1} \quad \text{for} \quad x \in (t_{i-1}^{(\delta)}, t_i^{(\delta)}), \quad i = 1, \dots, N \end{cases} \tag{3.4}$$

(see Figure 3.1).

We define the function $F_r : (-\delta_0, \delta_0) \to \mathbb{R}$ by

$$F_r(\delta) = E(u_\delta) \tag{3.5}$$

for every $\delta \in (-\delta_0, \delta_0)$. Clearly, we have $u_0 = \bar{u}_N$, and thus $F_r(0) = E(\bar{u}_N)$.

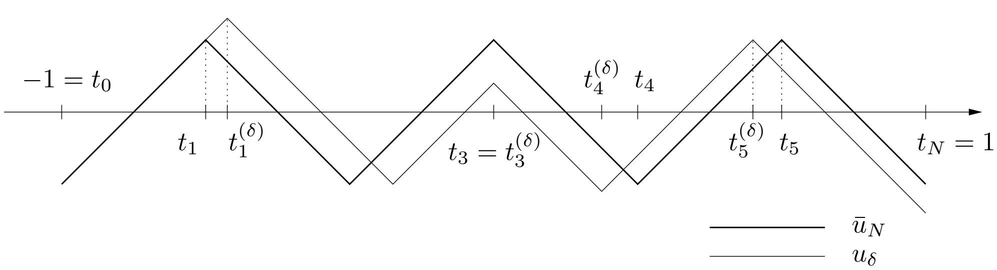


FIG. 3.1. *Example of a perturbation u_δ of $\bar{u}_N$ with $r_3 = 0$; case $N = 6$.*

Remarks.

(i) Since the corner at -1 is fixed, this definition does not cover all possible perturbations of $\bar{u}_N$ but indeed all relevant ones, which is due to the translation invariance of I.

(ii) To avoid ugly notation, we drop the dependence of u_δ on r.

(iii) The zero average condition $\int u = 0$ can be ignored due to the invariance of the functionals under the addition of constants.

We introduce a bilinear form by

$$(u, v)_g = \int_{-1}^{1} \int_{-1}^{1} g(x-y)(u(x)-u(y))(v(x)-v(y))dxdy. \tag{3.6}$$

Since we want to show that $\bar{u}_N$ is a local minimizer of E in the sense described above, we have to consider derivatives of F_r. To compute F_r', we need the following technical lemma.

LEMMA 3.1. *With the above definitions, let $\delta \in (-\delta_0, \delta_0)$. Then for a.e. $x \in (-1, 1)$, we have*

$$\begin{aligned} &\lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(u_{\delta+\varepsilon}(x) - u_\delta(x)) = 2 \sum_{i=1}^{N} \sum_{j=1}^{i-1} (-1)^{j-1} r_j \chi_{(t_{i-1}^{(\delta)}, t_i^{(\delta)})}(x), \\ &\lim_{\varepsilon \to 0} (u_{\delta+\varepsilon}(x) - u_\delta(x)) = 0. \end{aligned} \tag{3.7}$$

Furthermore, $|u_\delta(x)|$, $|u_{\delta+\varepsilon}(x)|$, $|\frac{1}{\varepsilon}(u_{\delta+\varepsilon}(x) - u_\delta(x))| \le C$ with C independent of δ, ε, x (but dependent on r).

Proof. For fixed $\delta \in (-\delta_0, \delta_0)$, choose $\varepsilon_0 > 0$ such that $(\delta - \varepsilon_0, \delta + \varepsilon_0) \subset (-\delta_0, \delta_0)$ and $t_{i-1}^{(\delta+\varepsilon)} \vee t_{i-1}^{(\delta)} < t_i^{(\delta+\varepsilon)} \wedge t_i^{(\delta)}$ for every $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$ and $i \in \{1, \dots, N\}$. Then, according to (3.4), we have

$$(u_{\delta+\varepsilon}' - u_\delta')(x) = 0 \ \text{ for } \ x \in (t_{i-1}^{(\delta)}, t_i^{(\delta)}) \cap (t_{i-1}^{(\delta+\varepsilon)}, t_i^{(\delta+\varepsilon)}) = (t_{i-1}^{(\delta+\varepsilon)} \vee t_{i-1}^{(\delta)}, t_i^{(\delta+\varepsilon)} \wedge t_i^{(\delta)})$$

for $i = 1, \dots, N$. Thus, up to a finite subset of $[-1, 1]$, $(u_{\delta+\varepsilon}' - u_\delta')(x) \neq 0$ can only hold if $x \in (t_i^{(\delta+\varepsilon)} \wedge t_i^{(\delta)}, t_i^{(\delta+\varepsilon)} \vee t_i^{(\delta)})$ for an $i \in \{1, \dots, N-1\}$. Consider a fixed $x \in (t_{i-1}^{(\delta)}, t_i^{(\delta)})$, $i \in \{1, \dots, N\}$. Then $x \in (t_{i-1}^{(\delta+\varepsilon)} \vee t_{i-1}^{(\delta)}, t_i^{(\delta+\varepsilon)} \wedge t_i^{(\delta)})$ for small $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$ since $t_j^{(\delta+\varepsilon)} \to t_j^{(\delta)}$ for $\varepsilon \to 0$. Thus, using $u_\delta(-1) = u_{\delta+\varepsilon}(-1)$ we obtain

$$u_{\delta+\varepsilon}(x) - u_\delta(x) = \int_{-1}^{x} (u_{\delta+\varepsilon}' - u_\delta')(\xi) d\xi = \sum_{j=1}^{i-1} \int_{t_j^{(\delta+\varepsilon)} \wedge t_j^{(\delta)}}^{t_j^{(\delta+\varepsilon)} \vee t_j^{(\delta)}} (u_{\delta+\varepsilon}' - u_\delta')(\xi) d\xi \tag{3.8}$$

for $x \in (t_{i-1}^{(\delta)}, t_i^{(\delta)})$ if ε is sufficiently small. Let $j \in \{1, \dots, N\}$. If $r_j \varepsilon > 0$, we have $t_j^{(\delta+\varepsilon)} > t_j^{(\delta)}$, so that by (3.4) and (3.2), we get

$$\int_{t_j^{(\delta+\varepsilon)} \wedge t_j^{(\delta)}}^{t_j^{(\delta+\varepsilon)} \vee t_j^{(\delta)}} (u_{\delta+\varepsilon}' - u_\delta')(\xi) d\xi = \int_{t_j^{(\delta)}}^{t_j^{(\delta+\varepsilon)}} (u_{\delta+\varepsilon}' - u_\delta')(\xi) d\xi = 2(-1)^{j-1} r_j \varepsilon$$

for ε sufficiently small since $(t_j^{(\delta)}, t_j^{(\delta+\varepsilon)}) \subset (t_j^{(\delta)}, t_{j+1}^{(\delta)}) \cap (t_{j-1}^{(\delta+\varepsilon)}, t_j^{(\delta+\varepsilon)})$. If $r_j \varepsilon < 0$, we analogously show the same formula, which obviously also holds if $r_j \varepsilon = 0$. Setting this into (3.8), we obtain

$$u_{\delta+\varepsilon}(x) - u_\delta(x) = 2\varepsilon \sum_{j=1}^{i-1} (-1)^{j-1} r_j \quad \text{for} \quad x \in (t_{i-1}^{(\delta)}, t_i^{(\delta)}), \quad i = 1, \dots, N,$$

if ε is small, and thus

$$u_{\delta+\varepsilon} - u_\delta = 2\varepsilon \sum_{i=1}^{N} \sum_{j=1}^{i-1} (-1)^{j-1} r_j \chi_{(t_{i-1}^{(\delta)}, t_i^{(\delta)})}$$

for a.e. $x \in (-1, 1)$ if ε is sufficiently small, which implies (3.7), while the other conditions are obvious. ∎

THEOREM 3.2. *For arbitrary* $r = (r_1, \ldots, r_{N-1}) \in \mathbb{R}^{N-1}$*, choose* $\delta_0 > 0$ *as above. Then the function* $F_r : (-\delta_0, \delta_0) \to \mathbb{R}$ *as defined in* (3.5) *is differentiable, and for every* $\delta \in (-\delta_0, \delta_0)$*, we have*

$$F_r'(\delta) = 4 \sum_{i=1}^{N} \alpha_i h_r^{(i)}(\delta), \tag{3.9}$$

where $\alpha_i \in \mathbb{R}$*,* $h_r^{(i)} : (-\delta_0, \delta_0) \to \mathbb{R}$ *are given by*

$$\alpha_i = \sum_{j=1}^{i-1} (-1)^{j-1} r_j, \quad h_r^{(i)}(\delta) = \left(\chi_{(t_{i-1}^{(\delta)}, t_i^{(\delta)})}, u_\delta \right)_g \tag{3.10}$$

for $i = 1, \ldots, N$.

Proof. Let $\delta \in (-\delta_0, \delta_0)$, $\varepsilon_0 > 0$ such that $(\delta - \varepsilon_0, \delta + \varepsilon_0) \subset (-\delta_0, \delta_0)$. Then for every $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$

$$F_r(\delta+\varepsilon) - F_r(\delta) = (u_{\delta+\varepsilon}, u_{\delta+\varepsilon})_g - (u_\delta, u_\delta)_g = (u_{\delta+\varepsilon} - u_\delta, u_{\delta+\varepsilon} + u_\delta)_g. \tag{3.11}$$

Hence

$$\frac{1}{\varepsilon}(F_r(\delta+\varepsilon) - F_r(\delta)) = \left(\frac{1}{\varepsilon} \left(u_{\delta+\varepsilon} - u_\delta \right), u_{\delta+\varepsilon} + u_\delta \right)_g,$$

which due to Lemma 3.1 and Lebesgue's convergence theorem yields

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(F_r(\delta+\varepsilon) - F_r(\delta)) = 4 \sum_{i=1}^{N} \sum_{j=1}^{i-1} (-1)^{j-1} r_j \left(\chi_{(t_{i-1}^{(\delta)}, t_i^{(\delta)})}, u_\delta \right)_g,$$

i.e.,

$$F_r'(\delta) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(F_r(\delta+\varepsilon) - F_r(\delta)) = 4 \sum_{i=1}^{N} \alpha_i h_r^{(i)}(\delta),$$

with $\alpha_i \in \mathbb{R}$, $h_r^{(i)} : (-\delta_0, \delta_0) \to \mathbb{R}$, $i = 1, \ldots, N$, as given in (3.10), which completes the proof of Theorem 3.2. ∎

As for the second derivative, we have the following result, the proof of which is extremely technical, so that we refer the reader to Theorem 5.3 in [24], where a detailed proof can be found.

THEOREM 3.3. *For arbitrary* $r = (r_1, \ldots, r_{N-1}) \in \mathbb{R}^{N-1}$*, choose* $\delta_0 > 0$ *as above. Then the function* $F_r : (-\delta_0, \delta_0) \to \mathbb{R}$ *as defined in* (3.5) *satisfies* $F_r \in C^2(-\delta_0, \delta_0)$*,* *and for every $\delta \in (-\delta_0, \delta_0)$, we have*

$$
\begin{aligned}
F_r''(\delta) &= 8\sum_{i=1}^N \sum_{k=1}^N \alpha_i\alpha_k \left(\chi_{(t_{k-1}^{(\delta)}, t_k^{(\delta)})}, \chi_{(t_{i-1}^{(\delta)}, t_i^{(\delta)})}\right)_g \\
&\quad + 8\sum_{i=1}^N \alpha_i r_i \int_{-1}^1 g(t_i^{(\delta)} - y)(u_\delta(t_i^{(\delta)}) - u_\delta(y))dy \\
&\quad - 8\sum_{i=1}^N \alpha_i r_{i-1} \int_{-1}^1 g(t_{i-1}^{(\delta)} - y)(u_\delta(t_{i-1}^{(\delta)}) - u_\delta(y))dy,
\end{aligned} \tag{3.12}
$$

with α_i, $i = 1, \ldots, N$, defined as in (3.10).

4. Criticality. From Theorem 3.2, we conclude that $\bar{u}_N$ is critical for the nonlocal energy in the following sense.

THEOREM 4.1. *Let $(r_1, \ldots, r_{N-1}) \in \mathbb{R}^{N-1}$, and choose $\delta_0 > 0$ as above. For every $\delta \in (-\delta_0, \delta_0)$, let u_δ be as defined by* (3.4). *Then*

$$F_r'(0) = 0. \tag{4.1}$$

Proof. Theorem 3.2 and (3.2) yield

$$F_r'(0) = 4\sum_{i=1}^N \alpha_i(\chi_{(t_{i-1}, t_i)}, \bar{u}_N)_g,$$

with $\alpha_1, \ldots, \alpha_N$ as defined in (3.10). We aim to show that

$$(\chi_{(t_{i-1}, t_i)}, \bar{u}_N)_g = 0$$

for $i = 1, \ldots, N$, which yields (4.1). The symmetry of g implies that the function $\varphi(x, y) = g(x - y)(\bar{u}_N(x) - \bar{u}_N(y))$ satisfies $\varphi(y, x) = -\varphi(x, y)$, so that

$$\int_{t_{i-1}}^{t_i} \int_{t_{i-1}}^{t_i} g(x - y)(\bar{u}_N(x) - \bar{u}_N(y))dxdy = 0.$$

Thus, using the symmetry and the 2-periodicity of g,

$$
\begin{aligned}
(\chi_{(t_{i-1}, t_i)}, \bar{u}_N)_g &= \int_{-1}^1 \int_{-1}^1 g(x - y)(\chi_{(t_{i-1}, t_i)}(x) - \chi_{(t_{i-1}, t_i)}(y))(\bar{u}_N(x) - \bar{u}_N(y))dxdy \\
&= 2\int_{-1}^{t_{i-1}} \int_{t_{i-1}}^{t_i} g(x - y)(\bar{u}_N(x) - \bar{u}_N(y))dxdy \\
&\quad + 2\int_{t_i}^1 \int_{t_{i-1}}^{t_i} g(x - y)(\bar{u}_N(x) - \bar{u}_N(y))dxdy \\
&= 2\int_{-1}^1 \int_{t_{i-1}}^{t_i} g(x - y)(\bar{u}_N(x) - \bar{u}_N(y))dxdy \\
&= 2\int_{t_{i-1}}^{t_i} \bar{u}_N(x)dx \int_{-1}^1 g(y)dy - 2\int_{t_{i-1}}^{t_i} \int_{-1}^1 g(y - x)\bar{u}_N(y)dydx \\
&= -2\int_{t_{i-1}}^{t_i} \int_{-1}^1 g(y - x)\bar{u}_N(y)dydx,
\end{aligned}
$$

since $\int_{t_{i-1}}^{t_i} \bar{u}_N(x)dx = 0$, which is easy to see. We claim that

$$x \mapsto \int_{-1}^{1} g(y-x)\bar{u}_N(y)dy$$

is antisymmetric on (t_{i-1}, t_i), which implies $(\chi_{(t_{i-1},t_i)}, \bar{u}_N)_g = 0$. To show this, we set $x_1 = t_{i-1} + \beta$, $x_2 = t_i - \beta$ with $\beta \in [0, \frac{t_{i-1}+t_i}{2}]$. The 2-periodicity of $\bar{u}_N$ and g and the antisymmetry of $\bar{u}_N$ with respect to the center of each interval $[t_{i-1}, t_i]$ yield

$$\begin{aligned}\int_{-1}^{1} g(y-x_1)\bar{u}_N(y)dy &= \int_{-1}^{1} g(y)\bar{u}_N(y+x_1)dy = \int_{-1}^{1} g(y)\bar{u}_N(y+t_{i-1}+\beta)dy \\ &= -\int_{-1}^{1} g(y)\bar{u}_N(t_i-\beta-y)dy = -\int_{-1}^{1} g(y)\bar{u}_N(x_2-y)dy \\ &= -\int_{-1}^{1} g(y)\bar{u}_N(x_2+y)dy = -\int_{-1}^{1} g(y-x_2)\bar{u}_N(y)dy,\end{aligned}$$

which implies the antisymmetry. $\square$

It is obvious that the perturbation u_δ of $\bar{u}_N$ as defined in (3.4) is not necessarily periodic (see Figure 3.1). However, to show the local minimizing property for an equidistant sawtooth function, we have to treat the periodicity as a constraint which is given as follows.

LEMMA 4.2. *Let $\delta \in (-\delta_0, \delta_0) \setminus \{0\}$, where $\delta_0 > 0$ is chosen as above. Then u_δ as given by* (3.4) *is periodic (i.e., $u_\delta(-1) = u_\delta(-1)$) if and only if*

$$\sum_{i=1}^{N-1} (-1)^{i-1} r_i = 0, \tag{4.2}$$

which in terms of (3.10) *is equivalent to $\alpha_N = 0$.*

Proof. Periodicity holds if and only if $u_\delta(1) = u_\delta(-1) = -\frac{1}{N}$, i.e.,

$$\begin{aligned}0 &= \int_{-1}^{1} u'_\delta(\xi)d\xi = \sum_{i=1}^{N} \int_{t_{i-1}^{(\delta)}}^{t_i^{(\delta)}} (-1)^{i-1} d\xi = \sum_{i=1}^{N} (-1)^{i-1} \left(t_i - t_{i-1} + (r_i - r_{i-1})\delta\right) \\ &= \sum_{i=1}^{N} (-1)^{i-1} \left(\frac{2}{N} + (r_i - r_{i-1})\delta\right) = \delta \sum_{i=1}^{N} (-1)^{i-1}(r_i - r_{i-1}) = 2\delta \sum_{i=1}^{N-1} (-1)^{i-1} r_i,\end{aligned}$$

where we used the fact that N is even and $r_0 = r_N = 0$. Since $\delta \neq 0$, the condition now follows. $\square$

Application of Theorem 3.3 leads to the following property on the second variation. Since the proof is a very technical but straightforward computation, we refer the reader to Theorem 6.4 in [24].

THEOREM 4.3. *For $N \in \mathbb{N}$ even, let $\bar{u}_N \in \mathcal{S}^0_{per}(-1,1)$ as defined in Theorem* 1.1. *Let $(r_1, \ldots, r_{N-1}) \in \mathbb{R}^{N-1}$ such that the periodicity condition* (4.2) *holds, and choose $\delta_0 > 0$ as before. Let $F_r : (-\delta_0, \delta_0) \to \mathbb{R}$ as given in* (3.5). *Let $\alpha = (\alpha_1, \ldots, \alpha_N)^t \in \mathbb{R}^N$ as defined in* (3.10) *(in particular, $\alpha_1 = \alpha_N = 0$). Then $F_r \in C^2(-\delta_0, \delta_0)$ with*

$$\frac{1}{8} F''_r(0) = \frac{2}{N} \alpha^t A_N \alpha,$$

where $A_N = (a_{|i-j|}^{(N)})_{i,j=1,\dots,N} \in \mathbb{R}^{N \times N}$ *is a symmetric circulant matrix, i.e.,* $a_k^{(N)} = a_{N-k}^{(N)}$ *for* $k = 1, \dots, N-1$. *The coefficients are given by*

$$
\begin{aligned}
a_0^{(N)} &= b_0^{(N)} + c_0^{(N)} + d_0^{(N)} \\
a_1^{(N)} = a_{N-1}^{(N)} &= b_1^{(N)} + c_1^{(N)} + d_1^{(N)} = \tfrac{1}{2}(b_0^{(N)} - d_0^{(N)}) + c_1^{(N)} \\
a_k^{(N)} &= c_k^{(N)} \quad \text{for} \quad k = 2, \dots, N-2,
\end{aligned} \tag{4.3}
$$

where

$$
\begin{aligned}
b_0^{(N)} &= \int_{-1}^{1} g(y)dy, \quad b_1^{(N)} = \frac{1}{2} b_0^{(N)}, \\
d_0^{(N)} &= N(-1)^{\frac{N}{2}+1} \int_{-1}^{1} g(y)\bar{u}_N(y)dy, \quad d_1^{(N)} = -\frac{1}{2} d_0^{(N)},
\end{aligned} \tag{4.4}
$$

and

$$
c_k^{(N)} = -N \int_0^{\frac{2}{N}} \int_{\frac{2}{N}k}^{\frac{2}{N}(k+1)} g(x-y)dydx \tag{4.5}
$$

for $k = 0, \dots, N-1$. *The circulant matrix* $C_N = (c_{|i-j|}^{(N)})_{i,j=1,\dots,N}$ *can alternatively be defined as* $C_N = (c_{ij}^{(N)})_{i,j=1,\dots,N} \in \mathbb{R}^{N \times N}$ *with*

$$
c_{ij}^{(N)} = -N \int_{\frac{2}{N}(i-1)}^{\frac{2}{N}i} \int_{\frac{2}{N}(j-1)}^{\frac{2}{N}j} g(x-y)dydx \tag{4.6}
$$

for $i, j = 1, \dots, N$.

5. Eigenvalues of $_N$**.** Theorem 4.3 shows that the positivity of the second variation is connected to a definiteness condition for the matrix A_N. Since the matrix is circulant, its eigenvalues and eigenvectors can be written down explicitly. Using Theorem 3.2.2 in [9] (or see Chapter 3 in [11]), the eigenvalues are

$$
\lambda_t^{(N)} = a_0^{(N)} + (-1)^t a_{\frac{N}{2}}^{(N)} + 2 \sum_{k=1}^{\frac{N}{2}-1} a_k^{(N)} \cos\left[\frac{2\pi kt}{N}\right] \quad \text{for} \quad t = 0, \dots, \frac{N}{2} \tag{5.1}
$$

with coefficients given in Theorem 4.3, where $\lambda_1^{(N)}, \dots, \lambda_{\frac{N}{2}-1}^{(N)}$ are double eigenvalues. For every $t \in \{0, \dots, \frac{N}{2}\}$, an eigenvector corresponding to $\lambda_t^{(N)}$ is given by

$$
\begin{aligned}
v_t &= (v_t^{(0)}, \dots, v_t^{(N-1)}) \in \mathbb{R}^N, \\
v_t^{(k)} &= \cos\left[\frac{2\pi kt}{N}\right], \quad k = 0, \dots, N-1.
\end{aligned} \tag{5.2}
$$

A long but straightforward calculation (see [24, proof of Theorem 7.3]) yields the following.

THEOREM 5.1. *For every even number* $N \in \mathbb{N}$, *we have*

$$
\lambda_0^{(N)} = \lambda_{\frac{N}{2}}^{(N)} = 0. \tag{5.3}
$$

Using this property and the fact that in view of (5.2), the associated eigenvectors are given by $v_0 = (1, \dots, 1)$ and $v_{\frac{N}{2}} = (1, -1, \dots, 1, -1)$, it is easy to show that if the double eigenvalues $\lambda_t^{(N)}$, $t = 1, \dots, \frac{N}{2} - 1$, are positive, then the quadratic form associated with A_N is positive definite on the $(N-2)$-dimensional subspace

$$X_N = \{\alpha = (\alpha_1, \dots, \alpha_N) \in \mathbb{R}^N \mid \alpha_1 = \alpha_N = 0\}$$

(see Theorem 7.4 in [24]). Thus, if we can show that $\lambda_t^{(N)} > 0$ for $t = 1, \dots, \frac{N}{2} - 1$, it follows from Theorem 4.3 that $F_r''(0) > 0$ for every $r \in \mathbb{R}^{N-1}$ satisfying (4.2), since this periodicity condition is equivalent to $\alpha_N = 0$ (note that $\alpha_1 = 0$ is always the case). Another very technical approach (see Chapter 8 in [24]) shows that the condition $F_r''(0) > 0$ for every r satisfying (4.2) indeed implies that $\bar{u}_N$ is an H^1-local minimizer of E on the set

$$\{u \in \mathcal{S}^0_{per}(-1,1) | \#(Su' \cap [-1,1)) = N\},$$

so that $\bar{u}_N$ is an H^1-local minimizer of I on $\mathcal{S}^0_{per}(-1,1)$, which together with Theorem 2.2 yields Theorem 1.1. The most important idea in the proof of this correlation is to rewrite the nonlocal energy E—viewed as a functional $E : \mathcal{S}(-1,1) \to \mathbb{R}$—as a function of the corners $t_1, \dots, t_N$ of sawtooth functions and to apply Ljusternik's sufficient condition for local minima under equality constraints (see [25, Theorem 43.D]). In this case, the constraint is given by a periodicity condition similar to (4.2).

6. Asymptotic behavior of eigenvalues. In view of the remarks in the previous section, our aim is to show that the existence of near-periodic local minimizers of I^ε for small $\varepsilon > 0$ is reduced to the positivity of the eigenvalues $\lambda_1^{(N)}, \dots, \lambda_{\frac{N}{2}-1}^{(N)}$ of A_N for large N. A first step is to assume the existence of arbitrarily large even numbers N for which at least one of these eigenvalues is nonpositive. We obtain the following "critical index" property.

THEOREM 6.1. *For every even number $N \in \mathbb{N}$, let $\lambda_t^{(N)}$, $t = 0, \dots, \frac{N}{2}$, be the eigenvalues of A_N as given in* (5.1). *Assume there is a sequence of even numbers $(N_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ with $N_n \nearrow \infty$ for $n \to \infty$ and a sequence $(t_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ with $t_n \in \{1, \dots, \frac{N_n}{2} - 1\}$ for every $n \in \mathbb{N}$ such that*

$$\limsup_{n\to\infty} \lambda_{t_n}^{(N_n)} \le 0.$$

Then we find a subsequence of $(N_n)_{n\in\mathbb{N}}$ (not relabelled) such that

$$\lim_{n\to\infty} \lambda_{t_n}^{(N_n)} = 0, \quad \lim_{n\to\infty} \frac{t_n}{N_n} = \frac{1}{2}.$$

The proof needs some preparations. First, we will compare A_N to a different circulant symmetric matrix $\bar{A}_N = (\bar{a}_{|i-j|}^{(N)})_{i,j=1,\dots,N} \in \mathbb{R}^{N\times N}$ with coefficients

$$\begin{array}{lcl} \bar{a}_0^{(N)} & = & b_0^{(N)} + c_0^{(N)}, \\ \bar{a}_1^{(N)} = \bar{a}_{N-1}^{(N)} & = & b_1^{(N)} + c_1^{(N)} = \frac{1}{2} b_0^{(N)} + c_1^{(N)}, \\ \bar{a}_k^{(N)} & = & c_k^{(N)} \quad \text{for} \quad k = 2, \dots, N-2, \end{array} \tag{6.1}$$

where the coefficients $c_i^{(N)}$ and $b_i^{(N)}$ are defined as in Theorem 4.3. Thus, $\bar{A}_N$ is the matrix that results from A_N by removing the $d_i^{(N)}$-terms. As remarked in the previous chapter, the eigenvalues are given explicitly by

$$\bar{\lambda}_t^{(N)} = \bar{a}_0^{(N)} + (-1)^t \bar{a}_{\frac{N}{2}}^{(N)} + 2 \sum_{k=1}^{\frac{N}{2}-1} \bar{a}_k^{(N)} \cos\left[\frac{2\pi kt}{N}\right] \quad \text{for} \quad t = 0, \ldots, \frac{N}{2} \tag{6.2}$$

for $t = 0, \ldots, \frac{N}{2}$ with corresponding eigenvectors $v_t = (v_t^{(0)}, \ldots, v_t^{(N-1)}) \in \mathbb{R}^N$, where

$$v_t^{(k)} = \cos\left[\frac{2\pi kt}{N}\right], \quad k = 0, \ldots, N-1. \tag{6.3}$$

For large N, the eigenvalues of $\bar{A}_N$ are close to those of A_N.

LEMMA 6.2. *There exists a constant $C > 0$, dependent only on g, such that for every even $N \in \mathbb{N}$, we have*

$$|\lambda_t^{(N)} - \bar{\lambda}_t^{(N)}| \leq \frac{C}{N} \tag{6.4}$$

for $t = 0, \ldots, \frac{N}{2}$.

Proof. From (5.1), (6.2), (6.1), and the definition of A_N in Theorem 4.3 we conclude that

$$\begin{aligned} \lambda_t^{(N)} - \bar{\lambda}_t^{(N)} &= d_0^{(N)} + 2d_1^{(N)} \cos\left[\frac{2\pi t}{N}\right] = d_0^{(N)}\left[1 - \cos\left[\frac{2\pi t}{N}\right]\right] \\ &= N(-1)^{\frac{N}{2}+1} \int_{-1}^{1} g(y)\bar{u}_N(y)dy \left[1 - \cos\left[\frac{2\pi t}{N}\right]\right]. \end{aligned}$$

Since $\int_{t_{i-1}}^{t_i} \bar{u}_N(y)dy = 0$ for $i = 1, \ldots, N$ and $|\bar{u}_N| \leq \frac{1}{N}$ by definition of $\bar{u}_N$, we obtain

$$\begin{aligned} |\lambda_t^{(N)} - \bar{\lambda}_t^{(N)}| &\leq 2N \left|\int_{-1}^{1} g(y)\bar{u}_N(y)dy\right| = 2N \left|\sum_{i=1}^{N} \int_{t_{i-1}}^{t_i} g(y)\bar{u}_N(y)dy\right| \\ &= 2N \left|\sum_{i=1}^{N} \int_{t_{i-1}}^{t_i} (g(y) - g(t_{i-1}))\bar{u}_N(y)dy\right| \leq 2 \sum_{i=1}^{N} \int_{t_{i-1}}^{t_i} |g(y) - g(t_{i-1})|dy \\ &\leq C \sum_{i=1}^{N} \int_{t_{i-1}}^{t_i} |y - t_{i-1}|dy \leq C \sum_{i=1}^{N} (t_i - t_{i-1})^2 \leq \frac{C}{N}, \end{aligned}$$

where we used the Lipschitz continuity of g and $t_i - t_{i-1} = \frac{2}{N}$ (cf. (3.1)). ∎

We also need the boundedness of the eigenvalues.

LEMMA 6.3. *There is a constant $C > 0$, dependent only on g, such that for every even $N \in \mathbb{N}$, we have*

$$|\lambda_t^{(N)}|, |\bar{\lambda}_t^{(N)}| \leq C \tag{6.5}$$

for $t = 0, \ldots, \frac{N}{2}$.

Proof. By Lemma 6.2, we only have to show the estimate for $\bar{\lambda}_t^{(N)}$. Let $\bar{\lambda}_t^{(N)}$, $t \in \{0, \ldots, \frac{N}{2}\}$, be an eigenvalue of $\bar{A}_N$ and $\alpha = (\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N$ a corresponding normed eigenvector of A_N (i.e., $\|\alpha\| = 1$) and set $\alpha_{N+1} = \alpha_1$. Then, by definition of $\bar{A}_N$ (cf. (6.1)) and the characterization of the matrix C_N in Theorem 4.3,

$$\bar{\lambda}_t^{(N)} = \bar{\lambda}_t^{(N)} \alpha^t \alpha = \alpha^t \bar{A}_N \alpha = b_0^{(N)} \sum_{i=1}^{N} (\alpha_i^2 + \alpha_i \alpha_{i+1}) + \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j c_{ij}^{(N)}.$$

From (4.6) and the boundedness of g, we deduce $|c_{ij}^{(N)}| \le \frac{C}{N}$ for every $i, j = 1, \ldots, N$, and we conclude, using the Cauchy–Schwarz inequality, $\alpha_{N+1} = \alpha_1$, and $\|\alpha\| = 1$, that

$$|\bar{\lambda}_t^{(N)}| \le b_0^{(N)} \sum_{i=1}^N (\alpha_i^2 + |\alpha_i \alpha_{i+1}|) + \sum_{i=1}^N \sum_{j=1}^N |\alpha_i \alpha_j c_{ij}^{(N)}| \le 2b_0^{(N)} \sum_{i=1}^N \alpha_i^2 + \frac{C}{N} \left(\sum_{i=1}^N |\alpha_i| \right)^2$$
$$\le 2b_0^{(N)} + \frac{C}{N} \left(\sum_{i=1}^N |\alpha_i|^2 \right) \left(\sum_{i=1}^N 1^2 \right) \le 2b_0^{(N)} + C \le C,$$

and the proof of the lemma is complete. ∎

Finally, we need the following property, which is easy to show (cf. [24, Lemma 9.3]).

LEMMA 6.4. *Let $\mathcal{B} : L^2(0,2) \times L^2(0,2) \to \mathbb{R}$, given by*

$$\mathcal{B}(\psi, \varphi) = 2 \int_{-1}^1 g(y) dy \int_0^2 \psi(x)\varphi(x) dx - 2 \int_0^2 \int_0^2 g(x-y)\psi(x)\varphi(y) dy dx. \tag{6.6}$$

Then $\mathcal{B}$ is a continuous, symmetric bilinear form.

We can now show Theorem 6.1.

Sketch of the proof of Theorem 6.1. Let $(N_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ with $N_n \nearrow \infty$ for $n \to \infty$, $(t_n)_{n\in\mathbb{N}} \subset \mathbb{N}$ as described. Due to Lemma 6.3, we may choose a subsequence (not relabelled) with $\lim_{n\to\infty} \lambda_{t_n}^{(N_n)} = \lambda \le 0$. For the eigenvalues $\bar{\lambda}_t^{(N)}$ of $\bar{A}_N$ (cf. (6.2)), Lemma 6.2 then yields

$$\lim_{n\to\infty} \bar{\lambda}_{t_n}^{(N_n)} = \lambda \le 0. \tag{6.7}$$

For every $n \in \mathbb{N}$, the vector $v_n = (v_n^{(0)}, \ldots, v_n^{(N_n-1)}) \in \mathbb{R}^{N_n}$ with

$$v_n^{(k)} = \sqrt{\frac{2}{N_n}} \cos\left[\frac{2\pi k t_n}{N_n}\right] \quad \text{for} \quad k = 0, \ldots, N_n - 1 \tag{6.8}$$

is an eigenvector of $\bar{A}_{N_n}$ corresponding to the eigenvalue $\bar{\lambda}_{t_n}^{(N_n)}$ (cf. (6.3)). We define a sequence of step functions $(\varphi_n)_{n\in\mathbb{N}} \subset L^2(0,2)$ by

$$\varphi_n = \sum_{k=0}^{N_n-1} \cos\left[\frac{2\pi k t_n}{N_n}\right] \chi_{\left(\frac{2}{N_n}k, \frac{2}{N_n}(k+1)\right)} = \sqrt{\frac{N_n}{2}} \sum_{k=0}^{N_n-1} v_n^{(k)} \chi_{\left(\frac{2}{N_n}k, \frac{2}{N_n}(k+1)\right)}.$$

Obviously, $|\varphi_n| \le 1$, so that $\|\varphi_n\|_{L^2(0,2)} \le \sqrt{2}$ for every $n \in \mathbb{N}$; hence we can choose a subsequence (not relabelled) weakly converging to a function $\varphi \in L^2(0,2)$, i.e.,

$$\varphi_n \underset{n\to\infty}{\rightharpoonup} \varphi \quad \text{in} \quad L^2(0,2). \tag{6.9}$$

Furthermore, since $\frac{t_n}{N_n} \notin \mathbb{Z}$, we can apply Euler's formula to obtain

$$\int_0^2 \varphi_n(x) dx = \sum_{k=0}^{N_n-1} \int_{\frac{2}{N_n}k}^{\frac{2}{N_n}(k+1)} \varphi_n(x) dx = \frac{2}{N_n} \sum_{k=0}^{N_n-1} \cos\left[\frac{2\pi k t_n}{N_n}\right] = 0 \tag{6.10}$$

for every $n \in \mathbb{N}$. Now, take an arbitrary $\psi \in C_0^\infty(0,2)$ and set

$$\psi_n = \sum_{k=0}^{N_n-1} \psi\left(\frac{2}{N_n}k\right) \chi_{\left(\frac{2}{N_n}k, \frac{2}{N_n}(k+1)\right)}$$

for every $n \in \mathbb{N}$. Then

$$\psi_n \underset{n\to\infty}{\longrightarrow} \psi \text{ in } L^2(0,2). \tag{6.11}$$

For every $n \in \mathbb{N}$, we define $w_n = (w_n^{(0)}, \dots, w_n^{(N_n-1)}) \in \mathbb{R}^{N_n}$ by

$$w_n^{(k)} = \sqrt{\frac{2}{N_n}}\,\psi\left(\frac{2}{N_n}k\right)$$

for $k = 0, \dots, N_n - 1$, and together with (6.8) we get

$$w_n^t v_n = \int_0^2 \psi_n(x)\varphi_n(x)dx, \tag{6.12}$$

and if φ_n and ψ_n also denote the periodic extensions of φ_n, ψ_n on $\mathbb{R}$, a careful calculation that makes use of $v_n^{(0)} = \sqrt{\frac{2}{N_n}}$ and $w_n^{(0)} = 0$ yields

$$\sum_{i=1}^{N_n-1} w_n^{(i-1)}v_n^{(i)} + w_n^{(N_n-1)}v_n^{(0)} = \int_0^2 \psi_n(x)\varphi_n\left(x + \frac{2}{N_n}\right)dx,$$
$$\sum_{i=1}^{N_n-1} w_n^{(i)}v_n^{(i-1)} + w_n^{(0)}v_n^{(N_n-1)} = \int_0^2 \psi_n(x)\varphi_n\left(x - \frac{2}{N_n}\right)dx.$$

Recalling (6.1) and the definitions in Theorem 4.3, a further computation leads to

$$\begin{aligned} w_n^t \bar{A}_{N_n} v_n &= \int_{-1}^1 g(y)dy \int_0^2 \psi_n(x)\left[\varphi_n(x) + \frac{1}{2}\varphi_n\left(x + \frac{2}{N_n}\right) + \frac{1}{2}\varphi_n\left(x - \frac{2}{N_n}\right)\right]dx \\ &\quad - 2\int_0^2\int_0^2 g(x-y)\psi_n(x)\varphi_n(y)dydx \\ &=: \mathcal{B}_n(\psi_n, \varphi_n). \end{aligned}$$

Since v_n is an eigenvector of $\bar{A}_{N_n}$ corresponding to the eigenvalue $\bar{\lambda}_{t_n}^{(N_n)}$, we derive from (6.12)

$$\bar{\lambda}_{t_n}^{(N_n)} \int_0^2 \psi_n(x)\varphi_n(x)dx = \bar{\lambda}_{t_n}^{(N_n)} w_n^t v_n = w_n^t \bar{A}_{N_n} v_n = \mathcal{B}_n(\psi_n, \varphi_n). \tag{6.13}$$

By straightforward methods, we obtain, using (6.9), (6.11), and the boundedness of g,

$$\lim_{n\to\infty} \mathcal{B}_n(\psi_n, \varphi_n) = \mathcal{B}(\psi, \varphi), \tag{6.14}$$

where $\mathcal{B} : L^2(0,2) \times L^2(0,2) \to \mathbb{R}$ is the symmetric, continuous bilinear form given by (6.6). Furthermore, we deduce from (6.9), (6.11), and $\bar{\lambda}_{t_n}^{(N_n)} \to \lambda$ that

$$\lim_{n\to\infty} \bar{\lambda}_{t_n}^{(N_n)} \int_0^2 \psi_n(x)\varphi_n(x)dx = \lambda(\psi, \varphi)_{L^2(0,2)},$$

which by (6.14) and (6.13) implies

$$\mathcal{B}(\psi, \varphi) = \lambda(\psi, \varphi)_{L^2(0,2)}. \tag{6.15}$$

Since $\psi \in C_0^\infty(0,2)$ was chosen arbitrarily, (6.15) holds for every $\psi \in C_0^\infty(0,2)$. Since $C_0^\infty(0,2)$ is dense in $L^2(0,2)$ and because of the $L^2(0,2)$-continuity of both sides, (6.15) holds for every $\psi \in L^2(0,2)$. Choosing $\psi = \varphi$ we obtain, using the 2-periodicity of g on $\mathbb{R}$ and the symmetry of g,

$$\begin{aligned}\lambda\|\varphi\|^2_{L^2(0,2)} = \mathcal{B}(\varphi,\varphi) &= 2\int_{-1}^1 g(y)dy \int_0^2 \varphi(x)^2 dx - 2\int_0^2\int_0^2 g(x-y)\varphi(x)\varphi(y)dydx \\ &= \int_0^2\int_0^2 g(x-y)(\varphi(x)-\varphi(y))^2 dydx.\end{aligned}$$

Equations (6.10) and (6.9) immediately imply $\int_0^2 \varphi(x)dx = 0$, so that due to (6.7) and the positivity of g (i.e., $g \geq g_0 > 0$ on $\mathbb{R}$; see the conditions on g in the introduction), we get

$$0 \geq \lambda\|\varphi\|^2_{L^2(0,2)} \geq g_0 \int_0^2\int_0^2 (\varphi(x)-\varphi(y))^2 dydx = 4g_0\|\varphi\|^2_{L^2(0,2)},$$

which, since $g_0 > 0$, can only hold if $\varphi = 0$, and due to (6.9), this implies

$$\varphi_n \xrightarrow[n\to\infty]{} 0 \quad \text{in} \quad L^2(0,2). \tag{6.16}$$

Recalling (6.2) and the definitions in Theorem 4.3, we obtain

$$\begin{aligned}\bar{\lambda}_{t_n}^{(N_n)} &= b_0^{(N_n)} + 2b_1^{(N_n)}\cos\left[\frac{2\pi t_n}{N_n}\right] + c_0^{(N_n)} + (-1)^{t_n} c_{\frac{N_n}{2}}^{(N_n)} + 2\sum_{k=1}^{\frac{N_n}{2}-1} c_k^{(N_n)}\cos\left[\frac{2\pi k t_n}{N_n}\right] \\ &= b_0^{(N_n)}\left(1+\cos\left[\frac{2\pi t_n}{N_n}\right]\right) + \sum_{k=0}^{\frac{N_n}{2}-1} c_k^{(N_n)}\cos\left[\frac{2\pi k t_n}{N_n}\right] + \sum_{k=1}^{\frac{N_n}{2}} c_k^{(N_n)}\cos\left[\frac{2\pi t_n}{N_n}\right].\end{aligned}$$

Setting

$$\tilde{c}_k^{(N_n)} = -2\int_{\frac{2}{N_n}k}^{\frac{2}{N_n}(k+1)} g(y)dy \quad \text{for} \quad k = 0,\dots,\frac{N_n}{2},$$

one can show that $|\tilde{c}_k^{(N_n)} - c_k^{(N_n)}| \leq \frac{C}{N_n^2}$. Thus, if we define

$$\mu_{t_n}^{(N_n)} = b_0^{(N_n)}\left(1+\cos\left[\frac{2\pi t_n}{N_n}\right]\right) + \sum_{k=0}^{\frac{N_n}{2}-1} \tilde{c}_k^{(N_n)}\cos\left[\frac{2\pi k t_n}{N_n}\right] + \sum_{k=1}^{\frac{N_n}{2}} \tilde{c}_k^{(N_n)}\cos\left[\frac{2\pi k t_n}{N_n}\right],$$

we observe that $|\mu_{t_n}^{(N_n)} - \bar{\lambda}_{t_n}^{(N_n)}| \leq \frac{C}{N_n}$; hence by Lemma 6.2

$$\lambda_{t_n}^{(N_n)} \geq \bar{\lambda}_{t_n}^{(N_n)} - \frac{C}{N_n} \geq \mu_{t_n}^{(N_n)} - \frac{C}{N_n}.$$

A simple calculation yields

$$\begin{aligned}\sum_{k=0}^{\frac{N_n}{2}-1} \tilde{c}_k^{(N_n)}\cos\left[\frac{2\pi k t_n}{N_n}\right] &= -2\int_0^1 g(y)\varphi_n(y)dy, \\ \sum_{k=1}^{\frac{N_n}{2}} \tilde{c}_k^{(N_n)}\cos\left[\frac{2\pi k t_n}{N_n}\right] &= -2\int_{\frac{2}{N_n}}^{1+\frac{2}{N_n}} g(y)\varphi_n(y)dy = -2\int_0^1 g(y)\varphi_n(y)dy + O\left(\frac{1}{N_n}\right),\end{aligned}$$

so that

$$\lambda_{t_n}^{(N_n)} \geq \mu_{t_n}^{(N_n)} - \frac{C}{N_n} = \int_0^2 g(y)dy \left(1 + \cos\left[\frac{2\pi t_n}{N_n}\right]\right) - 4\int_0^1 g(y)\varphi_n(y)dy - \frac{C}{N_n},$$

and due to (6.16), taking the limit for $n \to \infty$ yields

$$\begin{aligned} 0 \geq \lambda &\geq \int_0^2 g(y)dy \cdot \limsup_{n\to\infty} \left(1 + \cos\left[\frac{2\pi t_n}{N_n}\right]\right) \\ &\geq \int_0^2 g(y)dy \cdot \liminf_{n\to\infty} \left(1 + \cos\left[\frac{2\pi t_n}{N_n}\right]\right) \geq 0, \end{aligned}$$

which implies $\lambda = 0$. Furthermore, since $\int_{-1}^1 g(y)dy > 0$, we deduce that

$$\lim_{n\to\infty} \left(1 + \cos\left[\frac{2\pi t_n}{N_n}\right]\right) = 0,$$

which due to $\frac{2\pi t_n}{N_n} \in [0, \pi]$ implies that $\lim_{n\to\infty} \frac{t_n}{N_n} = \frac{1}{2}$, and thus the proof is complete. ∎

From Theorem 6.1 we easily conclude the following property.

COROLLARY 6.5. *For every even number $N \in \mathbb{N}$, let $\lambda_t^{(N)}$, $t = 0, \ldots, \frac{N}{2}$, be the eigenvalues of A_N as given in* (5.1). *Then for every $\delta > 0$, we find $\nu_0 > 0$ and $N_0 \in \mathbb{N}$ such that for every even number $N \geq N_0$, the following holds true:*

$$\lambda_t^{(N)} \geq \nu_0 \quad \textit{for every} \quad t \in \left\{1, \ldots, \frac{N}{2} - 1\right\} \quad \textit{satisfying} \quad \frac{t}{N} \leq \frac{1}{2} - \delta. \tag{6.17}$$

7. Estimation of the eigenvalues. In order to show the positivity of the eigenvalues $\lambda_t^{(N)}$ for $t = 1, \ldots, \frac{N}{2} - 1$ and N large, due to Corollary 6.5 we only have to consider the indices t with $\frac{t}{N} \geq \frac{1}{2} - \delta$, where δ is sufficiently small. We can do so by introducing the parameter

$$\delta = \delta_{t,N} = 1 - \frac{2t}{N} \tag{7.1}$$

and then estimating $\lambda_t^{(N)}$ in terms of $\frac{1}{N}$ and δ. Recall the eigenvalues of A_N given by (5.1) with coefficients as given in Theorem 4.3. While in the proof of the "index property" (Theorem 6.1) we could omit the $d_i^{(N)}$-terms (which was done by replacing A_N with $\bar{A}_N$), since they disappear in the limit $N \to \infty$ when $\frac{t}{N}$ is not close to $\frac{1}{2}$, they play an essential role in the case considered in this section. Unfortunately, since the calculations are extremely long and very technical, we can only mention the results here and have to refer the reader to [24] for details. First, a calculation yields

$$\begin{aligned} \lambda_t^{(N)} = &\int_{-1}^1 g(y)dy \left(1 + \cos\left[\frac{2\pi t}{N}\right]\right) + \left(c_0^{(N)} + (-1)^t c_{\frac{N}{2}}^{(N)} + 2\sum_{k=1}^{\frac{N}{2}-1} c_k^{(N)} \cos\left[\frac{2\pi k t}{N}\right]\right) \\ &+ 2N \sum_{k=0}^{\frac{N}{2}-1} (-1)^k \int_0^{\frac{1}{N}} x \left[g\left(\frac{2k+1}{N} - x\right) - g\left(\frac{2k+1}{N} + x\right)\right] dx \left(1 - \cos\left[\frac{2\pi t}{N}\right]\right). \end{aligned} \tag{7.2}$$

The middle part can be written in the form

$$c_0^{(N)} + (-1)^t c_{\frac{N}{2}}^{(N)} + 2\sum_{k=1}^{\frac{N}{2}-1} c_k^{(N)} \cos\left[\frac{2\pi kt}{N}\right] = T_{t,N}^{(1)} + T_{t,N}^{(2)},$$

where

$$T_{t,N}^{(1)} = \sum_{k=0}^{\frac{N}{2}-1} c_k^{(N)} \left(\cos\left[\frac{2\pi kt}{N}\right] + \cos\left[\frac{2\pi(k+1)t}{N}\right]\right), \tag{7.3}$$

$$T_{t,N}^{(2)} = \sum_{k=0}^{\frac{N}{2}-1} (c_{k+1}^{(N)} - c_k^{(N)}) \cos\left[\frac{2\pi(k+1)t}{N}\right]. \tag{7.4}$$

Then (cf. [24, Lemma 10.2 and remarks at the end of Chapter 10]) we have the following.

LEMMA 7.1. *Let $N \in \mathbb{N}$ be an even number, $t \in \{1, \ldots, \frac{N}{2} - 1\}$, δ as given by* (7.1). *Then*

$$|T_{t,N}^{(1)}| \le C\left(\delta^3 + \frac{\delta^2}{N} + \frac{\delta}{N^2}\right).$$

Idea of the proof. The first step is to observe that

$$\begin{aligned}\cos\left[\frac{2\pi kt}{N}\right] + \cos\left[\frac{2\pi(k+1)t}{N}\right] &= (-1)^{k+1}(\cos((k+1)\pi\delta) - \cos(k\pi\delta)) \\ &= (-1)^k \pi\delta \sin(k\pi\delta) + \frac{1}{2}(-1)^k \pi^2\delta^2 \cos(\xi_{k,N})\end{aligned}$$

and to show that

$$\begin{aligned}T_{t,N}^{(1)} &\approx \pi\delta \sum_{k=0}^{\frac{N}{2}-1} (-1)^k c_k^{(N)} \sin(k\pi\delta) = \pi\delta \sum_{\substack{k=0 \\ k \text{ even}}}^{\frac{N}{2}-1} (c_k^{(N)} \sin(k\pi\delta) - c_{k+1}^{(N)} \sin((k+1)\pi\delta)) \\ &= \pi\delta \sum_{\substack{k=0 \\ k \text{ even}}}^{\frac{N}{2}-1} (c_k^{(N)} - c_{k+1}^{(N)}) \sin(k\pi\delta) + \pi\delta \sum_{\substack{k=0 \\ k \text{ even}}}^{\frac{N}{2}-1} c_{k+1}^{(N)} (\sin(k\pi\delta) - \sin((k+1)\pi\delta)).\end{aligned}$$

As for the left-hand term, we have

$$\begin{aligned}c_k^{(N)} - c_{k+1}^{(N)} &= N\int_0^{\frac{2}{N}} \int_{\frac{2}{N}k}^{\frac{2}{N}(k+1)} \left[g\left(x - y + \frac{2}{N}\right) - g(x-y)\right] dxdy \\ &\approx 2\int_0^{\frac{2}{N}} \int_{\frac{2}{N}k}^{\frac{2}{N}(k+1)} g'(x-y)dxdy \\ &= 2\int_0^{\frac{2}{N}} \left[g\left(\frac{2}{N}(k+1) - y\right) - g\left(\frac{2}{N}k - y\right)\right] dy \\ &\approx \frac{4}{N}\int_0^{\frac{2}{N}} g'\left(\frac{2}{N}k - y\right) dy,\end{aligned} \tag{7.5}$$

from which one can conclude that

$$\pi\delta \sum_{\substack{k=2 \\ k\ even}}^{\frac{N}{2}-2} (c_k^{(N)} - c_{k+1}^{(N)}) \sin(k\pi\delta) \approx 4\pi\frac{\delta}{N} \sum_{\substack{k=2 \\ k\ even}}^{\frac{N}{2}-2} \sin(k\pi\delta) \int_{\frac{2}{N}(k-1)}^{\frac{2}{N}k} g'(y)dy,$$

$$\approx 4\pi\frac{\delta}{N} \sum_{\substack{k=2 \\ k\ even}}^{\frac{N}{2}-2} \int_{\frac{2}{N}(k-1)}^{\frac{2}{N}k} g'(y)\sin\left(\frac{N}{2}\pi\delta y\right) dy \approx 2\pi\frac{\delta}{N}\int_0^1 g'(y)\sin\left(\frac{N}{2}\pi\delta y\right) dy.$$

With similar techniques, the right-hand term in (7.5) can be approximated by a Fourier integral for g as

$$\pi\delta \sum_{\substack{k=0 \\ k\ even}}^{\frac{N}{2}-1} c_{k+1}^{(N)}(\sin(k\pi\delta) - \sin((k+1)\pi\delta)) \approx \pi^2\delta^2 \int_0^1 g(x)\cos\left(\frac{N}{2}\pi\delta x\right) dx,$$

and by partial integration both integrals sum up to zero.

Although the estimation of $T_{t,N}^{(2)}$ is done similarly, one has to be even more careful. So far, one could "delete" terms where a Taylor representation is not possible due to nondifferentiability of g in 0. In the above sketch, this was done by leaving out the index $k = 0$ in the sum over even indices. This is not possible here, and one obtains border terms.

LEMMA 7.2. *Let $N \in \mathbb{N}$ be even, δ as given by* (7.1)*. Then*

$$T_{t,N}^{(2)} = \frac{4}{3N^2}g'(0) + O\left(\frac{\delta^2}{N}\right) + O\left(\frac{\delta}{N^2}\right) + O\left(\frac{1}{N^3}\right).$$

As for the rear part of (7.2), one can show the following.

LEMMA 7.3. *Let $N \in \mathbb{N}$ be a multiple of* 4 *(i.e., $\frac{N}{2}$ is even), $t \in \{1, \dots, \frac{N}{2} - 1\}$, δ as given by* (7.1)*. Then*

$$2N \sum_{k=0}^{\frac{N}{2}-1} (-1)^k \int_0^{\frac{1}{N}} x \left[g\left(\frac{2k+1}{N} - x\right) - g\left(\frac{2k+1}{N} + x\right)\right] dx \left(1 - \cos(\frac{2\pi t}{N})\right)$$
$$= -\frac{4}{3N^2}g'(0) + O\left(\frac{\delta}{N^2}\right) + O\left(\frac{1}{N^3}\right).$$

Thus, (7.2) yields

$$\lambda_t^{(N)} = \int_{-1}^1 g(y)dy\left(1 + \cos\left[\frac{2\pi t}{N}\right]\right) + O\left(\delta^3\right) + O\left(\frac{\delta^2}{N}\right) + O\left(\frac{\delta}{N^2}\right) + O\left(\frac{1}{N^3}\right),$$

and using $\delta \geq \frac{2}{N}$ (cf. (7.1)), one obtains

$$\int_{-1}^1 g(y)dy\left(1 + \cos\left[\frac{2\pi t}{N}\right]\right) \geq C\delta^2 \geq C\left(\delta^2 + \frac{1}{N^2}\right),$$

so that the first term is the dominant one, which implies $\lambda_t^{(N)} > 0$ for $t = 1, \dots, \frac{N}{2} - 1$ if N is large and δ is small, the latter being equivalent to $\frac{t}{N} \geq \frac{1}{2} - \delta$. Thus, if δ is small enough, application of Corollary 6.5 yields positivity of all eigenvalues $\lambda_t^{(N)}$, $t = 1, \dots, \frac{N}{2} - 1$, if N is large.

Acknowledgments. I would like to thank Prof. Dr. Hansjörg Kielhöfer for constant support and good advice during the years I had been working on my thesis. My gratitude goes to Prof. Dr. Stefan Müller for inviting me to the Max-Planck-Institute for Mathematics in the Sciences twice, for pointing me to this subject, and for several stimulating discussions. I want to thank Prof. Dr. Timothy Healey for inviting me to the Department of Theoretical Applied Mechanics at Cornell University, where I was able to present first results and to exchange ideas. I am also indebted to Prof. Dr. Stanislaus Maier-Paape for inviting me to the RWTH Aachen and for inspiring discussions.

REFERENCES

[1] G. Alberti, *Variational models for phase transitions, an approach via* Γ*-convergence*, in L. Ambrosio and N. Dancer: Calculus of Variations and Partial Differential Equations. Topics on Geometrical Evolution Problems and Degree Theory, G. Buttazzo et al., eds., Springer-Verlag, Berlin, 2000, pp. 95–114.

[2] G. Alberti and G. Bellettini, *A nonlocal anisotropic model for phase transitions: Asymptotic behavior of rescaled energies*, European J. Appl. Math., 9 (1998), pp. 261–284.

[3] G. Alberti and G. Bellettini, *A nonlocal anisotropic model for phase transitions*, Math. Ann., 310 (1998), pp. 527–560.

[4] G. Alberti and S. Müller, *A new approach to variational problems with multiple scales*, Comm. Pure Appl. Math., 54 (2001), pp. 761–825.

[5] J. M. Ball and R. D. James, *Fine phase mixtures as minimizers of energy*, Arch. Rational Mech. Anal., 100 (1987), pp. 13–52.

[6] J. M. Ball and R. D. James, *Proposed experimental tests of a theory of fine microstructure and the two-well problem*, Philos. Trans. Roy. Soc. London, Seri. A, 338 (1992), pp. 389–450.

[7] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Applied Math. Sci., 78, Springer-Verlag, Berlin, 1989.

[8] D. Dal Maso, *An Introduction to* Γ*-Convergence*, Prog. in Nonlinear Differential Equations Appl. 8, Birkhäuser, Basel, 1993.

[9] P. Davis, *Circulant Matrices*, Pure and applied Mathematics, John Wiley and Sons, New York, 1979.

[10] E. De Giorgi, *Convergence problems for functionals and operators*, in Proceedings of the International Meeting on Recent Methods in Nonlinear Analysis, Pitagora, Bologna, 1979, pp. 131–188.

[11] R. M. Gray, *Toeplitz and Circulant Matrices: A Review*, http://ee.stanford.edu/˜gray/toeplitz.pdf.

[12] R. V. Kohn and S. Müller, *Branching of twins near an austenite/twinned-martensite interface*, Philosophical Magazine A, 66 (1992), pp. 697–715.

[13] R .V. Kohn and S. Müller, *Surface energy and microstructure in coherent phase transitions*, Comm. Pure Appl. Math., 47 (1994), pp. 405–435.

[14] R. V. Kohn and P. Sternberg, *Local minimisers and singular perturbations*, Proc. Roy. Soc. Edinburgh Sect. A, 111 (1989), pp. 69–84.

[15] L. Modica, *The gradient theory of phase transitions and the minimal interface criterion*, Arch. Rational Mech. Anal., 98 (1987), pp. 123–142.

[16] L. Modica and S. Mortola, *Il limite nella* Γ*-convergenza di una famiglia di funzionali ellittici.*, Boll. Un. Mat. Ital. A, 14 (1977), pp. 526–529.

[17] L. Modica and S. Mortola, *Un esempio di* Γ^-*-convergenza*, Boll. Un. Mat. Ital. B, 14 (1977), pp. 285–299.

[18] S. Müller, *Singular perturbations as a selection criterion for periodic minimizing sequences*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 169–204.

[19] Y. Nishiura and I. Ohnishi, *Some mathematical aspects of micro-phase separation in diblock copolymers*, Phys. D, 84 (1995), pp. 31–39.

[20] T. Ohta and K. Kawasaki, *Equilibrium morphology of block copolymer melts*, Macromolecules, 19 (1986), pp. 2621–2632.

[21] X. Ren and J. Wei, *On the multiplicity of solutions of two nonlocal variational problems*, SIAM J. Math. Anal., 31 (2000), pp. 909–924.

[22] X. Ren and J. Wei, *Concentrically layered energy equilibria of the di-block copolymer problem*, European J. Appl. Math., 13 (2002), pp. 479–496.
[23] X. Ren and J. Wei, *On energy minimizers of the di-block copolymer problem*, Interfaces Free Bound., 5 (2003), pp. 193–238.
[24] N. Tzoukmanis, *Local Minimizers of Singularly Perturbed Functionals with Nonlocal Term*, Augsburger Schriften zur Mathematik, Physik und Informatik Vol. 7, Logos Verlag, Berlin, 2004.
[25] E. Zeidler, *Nonlinear Functional Analysis and its Applications.* III: *Variational Methods and Optimization*, Springer-Verlag, New York, 1985.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1417–1434

A BLOW-UP CRITERION FOR THE NONHOMOGENEOUS INCOMPRESSIBLE NAVIER–STOKES EQUATIONS*

HYUNSEOK KIM†

Abstract. Let (ρ, u) be a strong or smooth solution of the nonhomogeneous incompressible Navier–Stokes equations in $(0, T^*) \times \Omega$, where T^* is a finite positive time and Ω is a bounded domain in $\mathbf{R}^3$ with smooth boundary or the whole space $\mathbf{R}^3$. We show that if (ρ, u) blows up at T^*, then $\int_0^{T^*} |u(t)|^s_{L^r_w(\Omega)}\, dt = \infty$ for any (r, s) with $\frac{2}{s} + \frac{3}{r} = 1$ and $3 < r \leq \infty$. As immediate applications, we obtain a regularity theorem and a global existence theorem for strong solutions.

Key words. blow-up criterion, nonhomogeneous incompressible Navier–Stokes equations

AMS subject classifications. 35Q30, 76D05

DOI. 10.1137/S0036141004442197

1. Introduction.

The motion of a nonhomogeneous incompressible viscous fluid in a domain Ω of $\mathbf{R}^3$ is governed by the *Navier–Stokes equations*

$$(\rho u)_t + \operatorname{div}(\rho u \otimes u) - \Delta u + \nabla p = \rho f, \tag{1.1}$$

$$\rho_t + \operatorname{div}(\rho u) = 0 \quad \text{in} \quad (0, \infty) \times \Omega, \tag{1.2}$$

$$\operatorname{div} u = 0 \tag{1.3}$$

and the initial and boundary conditions

$$\begin{aligned} &(\rho,\, \rho u)|_{t=0} = (\rho_0,\, \rho_0 u_0) \quad \text{in} \quad \Omega, \qquad u = 0 \quad \text{on} \quad (0, T) \times \partial\Omega, \\ &\rho(t, x) \to 0, \quad u(t, x) \to 0 \quad \text{as} \quad |x| \to \infty, \quad (t, x) \in (0, T) \times \Omega. \end{aligned} \tag{1.4}$$

Here we denote by ρ, u, and p the unknown density, velocity, and pressure fields of the fluid, respectively. f is a given external force driving the motion. Ω is either a bounded domain in $\mathbf{R}^3$ with smooth boundary or the whole space $\mathbf{R}^3$. Throughout this paper, we adopt the following simplified notation for standard homogeneous and inhomogeneous Sobolev spaces:

$$\begin{aligned} &L^r = L^r(\Omega), \quad D^{k,r} = \{v \in L^1_{loc}(\Omega) : |v|_{D^{k,r}} < \infty\}; \\ &H^{k,r} = L^r \cap D^{k,r}, \quad D^k = D^{k,2}, \quad H^k = H^{k,2}; \\ &D^1_0 = \{v \in L^6 : |v|_{D^1_0} < \infty \ \text{and} \ v = 0 \ \text{on} \ \partial\Omega\}; \\ &H^1_0 = L^2 \cap D^1_{0,}, \quad D^1_{0,\sigma} = \{v \in D^1_0 : \operatorname{div} v = 0 \text{ in } \Omega\}; \\ &|v|_{D^{k,r}} = |\nabla^k v|_{L^r} \quad \text{and} \quad |v|_{D^1_0} = |v|_{D^1_{0,\sigma}} = |\nabla v|_{L^2}. \end{aligned}$$

Note that the space D^1_0 is the completion of $C^\infty_c(\Omega)$ in D^1, and thus there holds the following Sobolev inequality:

$$|v|_{L^6} \leq \frac{2}{\sqrt{3}} |v|_{D^1_0} \qquad \text{for all } v \in D^1_0. \tag{1.5}$$

*Received by the editors March 17, 2004; accepted for publication (in revised form) May 9, 2005; published electronically January 10, 2006. This work was supported by the Japan Society for the Promotion of Science under the JSPS Postdoctoral Fellowship for Foreign Researchers.
http://www.siam.org/journals/sima/37-5/44219.html
†School of Mathematics, Korea Institute for Advanced Study 207-43 Cheongnyangni 2-dong, Dongdaemun-gu, Seoul, 130-722, Korea (khs319@postech.ac.kr, khs319@kias.re.kr).

For a proof of (1.5), see sections II.5 and II.6 in the book by Galdi [11].

The global existence of weak solutions has been established by Antontsev and Kazhikov [1], Fernandez-Cara and Guillen [10], Kazhikov [13], Lions [21], and Simon [26, 27]. From these results (see [21] in particular), it follows that for any data (ρ_0, u_0, f) with the regularity

$$0 \le \rho_0 \in L^{\frac{3}{2}} \cap L^\infty, \quad u_0 \in L^6, \quad \text{and} \quad f \in L^2(0,\infty; L^2),$$

there exists at least one weak solution (ρ, u) to the initial boundary value problem (1.1)–(1.4) satisfying the regularity

$$\rho \in L^\infty(0,\infty; L^{\frac{3}{2}} \cap L^\infty), \quad \sqrt{\rho} u \in L^\infty(0,\infty; L^2), \quad \text{and} \quad u \in L^2(0,\infty; D^1_{0,\sigma}) \tag{1.6}$$

as well as the natural energy inequality. Then an associated pressure p is determined as a distribution in $(0,\infty) \times \Omega$.

But the global existence of strong or smooth solutions is still an open problem and only local existence results have been obtained for sufficiently regular data satisfying some compatibility conditions. For details, we refer to the papers by Choe and the author [6], Kim [15], Ladyzhenskaya and Solonnikov [19], Okamoto [22], Padula [23], and Salvi [24]. In particular, it is shown in [6] (see also [7]) that if the data ρ_0, u_0, and f satisfy

$$\begin{gathered} 0 \le \rho_0 \in L^{\frac{3}{2}} \cap H^2, \quad u_0 \in D^1_{0,\sigma} \cap D^2, \quad -\Delta u_0 + \nabla p_0 = \rho_0^{\frac{1}{2}} g, \\ f \in L^2(0,\infty; H^1), \quad \text{and} \quad f_t \in L^2(0,\infty; L^2) \end{gathered} \tag{1.7}$$

for some $(p_0, g) \in D^1 \times L^2$, then there exist a positive time T and a unique strong solution (ρ, u) to the problem (1.1)–(1.4) such that

$$\begin{gathered} \rho \in C([0,T]; L^{\frac{3}{2}} \cap H^2), \quad u \in C([0,T]; D^1_{0,\sigma} \cap D^2) \cap L^2(0,T; D^3), \\ u_t \in L^2(0,T; D^1_{0,\sigma}), \quad \text{and} \quad \sqrt{\rho} u_t \in L^\infty(0,T; L^2). \end{gathered} \tag{1.8}$$

Moreover, the existence of a pressure p in $C([0,T]; D^1) \cap L^2(0,T; D^2)$ can be deduced from (1.1)–(1.3). See [5] for a detailed argument.

Let (ρ, u) be a global weak solution to the problem (1.1)–(1.4) with the data (ρ_0, u_0, f) satisfying condition (1.7). Then from the above local existence result and weak-strong uniqueness results in [6] and [21], we conclude that the solution (ρ, u) satisfies the regularity (1.8) for some positive time T. One fundamental problem in mathematical fluid mechanics is to determine whether or not (ρ, u) satisfies (1.8) for all time T. As an equivalent formulation, we may ask the following.

Fundamental question 1.1. Does the solution (ρ, u) blow up at some finite time T^*? Such a time T^*, if it exists, is called the finite *blow-up time* of the solution (ρ, u) in the class H^2.

In spite of great efforts since the pioneering works by Leray [20] in 1930s, there have been no definite answers to the fundamental question even for the case of the *homogenous Navier–Stokes equations* with only some blow-up criteria available. The first criterion is due to Leray [20] who proved, among other things, that if T^* is the finite blow-up time of a strong solution u to the Cauchy problem for the homogeneous Navier–Stokes equations, then for each r with $3 < r \le \infty$, there exists a constant $C = C(r) > 0$ such that

$$|u(t)|_{L^r} \ge C(T^* - t)^{-\frac{1}{2}(1-\frac{3}{r})} \qquad \text{for all near } t < T^*. \tag{1.9}$$

This estimate near the blow-up time was extended by Giga [12] to the case of bounded domains. An immediate consequence of (1.9) is the following well-known blow-up criterion in terms of the so-called Serrin class (see [9, 25, 29]):

$$\int_0^{T^*} |u(t)|_{L^r}^s \, dt = \infty \quad \text{for any } (r,s) \text{ with } \frac{2}{s} + \frac{3}{r} = 1, \qquad 3 < r \le \infty. \tag{1.10}$$

By virtue of Sobolev inequality (1.5), we deduce from (1.10) that

$$\int_0^{T^*} |\nabla u(t)|_{L^2}^4 \, dt = \infty. \tag{1.11}$$

Further generalizations of (1.10) and (1.11) have been obtained by Beirao da Veiga [2], Berselli [3], Chae and Choe [4], and Kozono and Taniuchi [17].

The major purpose of this paper is to prove the blow-up criterion (1.10) for strong solutions of the nonhomogeneous Navier–Stokes equations (1.1)–(1.3). In fact, we establish a more general result. To state our main result precisely, we first introduce the notion of the blow-up time of solutions in the class H^{2m} with $m \ge 1$. Let (ρ, u) be a strong solution to the initial boundary value problem (1.1)–(1.4) with the regularity

$$\begin{aligned} &\rho \in C([0,T]; L^{\frac{3}{2}} \cap H^{2m}), \quad u \in C([0,T]; D^1_{0,\sigma} \cap D^{2m}) \cap L^2(0,T; D^{2m+1}), \\ &\partial_t^j u \in C([0,T]; D^1_{0,\sigma} \cap D^{2m-2j}) \cap L^2(0,T; D^{2m-2j+1}) \quad \text{for} \quad 1 \le j < m, \\ &\quad \partial_t^m u \in L^2(0,T; D^1_{0,\sigma}), \qquad \text{and} \qquad \sqrt{\rho}\, \partial_t^m u \in L^\infty(0,T; L^2) \end{aligned} \tag{1.12}$$

for any $T < T^*$, where T^* is a finite positive time. Then we can define

$$\begin{aligned} \Phi_m(T) = {} & 1 + \sup_{0 \le t \le T} \left(|\nabla \rho(t)|_{H^{2m-1}} + |u(t)|_{D_0^1 \cap D^{2m}} \right) + \int_0^T |u(t)|_{D^{2m+1}}^2 \, dt \\ & + \sup_{1 \le j < m} \left(\sup_{0 \le t \le T} |\partial_t^j u(t)|_{D_0^1 \cap D^{2m-2j}} + \int_0^T |\partial_t^j u(t)|_{D^{2m-2j+1}}^2 \, dt \right) \\ & + \operatorname*{ess\,sup}_{0 \le t \le T} |\sqrt{\rho}\, \partial_t^m u(t)|_{L^2} + \int_0^T |\partial_t^m u(t)|_{D_0^1}^2 \, dt \end{aligned} \tag{1.13}$$

for any $T < T^*$. Hereafter we use the obvious notation

$$|\cdot|_{X \cap Y} = |\cdot|_X + |\cdot|_Y \quad \text{for (semi-)normed spaces } X,\, Y.$$

Definition 1.2. A finite positive number T^* is called the finite *blow-up time* of the solution (ρ, u) in the class H^{2m}, provided that

$$\Phi_m(T) < \infty \quad \text{for} \quad 0 < T < T^* \quad \text{and} \quad \lim_{T \to T^*} \Phi_m(T) = \infty.$$

We are now ready to state the main result of this paper.

Theorem 1.3. *For a given integer $m \ge 1$, we assume that*

$$\partial_t^m f \in L^2(0,\infty; L^2) \quad \textit{and} \quad \partial_t^j f \in L^2(0,\infty; H^{2m-2j-1}) \quad \textit{for} \quad 0 \le j < m.$$

Let (ρ, u) be a strong solution of the nonhomogeneous Navier–Stokes equations (1.1)–(1.3) *satisfying the regularity* (1.12) *for any $T < T^*$. If T^* is the finite blow-up time of (ρ, u) in the class H^{2m}, then we have*

$$\int_0^{T^*} |u(t)|_{L_w^r}^s \, dt = \infty \quad \textit{for any } (r,s) \textit{ with } \frac{2}{s} + \frac{3}{r} = 1, \qquad 3 < r \le \infty. \tag{1.14}$$

Here L^r_w denotes the weak L^r-space, that is, the space consisting of all vector fields $v \in L^1_{loc}(\Omega)$ such that $|v|_{L^r_w} = \sup_{\alpha>0} \alpha \left|\{x \in \Omega : |v(x)| > \alpha\}\right|^{\frac{1}{r}} < \infty$ for $3 < r < \infty$ and $|v|_{L^\infty_w} = |v|_{L^\infty} < \infty$. In the case when $3 < r < \infty$, L^r is a proper subspace of L^r_w ($|x|^{-3/r} \in L^r_w(\mathbf{R}^3)$ for instance) and so Theorem 1.3 is in fact a generalization of the blow-up criterion (1.10) due to Leray and Giga even for the homogeneous Navier–Stokes equations.

Theorem 1.3 is proved in the next two sections. In section 2, we provide a proof of the theorem for the very special case $m = 1$. Our method of the proof is quite well known in the case of the homogeneous Navier–Stokes equations and was also applied in an earlier paper [6] by Choe and the author to the nonhomogeneous case: combining classical regularity results on the Stokes equations with Hölder and Sobolev inequalities, we show that $\Phi_1(T)$ is bounded in a double exponential way by $\int_0^T |u(t)|^s_{L^r_w}\, dt$ for any T less than the blow-up time T^*. But the use of weak Lebesgue spaces in space variables makes it more difficult to estimate the nonlinear convection term. To overcome this technical difficulty, we utilize some basic theory of the Lorenz spaces developed in [18] and [30]. See the derivations of (2.6) and (2.7) for details. Concerning the proof for the general case $m \geq 2$, the basic idea is also to show that $\Phi_m(T)$ is bounded in some specific way by $\int_0^T |u(t)|^s_{L^r_w}\, dt$ for any $T < T^*$. Such an approach is more or less standard in the case of the homogeneous Navier–Stokes equations, but its extension to the nonhomogeneous case is not straightforward and indeed much complicated due to the evolution of the density. A detailed argument is provided in section 3.

Some corollaries of Theorem 1.3 can be deduced from a local existence result on strong solutions in the class H^{2m}. For instance, as an immediate consequence of Theorem 1.3, the local existence result in the class H^2, and the weak-strong uniqueness result, we obtain the following regularity result whose obvious proof may be omitted.

Corollary 1.4. *Let* (ρ, u) *be a global weak solution to the initial boundary value problem* (1.1)–(1.4) *with the data satisfying condition* (1.7). *If there exists a finite positive time* T_* *such that*

$$u \in L^s(0, T_*; L^r_w) \quad \text{for some } (r, s) \text{ with } \frac{2}{s} + \frac{3}{r} = 1, \qquad 3 < r \leq \infty, \tag{1.15}$$

then the solution (ρ, u) *satisfies regularity* (1.8) *for some* $T > T_*$.

A similar result was obtained by Choe and the author [6] assuming, however, a stronger condition on u, that is, $u \in L^4(0, T_*; D^1_0)$. By virtue of Corollary 1.4, we may conclude that the class (which we call a *weak Serrin class*) in (1.15) is a regularity class for weak solutions of the nonhomogeneous Navier–Stokes equations, which was already proved by Sohr [28] for the homogeneous case. See also a local version of Sohr's result in [14]. Moreover, thanks to a recent result by Dubois [8], the weak Serrin class is a uniqueness class for the homogeneous Navier–Stokes equations. It is also noticed that the same results can be easily derived from regularity and uniqueness results due to Kozono by adapting the arguments in the remarks of Theorem 3 in [16].

Theorem 1.3 and its proof can be used to obtain a global existence result on solutions in the class H^2 under some smallness condition on u_0 and f (but not on ρ_0).

Theorem 1.5. *For each* $K > 1$*, there exists a small constant* $\varepsilon > 0$*, depending only on* K *and the domain* Ω*, with the following property: if the data* ρ_0*,* u_0*, and* f *satisfy*

$$|\rho_0|_{L^{\frac{3}{2}} \cap L^\infty} \leq K, \quad |u_0|_{D^1_0} \leq \varepsilon, \quad \text{and} \quad \int_0^\infty |f(t)|^2_{L^2}\, dt \leq \varepsilon^2 \tag{1.16}$$

in addition to condition (1.7)*, then there exists a unique global strong solution* (ρ, u) *to problem* (1.1)–(1.4) *satisfying regularity* (1.8) *for any* $T < \infty$*.*

A rather simple proof of Theorem 1.5 is provided in section 4. Finally, we recall that in the case when ρ_0 is bounded away from zero and Ω is a bounded domain in $\mathbf{R}^3$ with smooth boundary, Salvi [24] proved the local existence of strong solutions in the class H^{2m} for every $m \geq 1$. Hence adapting the proofs of Corollary 1.4 and Theorem 1.5, we can also prove analogous regularity and global existence results on strong solutions in every class H^{2m} provided that $\Omega \subset\subset \mathbf{R}^3$ and $\rho_0 > 0$ on $\overline{\Omega}$.

2. Proof of Theorem 1.3 with $\;\;= 1$. In this section, we prove Theorem 1.3 for the special case $m = 1$. Let t_0 be a fixed time with $0 < t_0 < T^*$ and let us denote

$$\Phi_0(T) = \int_0^T |u(t)|_{L^r_w}^s \, dt \quad \text{for} \quad t_0 \leq T < T^*,$$

where (r, s) is any pair satisfying condition (1.14). To prove Theorem 1.3, we have only to show that

$$\Phi_1(T) \leq C \exp\left(C \exp\left(C\Phi_0(T)\right)\right) \quad \text{for} \quad t_0 \leq T < T^*. \tag{2.1}$$

Throughout this paper, we denote by C a generic positive constant depending only on r, m, $\Phi_m(t_0)$, T^*, Ω, $|\rho(0)|_{L^{\frac{3}{2}} \cap L^\infty}$, $|u(0)|_{L^6}$, and the norm of f, but independent of T in particular. To begin with, we recall from (1.6) that

$$\sup_{0 \leq t \leq T} \left(|\rho(t)|_{L^{\frac{3}{2}} \cap L^\infty} + |\sqrt{\rho} u(t)|_{L^2} \right) + \int_0^T |u(t)|_{D_0^1}^2 \, dt \leq C \tag{2.2}$$

for $t_0 \leq T < T^*$.

2.1. Estimates for $\int_0^T |\sqrt{\;\;}\;\;_t(\;\;)|_{L^2}^2$ and $\sup_{0 \leq t \leq T} |\;\;(\;\;)|_{D_0^1}$. Next, we will show that

$$\int_0^T \left(|\sqrt{\rho} u_t(t)|_{L^2}^2 + |u(t)|_{D_0^1 \cap D^2}^2 \right) dt + \sup_{0 \leq t \leq T} |u(t)|_{D_0^1}^2 \leq C \exp\left(C\Phi_0(T)\right) \tag{2.3}$$

for $t_0 \leq T < T^*$. To show this, we multiply the momentum equation (1.1) by u_t and integrate over Ω. Then using (1.2) and (1.3), we easily derive

$$\int \rho |u_t|^2 \, dx + \frac{1}{2} \frac{d}{dt} \int |\nabla u|^2 \, dx = \int \rho \left(f - u \cdot \nabla u \right) \cdot u_t \, dx$$

and

$$\int \rho |u_t|^2 \, dx + \frac{d}{dt} \int |\nabla u|^2 \, dx \leq 2 \int \rho |f|^2 \, dx + 2 \int \rho |u \cdot \nabla u|^2 \, dx. \tag{2.4}$$

On the other hand, since (u, p) is a solution of the stationary Stokes equations

$$-\Delta u + \nabla p = F \quad \text{and} \quad \operatorname{div} u = 0 \quad \text{in} \quad \Omega,$$

where $F = \rho(f - u \cdot \nabla u - u_t)$, it follows from the classical regularity theory that

$$\begin{aligned} |\nabla u|_{H^1} &\leq C \left(|F|_{L^2} + |\nabla u|_{L^2} \right) \\ &\leq C \left(|f|_{L^2} + |\sqrt{\rho} u_t|_{L^2} + |u \cdot \nabla u|_{L^2} + |\nabla u|_{L^2} \right). \end{aligned} \tag{2.5}$$

To estimate the right-hand sides of (2.4) and (2.5), we first observe that

$$|u\cdot\nabla u|_{L^2} = |u\cdot\nabla u|_{L^{2,2}} \le C|u|_{L^r_w}|\nabla u|_{L^{\frac{2r}{r-2},2}}, \tag{2.6}$$

which follows from Hölder inequality in the Lorenz spaces. See Proposition 2.1 in [18]. Next, we will show that

$$|\nabla u|_{L^{\frac{2r}{r-2},2}} \le C|\nabla u|_{L^2}^{1-\frac{3}{r}}|\nabla u|_{H^1}^{\frac{3}{r}}. \tag{2.7}$$

If $r=\infty$, then (2.7) is obvious. Assuming that $3<r<\infty$, we choose r_1 and r_2 such that $3<r_1<r<r_2<\infty$ and $\frac{2}{r}=\frac{1}{r_1}+\frac{1}{r_2}$. Then in view of Hölder and Sobolev inequalities, we have

$$|\nabla u|_{L^{\frac{2r_i}{r_i-2}}} \le |\nabla u|_{L^2}^{1-\frac{3}{r_i}}|\nabla u|_{L^6}^{\frac{3}{r_i}} \le |\nabla u|_{L^2}^{1-\frac{3}{r_i}}\left(C|\nabla u|_{H^1}\right)^{\frac{3}{r_i}}$$

for each $i=1,2$. Since $L^{\frac{2r}{r-2},2}$ is a real interpolation space of $L^{\frac{2r_2}{r_2-2}}$ and $L^{\frac{2r_1}{r_1-2}}$, more precisely, $L^{\frac{2r}{r-2},2}=(L^{\frac{2r_2}{r_2-2}},L^{\frac{2r_1}{r_1-2}})_{\frac12,2}$, it thus follows that

$$\begin{aligned}|\nabla u|_{L^{\frac{2r}{r-2},2}} &\le C|\nabla u|^{\frac12}_{L^{\frac{2r_2}{r_2-2}}}|\nabla u|^{\frac12}_{L^{\frac{2r_1}{r_1-2}}}\\ &\le C\left(|\nabla u|_{L^2}^{1-\frac{3}{r_2}}\left(C|\nabla u|_{H^1}\right)^{\frac{3}{r_2}}\right)^{\frac12}\left(|\nabla u|_{L^2}^{1-\frac{3}{r_1}}\left(C|\nabla u|_{H^1}\right)^{\frac{3}{r_1}}\right)^{\frac12},\end{aligned}$$

which proves (2.7). For some facts on the real interpolation theory and Lorenz spaces used above, we refer to sections 1.3.3 and 1.18.6 in Triebel's book [30].

The estimates (2.6) and (2.7) yield

$$|u\cdot\nabla u|_{L^2} \le C|u|_{L^r_w}|\nabla u|_{L^2}^{\frac{2}{s}}|\nabla u|_{H^1}^{\frac{3}{r}} \le \eta^{-\frac{3s}{2r}}C|u|_{L^r_w}^{\frac{s}{2}}|\nabla u|_{L^2}+\eta|\nabla u|_{H^1}$$

for any small number $\eta\in(0,1)$. Substituting this into (2.5), we obtain

$$|\nabla u|_{H^1} \le C\left(|f|_{L^2}+|\sqrt{\rho}u_t|_{L^2}+|u|_{L^r_w}^{\frac{s}{2}}|\nabla u|_{L^2}+|\nabla u|_{L^2}\right), \tag{2.8}$$

and thus

$$|u\cdot\nabla u|_{L^2} \le \eta^{-\frac{3s}{2r}}C\left(|u|_{L^r_w}^{\frac{s}{2}}+1\right)|\nabla u|_{L^2}+C|f|_{L^2}+\eta|\sqrt{\rho}u_t|_{L^2}.$$

Therefore, substituting this estimate into (2.4) and choosing a sufficiently small $\eta>0$, we conclude that

$$\frac12|\sqrt{\rho}u_t(t)|_{L^2}^2+\frac{d}{dt}|\nabla u(t)|_{L^2}^2 \le C|f(t)|_{L^2}^2+C\left(|u(t)|_{L^r_w}^s+1\right)|\nabla u(t)|_{L^2}^2 \tag{2.9}$$

for $t_0\le t<T^*$. In view of Gronwall's inequality, we have

$$\int_0^T|\sqrt{\rho}u_t(t)|_{L^2}^2\,dt+\sup_{0\le t\le T}|\nabla u(t)|_{L^2}^2 \le C\exp\left(C\Phi_0(T)\right)$$

for any T with $t_0\le T<T^*$. Combining this and (2.8), we obtain the desired estimate (2.3).

2.2. Estimates for $\operatorname{ess\,sup}_{0\le t\le T}|\sqrt{\rho}u_t(\cdot)|^2_{L^2}$ and $\int_0^T|\nabla u_t(\cdot)|^2_{D_0^1}$. To derive these estimates, we differentiate the momentum equation (1.1) with respect to time t and obtain

$$\rho u_{tt}+\rho u\cdot\nabla u_t-\Delta u_t+\nabla p_t=\rho_t(f-u_t-u\cdot\nabla u)+\rho(f_t-u_t\cdot\nabla u).$$

Then multiplying this by u_t, integrating over Ω, and using (1.2) and (1.3), we have

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int\rho|u_t|^2\,dx+\int|\nabla u_t|^2\,dx\\&\qquad=\int(\rho_t(f-u_t-u\cdot\nabla u)+\rho(f_t-u_t\cdot\nabla u))\cdot u_t\,dx.\end{aligned}\tag{2.10}$$

Note that since $\rho\in C([0,T];L^{\frac{3}{2}}\cap L^\infty)$, $\rho_t\in C([0,T];L^{\frac{3}{2}})$, and $u_t\in L^2(0,T;D_0^1)$ for any $T<T^*$, the right-hand side of (2.10) is well defined for almost all $t\in(0,T^*)$. Hence using finite differences in time, we can easily show that the identity (2.10) holds for almost all $t\in(0,T^*)$.

In view of the continuity equation (1.2) again, we deduce from (2.10) that

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int\rho|u_t|^2\,dx+\int|\nabla u_t|^2\,dx\\&\quad\le\int 2\rho|u||u_t||\nabla u_t|+\rho|u||u_t||\nabla u|^2+\rho|u|^2|u_t||\nabla^2u|\\&\qquad+\rho|u|^2|\nabla u||\nabla u_t|+\rho|u_t|^2|\nabla u|+\rho|u||u_t||\nabla f|\\&\qquad+\rho|u||f||\nabla u_t|+\rho|f_t||u_t|\ dx\equiv\sum_{j=1}^{8}I_j.\end{aligned}\tag{2.11}$$

Following the arguments in [6], we can estimate each term I_j:

$$\begin{aligned}I_1,\,I_5&\le C|\rho|^{\frac{1}{2}}_{L^\infty}|\nabla u|_{L^2}|\sqrt{\rho}u_t|_{L^3}|\nabla u_t|_{L^2}\le C|\rho|^{\frac{3}{4}}_{L^\infty}|\nabla u|_{L^2}|\sqrt{\rho}u_t|^{\frac{1}{2}}_{L^2}|\nabla u_t|^{\frac{3}{2}}_{L^2}\\&\le C|\nabla u|^4_{L^2}|\sqrt{\rho}u_t|^2_{L^2}+\frac{1}{16}|\nabla u_t|^2_{L^2},\end{aligned}$$

$$I_2,\,I_3,\,I_4\le C|\rho|_{L^\infty}|\nabla u|^2_{L^2}|\nabla u_t|_{L^2}|\nabla u|_{H^1}\le C|\nabla u|^4_{L^2}|\nabla u|^2_{H^1}+\frac{1}{16}|\nabla u_t|^2_{L^2},$$

$$I_6,\,I_7\le C|\rho|_{L^6}|\nabla u|_{L^2}|f|_{H^1}|\nabla u_t|_{L^2}\le C|\nabla u|^2_{L^2}|f|^2_{H^1}+\frac{1}{16}|\nabla u_t|^2_{L^2},$$

and finally

$$I_8\le C|\rho|_{L^3}|f_t|_{L^2}|\nabla u_t|_{L^2}\le C|f_t|^2_{L^2}+\frac{1}{16}|\nabla u_t|^2_{L^2}.$$

Substitution of these estimates into (2.11) yields

$$\begin{aligned}&\frac{d}{dt}|\sqrt{\rho}u_t|^2_{L^2}+|\nabla u_t|^2_{L^2}\\&\qquad\le C|\nabla u|^4_{L^2}\left(|\sqrt{\rho}u_t|^2_{L^2}+|\nabla u|^2_{H^1}+|f|^2_{H^1}\right)+C\left(|f|^2_{H^1}+|f_t|^2_{L^2}\right).\end{aligned}$$

Therefore, by virtue of estimate (2.3), we conclude that

$$\operatorname*{ess\,sup}_{0\le t\le T}|\sqrt{\rho}u_t(t)|_{L^2}^2+\int_0^T|\nabla u_t(t)|_{L^2}^2\,dt\le C\exp\left(C\Phi_0(T)\right) \tag{2.12}$$

for $t_0\le T<T^*$. On the other hand, using the regularity theory of the Stokes equations again, we have

$$\begin{aligned}|\nabla u|_{H^1}&\le C\left(|f|_{L^2}+|\sqrt{\rho}u_t|_{L^2}+|u\cdot\nabla u|_{L^2}+|\nabla u|_{L^2}\right)\\&\le C\left(|f|_{L^2}+|\sqrt{\rho}u_t|_{L^2}+|\nabla u|_{L^2}^{\frac{3}{2}}|\nabla u|_{H^1}^{\frac{1}{2}}+|\nabla u|_{L^2}\right)\end{aligned}$$

and

$$\begin{aligned}|\nabla u|_{H^{1,6}}&\le C\left(|u_t|_{L^6}+|u\cdot\nabla u|_{L^6}+|f|_{L^6}+|\nabla u|_{L^6}\right)\\&\le C\left(|\nabla u_t|_{L^2}+|\nabla u|_{H^1}^2+|f|_{H^1}+|\nabla u|_{H^1}\right).\end{aligned}$$

Hence it follows immediately from (2.3) and (2.12) that

$$\sup_{0\le t\le T}|u(t)|_{D_0^1\cap D^2}^2+\int_0^T|\nabla u(t)|_{H^{1,6}}^2\,dt\le C\exp\left(C\Phi_0(T)\right) \tag{2.13}$$

for $t_0\le T<T^*$.

2.3. Estimates for $\sup_{0\le t\le T}|\nabla\ (\ )|_{H^1}$ and $\int_0^T|\ (\ )|_{D^3}^2$. To derive these, we first observe that each ρ_{x_j} $(j=1,2,3)$ satisfies

$$\left(\rho_{x_j}\right)_t+u\cdot\nabla\rho_{x_j}=-u_{x_j}\cdot\nabla\rho.$$

Then multiplying this by ρ_{x_j}, integrating over Ω, and summing up, we obtain

$$\frac{d}{dt}\int|\nabla\rho|^2\,dx\le C\int|\nabla u||\nabla\rho|^2\,dx\le C|\nabla u|_{L^\infty}|\nabla\rho|_{L^2}^2.$$

A similar argument shows that

$$\begin{aligned}\frac{d}{dt}\int|\nabla^2\rho|^2\,dx&\le C\int\left(|\nabla u||\nabla^2\rho|^2+|\nabla^2u||\nabla\rho||\nabla^2\rho|\right)dx\\&\le C|\nabla u|_{L^\infty}|\nabla^2\rho|_{L^2}^2+|\nabla^2u|_{L^6}|\nabla\rho|_{L^3}|\nabla^2\rho|_{L^2}.\end{aligned}$$

Hence using Sobolev embedding results and then Gronwall's inequality, we derive the well-known estimate

$$\begin{aligned}|\nabla\rho(t)|_{H^1}&\le C\exp\left(C\int_0^t|\nabla u(\tau)|_{H^{1,6}}\,d\tau\right)\\&\le C\exp\left(\int_0^t\left(C|\nabla u(\tau)|_{H^{1,6}}^2+1\right)\,d\tau\right).\end{aligned}$$

Therefore by virtue of (2.13), we conclude that

$$\sup_{0\le t\le T}|\nabla\rho(t)|_{H^1}\le C\exp\left(C\exp\left(C\Phi_0(T)\right)\right) \tag{2.14}$$

for $t_0 \leq T < T^*$. Finally, observing from the regularity theory on the Stokes equations that

$$\begin{aligned} |u|_{D^3} &\leq C\left(|\rho u_t|_{H^1} + |\rho u \cdot \nabla u|_{H^1} + |\rho f|_{H^1}\right) \\ &\leq C\left(|\nabla\rho|_{L^3} + 1\right)\left(|\nabla u_t|_{L^2} + |\nabla u|_{H^1}^2 + |f|_{H^1}\right), \end{aligned}$$

we easily deduce from (2.3), (2.12), (2.13), and (2.14) that

$$\int_0^T |u(t)|_{D^3}^2\,dt \leq C\exp\left(C\exp\left(C\Phi_0(T)\right)\right) \tag{2.15}$$

for $t_0 \leq T < T^*$. This completes the proof of (2.1) and thus the proof of Theorem 1.3 with $m = 1$.

3. Proof of Theorem 1.3 with $\geq$ 2. Assume that $m \geq 2$. Then to prove Theorem 1.3, it suffices to show that the following estimate holds for each k, $0 \leq k < m$:

$$\Phi_{k+1}(T) \leq C\exp\left(C\exp\left(C\Phi_k(T)^{10m}\right)\right) \quad \text{for} \quad t_0 \leq T < T^*. \tag{3.1}$$

The case $k = 0$ was already proved in section 2 and so it remains to prove (3.1) for the case $1 \leq k < m$.

Let k be a fixed integer with $1 \leq k < m$. From (1.13), we recall that

$$\begin{aligned} \Phi_k(T) &= 1 + \sup_{0\leq j<k}\left(\sup_{0\leq t\leq T}|\partial_t^j u(t)|_{D_0^1\cap D^{2k-2j}} + \int_0^T |\partial_t^j u(t)|_{D^{2k-2j+1}}^2\,dt\right) \\ &\quad + \sup_{0\leq t\leq T}|\nabla\rho(t)|_{H^{2k-1}} + \operatorname{ess}\sup_{0\leq t\leq T}|\sqrt{\rho}\,\partial_t^k u(t)|_{L^2} + \int_0^T |\partial_t^k u(t)|_{D_0^1}^2\,dt \end{aligned} \tag{3.2}$$

for any $T < T^*$.

3.1. Estimates for $^{j}_{t}($ $\cdot\,\nabla$ $)$, $^{j+1}_{t}$, and $^{j}_{t}($ $)$ with $0 \leq$ $\leq$. To estimate nonlinear terms, we will make repeated use of the following simple lemma whose proof is omitted.

LEMMA 3.1. *If $g \in D_0^1 \cap D^j$, $h \in H^i$, $0 \leq i \leq j$, and $j \geq 2$, then*

$$gh \in H^i \qquad \text{and} \qquad |gh|_{H^i} \leq C|g|_{D_0^1\cap D^j}|h|_{H^i}$$

for some constant $C > 0$ depending only on j and Ω.

Using this lemma together with the fact that

$$\partial_t^j(u\cdot\nabla u) = \sum_{i=0}^{j}\frac{j!}{i!(j-i)!}\partial_t^i u\cdot\nabla\partial_t^{j-i}u,$$

we can estimate $\partial_t^j(u\cdot\nabla u)$ as follows: for $0 \leq j < k$,

$$\begin{aligned} |\partial_t^j(u\cdot\nabla u)|_{H^{2k-2j-1}} &\leq C\sum_{i=0}^{j}|\partial_t^i u\cdot\nabla\partial_t^{j-i}u|_{H^{2k-2j-1}} \\ &\leq C\sum_{i=0}^{j}|\partial_t^i u|_{D_0^1\cap D^{2k-2j}}|\nabla\partial_t^{j-i}u|_{H^{2k-2j-1}} \\ &\leq C\sum_{i=0}^{j}|\partial_t^i u|_{D_0^1\cap D^{2k-2i}}|\partial_t^{j-i}u|_{D_0^1\cap D^{2k-2(j-i)}} \end{aligned}$$

and

$$\begin{aligned}
|\partial_t^k(u\cdot\nabla u)|_{L^2} &\le C\sum_{i=0}^{k-1}|\partial_t^i u\cdot\nabla\partial_t^{k-i}u|_{L^2}+|\partial_t^k u\cdot\nabla u|_{L^2}\\
&\le C\sum_{i=0}^{k-1}|\partial_t^i u|_{D_0^1\cap D^2}|\nabla\partial_t^{k-i}u|_{L^2}+|\partial_t^k u|_{D_0^1}|\nabla u|_{H^1}\\
&\le C\sum_{i=0}^{k-1}|\partial_t^i u|_{D_0^1\cap D^{2k-2i}}|\partial_t^{k-i}u|_{D_0^1}+|\partial_t^k u|_{D_0^1}|u|_{D_0^1\cap D^2}.
\end{aligned}$$

Hence it follows from (3.2) that

$$\sup_{0\le j<k}\sup_{0\le t\le T}|\partial_t^j(u\cdot\nabla u)(t)|_{H^{2k-2j-1}}+\int_0^T|\partial_t^k(u\cdot\nabla u)(t)|_{L^2}^2\,dt\le C\Phi_k(T)^4 \tag{3.3}$$

for $t_0\le T<T^*$. Applying Lemma 3.1 to the continuity equation

$$\rho_t=-\mathrm{div}(\rho u)=-u\cdot\nabla\rho, \tag{3.4}$$

we also deduce that

$$\sup_{0\le t\le T}|\rho_t(t)|_{H^{2k-1}}\le C\Phi_k(T)^2\quad\text{for}\quad t_0\le T<T^*. \tag{3.5}$$

Using (3.4) and (3.5), we can show that

$$\sup_{1\le j<k}\sup_{0\le t\le T}|\partial_t^{j+1}\rho(t)|_{H^{2k-2j}}+\int_0^T|\partial_t^{k+1}\rho(t)|_{L^2}^2\,dt\le C\Phi_k(T)^{2k+4} \tag{3.6}$$

for $t_0\le T<T^*$. A simple inductive proof of (3.6) may be based on the observation that for $1\le j<k$,

$$\begin{aligned}
|\partial_t^{j+1}\rho|_{H^{2k-2j}} &= |-\partial_t^j(u\cdot\nabla\rho)|_{H^{2k-2j}}\\
&\le C\sum_{i=0}^{j}|\partial_t^{j-i}u\cdot\nabla\partial_t^i\rho|_{H^{2k-2j}}\\
&\le C\sum_{i=0}^{j}|\partial_t^{j-i}u|_{D_0^1\cap D^{2k-2j}}|\nabla\partial_t^i\rho|_{H^{2k-2j}}\\
&\le C\sum_{i=0}^{j}|\partial_t^{j-i}u|_{D_0^1\cap D^{2k-2(j-i)}}|\partial_t^i\rho|_{H^{2k-2(i-1)}}
\end{aligned}$$

and

$$|\partial_t^{k+1}\rho|_{L^2}\le C\sum_{i=0}^{k}|\partial_t^{k-i}u\cdot\nabla\partial_t^i\rho|_{L^2}\le C\sum_{i=0}^{k}|\partial_t^{k-i}u|_{D_0^1}|\partial_t^i\rho|_{H^2}.$$

Moreover, it follows easily from (3.6) that

$$\sup_{0\le j<k}\sup_{0\le t\le T}|\partial_t^j(\rho u)(t)|_{H^{2k-2j}}+\int_0^T|\partial_t^k(\rho u)(t)|_{H^1}^2\,dt\le C\Phi_k(T)^{4k+10} \tag{3.7}$$

for $t_0 \le T < T^*$. Finally, recalling that

$$\partial_t^{k+1} f \in L^2(0,\infty; L^2) \quad \text{and} \quad \partial_t^j f \in L^2(0,\infty; H^{2k-2j+1}) \quad \text{for} \quad 0 \le j \le k,$$

we deduce from standard embedding results that

$$\partial_t^j f \in C([0,\infty); H^{2k-2j}) \quad \text{for} \quad 0 \le j \le k.$$

3.2. Estimates for $\int_0^T |\sqrt{\ }\ {}^{k+1}_{t}\ (\ )|^2_{L^2}$ and $\sup_{0\le t\le T} |\ {}^{k}_{t}\ (\ )|_{D_0^1}$. From the momentum equation (1.1), we derive

$$\rho \left(\partial_t^k u\right)_t - \Delta \partial_t^k u + \nabla \partial_t^k p = \partial_t^k(\rho f - \rho u \cdot \nabla u) + \left(\rho \partial_t^k u_t - \partial_t^k(\rho u_t)\right).$$

Hence multiplying this by $\partial_t^{k+1} u$ and integrating over Ω, we have

$$\begin{aligned} &\int \rho |\partial_t^{k+1} u|^2 \, dx + \frac{1}{2} \frac{d}{dt} \int |\nabla \partial_t^k u|^2 \, dx \\ &\quad = \int \left(\partial_t^k(\rho f - \rho u \cdot \nabla u) + \left(\rho \partial_t^k u_t - \partial_t^k(\rho u_t)\right)\right) \cdot \partial_t^{k+1} u \, dx \\ &\quad = I_{0,1} + \sum_{j=1}^{k} \frac{k!}{j!(k-j)!} \left(I_{j,1} + I_{j,2}\right), \end{aligned} \tag{3.8}$$

where

$$I_{j,1} = \int \partial_t^j \rho \, \partial_t^{k-j}(f - u \cdot \nabla u) \cdot \partial_t^{k+1} u \, dx, \quad I_{j,2} = -\int \partial_t^j \rho \, \partial_t^{k-j} u_t \cdot \partial_t^{k+1} u \, dx.$$

We easily estimate $I_{0,1}$ as follows:

$$\begin{aligned} I_{0,1} &\le |\rho|_{L^\infty}^{\frac{1}{2}} \left(|\partial_t^k f|_{L^2} + |\partial_t^k (u \cdot \nabla u)|_{L^2}\right) |\sqrt{\rho} \partial_t^{k+1} u|_{L^2} \\ &\le C \left(|\partial_t^k f|_{L^2}^2 + |\partial_t^k (u \cdot \nabla u)|_{L^2}^2\right) + \frac{1}{2} |\sqrt{\rho} \partial_t^{k+1} u|_{L^2}^2. \end{aligned}$$

To estimate $I_{j,1}$ for $1 \le j \le k$, we rewrite it as

$$\begin{aligned} I_{j,1} = &\frac{d}{dt} \int \partial_t^j \rho \, \partial_t^{k-j}(f - u \cdot \nabla u) \cdot \partial_t^k u \, dx - \int \partial_t^{j+1} \rho \, \partial_t^{k-j}(f - u \cdot \nabla u) \cdot \partial_t^k u \, dx \\ &- \int \partial_t^j \rho \, \partial_t^{k-j+1}(f - u \cdot \nabla u) \cdot \partial_t^k u \, dx \end{aligned}$$

and observe that

$$\begin{aligned} &- \int \partial_t^{j+1} \rho \, \partial_t^{k-j}(f - u \cdot \nabla u) \cdot \partial_t^k u \, dx \\ &\quad \le C \left(|\partial_t^{k-j} f|_{H^1}^2 + |\partial_t^{k-j}(u \cdot \nabla u)|_{H^1}^2\right) |\partial_t^{j+1} \rho|_{L^2}^2 + |\nabla \partial_t^k u|_{L^2}^2 \end{aligned}$$

and

$$\begin{aligned} &- \int \partial_t^j \rho \, \partial_t^{k-j+1}(f - u \cdot \nabla u) \cdot \partial_t^k u \, dx \\ &\quad \le C |\partial_t^j \rho|_{H^1}^2 \left(|\partial_t^{k-j+1} f|_{L^2}^2 + |\partial_t^{k-j+1}(u \cdot \nabla u)|_{L^2}^2\right) + |\nabla \partial_t^k u|_{L^2}^2. \end{aligned}$$

Using the continuity equation (1.2), we can also estimate $I_{j,2}$ as follows:

$$\begin{aligned} I_{1,2} &= -\int \rho_t \left(\frac{1}{2}|\partial_t^k u|^2\right)_t dx = -\frac{d}{dt}\int \rho_t \frac{1}{2}|\partial_t^k u|^2\,dx + \int \partial_t^2\rho \frac{1}{2}|\partial_t^k u|^2\,dx \\ &= -\frac{d}{dt}\int \rho u\cdot\nabla\left(\frac{1}{2}|\partial_t^k u|^2\right)dx + \int \partial_t(\rho u)\cdot\nabla\left(\frac{1}{2}|\partial_t^k u|^2\right)dx \\ &\leq -\frac{d}{dt}\int (\rho u\cdot\nabla\partial_t^k u)\cdot\partial_t^k u\,dx + C|\partial_t(\rho u)|_{H^1}|\nabla\partial_t^k u|_{L^2}^2 \end{aligned}$$

and similarly

$$\begin{aligned} I_{j,2} &= -\frac{d}{dt}\int \partial_t^j\rho\,\partial_t^{k-j}u_t\cdot\partial_t^k u\,dx + \int\left(\partial_t^{j+1}\rho\,\partial_t^{k-j}u_t + \partial_t^j\rho\,\partial_t^{k-j+1}u_t\right)\cdot\partial_t^k u\,dx \\ &\leq -\frac{d}{dt}\int \partial_t^{j-1}(\rho u)\cdot\nabla\left(\partial_t^{k-j+1}u\cdot\partial_t^k u\right)dx \\ &\quad + C\left(|\partial_t^j(\rho u)|_{H^1}|\nabla\partial_t^{k-j+1}u|_{L^2} + |\partial_t^{j-1}(\rho u)|_{H^1}|\nabla\partial_t^{k-j+2}u|_{L^2}\right)|\nabla\partial_t^k u|_{L^2} \end{aligned}$$

for $2\leq j\leq k$. Substituting all the estimates into (3.8), we have

$$\begin{aligned} &\frac{1}{2}\int \rho|\partial_t^{k+1}u|^2\,dx + \frac{1}{2}\frac{d}{dt}\int|\nabla\partial_t^k u|^2\,dx \\ &\quad\leq \frac{d}{dt}\int\left(\sum_{j=1}^k \frac{k!}{j!(k-j)!}\partial_t^j\rho\partial_t^{k-j}(f-u\cdot\nabla u)\cdot\partial_t^k u - (\rho u\cdot\nabla\partial_t^k u)\cdot\partial_t^k u\right)dx \\ &\quad\quad -\frac{d}{dt}\int\sum_{j=2}^k\frac{k!}{j!(k-j)!}\partial_t^{j-1}(\rho u)\cdot\nabla\left(\partial_t^{k-j+1}u\cdot\partial_t^k u\right)dx \\ &\quad\quad + C\sum_{j=1}^{k-1}\left(|\partial_t^j(\rho u)|_{H^1}^2 + |\partial_t^{j+1}\rho|_{H^1}^2\right)\left(|\partial_t^{k-j}f|_{H^1}^2 + |\partial_t^{k-j}(u\cdot\nabla u)|_{H^1}^2 + |\partial_t^{k-j+1}u|_{D_0^1}^2\right) \\ &\quad\quad + C|\partial_t^{k+1}\rho|_{L^2}^2\left(|f|_{H^1}^2 + |u|_{D_0^1\cap D^2}^2\right) + C\left(1+|\partial_t\rho|_{H^1}^2\right)\left(|\partial_t^k f|_{L^2}^2 + |\partial_t^k(u\cdot\nabla u)|_{L^2}^2\right) \\ &\quad\quad + C|\partial_t^k(\rho u)|_{H^1}^2 + C\left(1+|\partial_t(\rho u)|_{H^1}^2 + |\nabla\partial_t u|_{L^2}^2\right)|\nabla\partial_t^k u|_{L^2}^2. \end{aligned}$$

Hence, integrating this in time over (t_0,T) and using (3.3), (3.5), (3.6), and (3.7) together with the estimates

$$\begin{aligned} &\int |\partial_t^j\rho||\partial_t^{k-j}(f-u\cdot\nabla u)||\partial_t^k u|\,dx \\ &\quad\leq \eta^{-1}|\partial_t^j\rho|_{H^1}^2|\partial_t^{k-j}(f-u\cdot\nabla u)|_{L^2}^2 + \eta|\nabla\partial_t^k u|_{L^2}, \end{aligned}$$

$$\int\rho|u||\nabla\partial_t^k u||\partial_t^k u|\,dx \leq \eta^{-3}C|\rho|_{L^\infty}^3|\nabla u|_{L^2}^4|\sqrt{\rho}\partial_t^k u|_{L^2}^2 + \eta|\nabla\partial_t^k u|_{L^2}^2$$

and

$$\begin{aligned} &\int|\partial_t^{j-1}(\rho u)|\left(|\nabla\partial_t^{k-j+1}u||\partial_t^k u| + |\partial_t^{k-j+1}u||\nabla\partial_t^k u|\right)dx \\ &\quad\leq \eta^{-1}C|\partial_t^{j-1}(\rho u)|_{H^1}^2|\nabla\partial_t^{k-j+1}u|_{L^2}^2 + \eta|\nabla\partial_t^k u|_{L^2}, \end{aligned}$$

where η is any small positive number, we deduce that

$$\begin{aligned}&\int_{t_0}^{T} |\sqrt{\rho}\partial_t^{k+1}u(t)|_{L^2}^2\,dt + |\nabla\partial_t^k u(T)|_{L^2}^2\\ &\qquad\le C\Phi_k(T)^{20m} + C\int_{t_0}^{T}\left(1+|\partial_t(\rho u)(t)|_{H^1}^2 + |u_t(t)|_{D_0^1}^2\right)|\nabla\partial_t^k u(t)|_{L^2}^2\,dt\end{aligned}$$

for $t_0 \le T < T^*$. Note that

$$\int_{t_0}^{T}\left(1+|\partial_t(\rho u)(t)|_{H^1}^2 + |u_t(t)|_{D_0^1}^2\right)dt \le C\Phi_k(T)^{10m}.$$

Therefore, in view of Gronwall's inequality, we conclude that

$$\int_0^T |\sqrt{\rho}\partial_t^{k+1}u(t)|_{L^2}^2\,dt + \sup_{0\le t\le T}|\partial_t^k u(t)|_{D_0^1}^2 \le C\exp\left(C\Phi_k(T)^{10m}\right) \tag{3.9}$$

for any T with $t_0 \le T < T^*$.

3.3. Estimates for $\operatorname{ess\,sup}_{0\le t\le T}|\sqrt{\ }\ {}_t^{k+1}\ (\ )|_{L^2}$ and $\int_0^T|\ {}_t^{k+1}\ (\ )|_{D_0^1}^2$.
From the momentum equation (1.1), it follows that

$$\begin{aligned}&\rho\left(\partial_t^{k+1}u\right)_t + \rho u\cdot\nabla\partial_t^{k+1}u - \Delta\partial_t^{k+1}u + \nabla\partial_t^{k+1}p\\ &\quad= \partial_t^{k+1}(\rho f) + \left(\rho\partial_t^{k+1}u_t - \partial_t^{k+1}(\rho u_t)\right) + \left(\rho u\cdot\nabla\partial_t^{k+1}u - \partial_t^{k+1}(\rho u\cdot\nabla u)\right).\end{aligned}$$

Multiplying this by $\partial_t^{k+1}u$ and integrating over Ω, we have

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\int\rho|\partial_t^{k+1}u|^2\,dx + \int|\nabla\partial_t^{k+1}u|^2\,dx\\ &\quad= \int\partial_t^{k+1}(\rho f)\cdot\partial_t^{k+1}u\,dx + \int\left(\rho\partial_t^{k+1}u_t - \partial_t^{k+1}(\rho u_t)\right)\cdot\partial_t^{k+1}u\,dx\\ &\qquad+\int\left(\rho u\cdot\nabla\partial_t^{k+1}u - \partial_t^{k+1}(\rho u\cdot\nabla u)\right)\cdot\partial_t^{k+1}u\,dx.\end{aligned}\tag{3.10}$$

This identity can be derived rigorously by using a standard finite difference method because if $0 < T < T^*$, then $\partial_t^m\rho \in L^2(0,T;L^{\frac{3}{2}}\cap L^2)$ and $\partial_t^j\rho \in C([0,T];L^{\frac{3}{2}}\cap L^\infty)$ for $0 \le j < m$. The first term of the right-hand side in (3.10) is bounded by

$$\begin{aligned}&C\sum_{j=0}^{k}\int|\partial_t^j\rho||\partial_t^{k-j+1}f||\partial_t^{k+1}u|\,dx + \int|\partial_t^{k+1}\rho||f||\partial_t^{k+1}u|\,dx\\ &\qquad\le C\sum_{j=0}^{k}|\partial_t^j\rho|_{H^1}^2|\partial_t^{k-j+1}f|_{L^2}^2 + C|\partial_t^{k+1}\rho|_{L^2}^2|f|_{H^1}^2 + \frac{1}{6}|\nabla\partial_t^{k+1}u|_{L^2}^2.\end{aligned}$$

In view of the continuity equation (1.2), we can rewrite the second term as

$$\begin{aligned}&-\sum_{j=1}^{k+1}\frac{(k+1)!}{j!(k-j+1)!}\int\partial_t^j\rho\,\partial_t^{k-j+1}u_t\cdot\partial_t^{k+1}u\,dx\\ &\qquad= -\sum_{j=1}^{k+1}\frac{(k+1)!}{j!(k-j+1)!}\int\partial_t^{j-1}(\rho u)\cdot\nabla\left(\partial_t^{k-j+2}u\cdot\partial_t^{k+1}u\right)dx,\end{aligned}$$

which is bounded by

$$C|\rho|_{L^\infty}|u|^2_{D_0^1\cap D^2}|\sqrt{\rho}\partial_t^{k+1}u|^2_{L^2}+C\sum_{j=1}^{k}|\partial_t^j(\rho u)|^2_{H^1}|\partial_t^{k-j+1}u|^2_{D_0^1}+\frac{1}{6}|\nabla\partial_t^{k+1}u|^2_{L^2}.$$

Finally, the last term is bounded by

$$\begin{aligned}
C\sum_{j=1}^{k}\int|\partial_t^j(\rho u)||\nabla\partial_t^{k-j+1}u||\partial_t^{k+1}u|\,dx+\int|\partial_t^{k+1}(\rho u)||\nabla u||\partial_t^{k+1}u|\,dx&\\
\le C\sum_{j=1}^{k}\left(|\partial_t^j(\rho u)|^2_{H^1}+|\partial_t^j\rho|^2_{H^1}|u|^2_{D_0^1\cap D^2}\right)|\partial_t^{k-j+1}u|^2_{D_0^1}+C|\partial_t^{k+1}\rho|^2_{L^2}|u|^4_{D_0^1\cap D^2}&\\
+C|\rho|_{L^\infty}|u|^2_{D_0^1\cap D^2}|\sqrt{\rho}\partial_t^{k+1}u|^2_{L^2}+\frac{1}{6}|\nabla\partial_t^{k+1}u|^2_{L^2}.&
\end{aligned}$$

Hence substituting these estimates into (3.10), we have

$$\begin{aligned}
&\frac{d}{dt}\int\rho|\partial_t^{k+1}u|^2\,dx+\int|\nabla\partial_t^{k+1}u|^2\,dx\\
&\quad\le C\left(1+|u|^2_{D_0^1\cap D^2}\right)|\sqrt{\rho}\partial_t^{k+1}u|^2_{L^2}+C\sum_{j=0}^{k}|\partial_t^j\rho|^2_{H^1}|\partial_t^{k-j+1}f|^2_{L^2}+C|\partial_t^{k+1}\rho|^2_{L^2}|f|^2_{H^1}\\
&\qquad+C\sum_{j=1}^{k}\left(|\partial_t^j(\rho u)|^2_{H^1}+|\partial_t^j\rho|^2_{H^1}|u|^2_{D_0^1\cap D^2}\right)|\partial_t^{k-j+1}u|^2_{D_0^1}+C|\partial_t^{k+1}\rho|^2_{L^2}|u|^4_{D_0^1\cap D^2}.
\end{aligned}$$

Therefore, by virtue of (3.5), (3.6), (3.7), and (3.9), we conclude that

$$\text{(3.11)}\quad \operatorname*{ess\,sup}_{0\le t\le T}|\sqrt{\rho}\,\partial_t^{k+1}u(t)|_{L^2}+\int_0^T|\partial_t^{k+1}u(t)|^2_{D_0^1}\,dt\le C\exp\left(C\Phi_k(T)^{10m}\right)$$

for any T with $t_0\le T<T^*$.

3.4. Estimates for $\sup_{0\le t\le T}|\ {}^{j}_{t}\ (\ )|_{D_0^1\cap D^{2k-2j+2}}$ with $0\le\ \ \le\ $. To derive these estimates, we observe that

$$\text{(3.12)}\quad \partial_t^j u\in C([0,T^*);D^1_{0,\sigma})\quad\text{and}\quad -\Delta\partial_t^j u+\nabla\partial_t^j p=\partial_t^j(\rho f-\rho u\cdot\nabla u-\rho u_t)$$

for each $j\le k$. From (3.5), (3.6), (3.9), and (3.11), it follows easily that

$$\operatorname*{ess\,sup}_{0\le t\le T}\left(|\partial_t^k(\rho u\cdot\nabla u)(t)|_{L^2}+|\partial_t^k(\rho u_t)(t)|_{L^2}\right)\le C\exp\left(C\Phi_k(T)^{10m}\right)$$

for $t_0\le T<T^*$. Hence applying the regularity theory of the Stokes equations to (3.12) with $j=k$, we obtain

$$\sup_{0\le t\le T}|\partial_t^k u(t)|_{D_0^1\cap D^2}\le C\exp\left(C\Phi_k(T)^{10m}\right)\quad\text{for } t_0\le T<T^*.$$

It also follows from the Stokes regularity theory that for $0\le j<k$,

$$\text{(3.13)}\qquad |\partial_t^j u|_{D_0^1\cap D^{2k-2j+2}}\le C|\partial_t^j(\rho f-\rho u\cdot\nabla u-\rho u_t)|_{H^{2k-2j}}+C|\partial_t^j u|_{D_0^1}.$$

Using the estimates in section 3.1, we can estimate the first term of the right-hand side in (3.13) as follows:

$$\begin{aligned} &C|\partial_t^j(\rho f - \rho u\cdot\nabla u - \rho u_t)(t) + \rho\partial_t^j(u\cdot\nabla u + u_t)(t)|_{H^{2k-2j}} \\ &\quad\le C|\rho\partial_t^j f(t)|_{H^{2k-2j}} + C\sum_{i=1}^{j}|\partial_t^i\rho\partial_t^{j-i}(f - u\cdot\nabla u - u_t)(t)|_{H^{2k-2j}} \\ &\quad\le C\exp\left(C\Phi_k(T)^{10m}\right), \end{aligned}$$

$$\begin{aligned} &C|\rho\partial_t^j(u\cdot\nabla u)(t)|_{H^{2k-2j}} \\ &\quad\le C\sum_{i=1}^{j}|\rho\partial_t^i u\cdot\nabla\partial_t^{j-i}u(t)|_{H^{2k-2j}} + C|\rho u\cdot\nabla\partial_t^j u(t)|_{H^{2k-2j}} \\ &\quad\le C\exp\left(C\Phi_k(T)^{10m}\right) + C|\rho(t)|_{H^{2k}}|u(t)|_{D_0^1\cap D^{2k}}|\partial_t^j u(t)|_{D_0^1\cap D^{2k-2j+1}} \\ &\quad\le C\exp\left(C\Phi_k(T)^{10m}\right) + \frac{1}{2}|\partial_t^j u(t)|_{D_0^1\cap D^{2k-2j+2}} \end{aligned}$$

and

$$C|\rho\partial_t^j u_t(t)|_{H^{2k-2j}} \le C|\rho(t)|_{H^{2k}}|\partial_t^{j+1}u(t)|_{D_0^1\cap D^{2k-2j}}.$$

Substituting these into (3.13), we deduce that

$$|\partial_t^j u(t)|_{D_0^1\cap D^{2k-2j+2}} \le C\exp\left(C\Phi_k(T)^{10m}\right)\left(1 + |\partial_t^{j+1}u(t)|_{D_0^1\cap D^{2k-2j}}\right).$$

Therefore, by a backward induction on j, we conclude that

$$\sup_{0\le j<k+1}\sup_{0\le t\le T}|\partial_t^j u(t)|_{D_0^1\cap D^{2k-2j+2}} \le C\exp\left(C\Phi_k(T)^{10m}\right). \tag{3.14}$$

3.5. Estimates for $\sup_{0\le t\le T}|\nabla\ (\ )|_{H^{2k+1}}$ and $\int_0^T|\ {}_t^j\ (\ )|^2_{D^{2k-2j+3}}$ with $\le$. Let α be a multi-index with $1\le|\alpha|\le 2k+2$. Then taking the differential operator D^α to the continuity equation (1.2), we have

$$(D^\alpha\rho)_t + u\cdot\nabla(D^\alpha\rho) = u\cdot\nabla(D^\alpha\rho) - D^\alpha(u\cdot\nabla\rho).$$

Multiplying this by $D^\alpha\rho$ and integrating over Ω, we obtain

$$\frac{d}{dt}\int|D^\alpha\rho|^2\,dx \le C\int|u\cdot\nabla(D^\alpha\rho) - D^\alpha(u\cdot\nabla\rho)||D^\alpha\rho|\,dx. \tag{3.15}$$

But since

$$|u\cdot\nabla(D^\alpha\rho) - D^\alpha(u\cdot\nabla\rho)| \le C\sum_{l=1}^{|\alpha|}\left|\nabla^{|\alpha|+1-l}u\right|\left|\nabla^l\rho\right|,$$

Hölder and Sobolev inequalities yield

$$\begin{aligned} |u\cdot\nabla(D^\alpha\rho) - D^\alpha(u\cdot\nabla\rho)|_{L^2} &\le C\left(|\nabla u|_{H^{2k+1}}|\nabla\rho|_{H^{2k+1}} + |\nabla u|_{L^\infty}|\nabla\rho|_{H^{2k+1}}\right) \\ &\le C|\nabla u|_{H^{2k+1}}|\nabla\rho|_{H^{2k+1}}. \end{aligned}$$

Hence from (3.15), we derive the standard estimate

$$\frac{d}{dt}|\nabla\rho|^2_{H^{2k+1}} \le C|u|_{D_0^1\cap D^{2k+2}}|\nabla\rho|^2_{H^{2k+1}},$$

which implies then that

$$\sup_{0\le t\le T}|\nabla\rho(t)|_{H^{2k+1}} \le |\nabla\rho_0|_{H^{2k+1}}\exp\left(C\int_0^T |u(t)|_{D_0^1\cap D^{2k+2}}\,dt\right).$$

Therefore, by virtue of (3.14), we conclude that

$$\sup_{0\le t\le T}|\nabla\rho(t)|_{H^{2k+1}} \le C\exp\left(C\exp\left(C\Phi_k(T)^{10m}\right)\right)$$

for any T with $t_0 \le T < T^*$. Finally, applying the Stokes regularity theory to (3.12) for each j and arguing by induction on $k-j$, we easily show that

$$\sup_{0\le j<k+1}\int_0^T |\partial_t^j u(t)|^2_{D^{2k-2j+3}}\,dt \le C\exp\left(C\exp\left(C\Phi_k(T)^{10m}\right)\right)$$

for any T with $t_0 \le T < T^*$. This completes the proof of (3.1) and so we have completed the proof of Theorem 1.3.

4. Proof of Theorem 1.5. Let (ρ,u) be a strong solution satisfying the regularity (1.8) for some $T>0$. Assume that $|\rho_0|_{L^{\frac32}\cap L^\infty} \le K$ for some constant $K>1$. Then it follows easily from (1.2) and (1.3) that

$$|\rho(t)|_{L^{\frac32}\cap L^\infty} = |\rho_0|_{L^{\frac32}\cap L^\infty} \le K \quad \text{for} \quad 0\le t\le T.$$

Moreover, from the energy equality

$$\frac12\frac{d}{dt}\int\rho|u|^2\,dx + \int|\nabla u|^2\,dx = \int\rho f\cdot u\,dx,$$

we derive

$$|\sqrt{\rho}u(t)|^2_{L^2} + \int_0^t |\nabla u(\tau)|^2_{L^2}\,d\tau \le |\sqrt{\rho}_0u_0|^2_{L^2} + C\int_0^t |\rho(\tau)|^2_{L^3}|f(\tau)|^2_{L^2}\,d\tau,$$

and thus

$$\int_0^T |\nabla u(\tau)|^2\,d\tau \le C_K\left(|\nabla u_0|^2_{L^2} + \int_0^\infty |f(\tau)|^2_{L^2}\,d\tau\right). \tag{4.1}$$

Throughout the proof, we denote by $C_K>1$ a generic constant dependent only on K and Ω but independent of time T. On the other hand, from the estimate (2.9) with $(r,s)=(6,4)$, it follows that

$$|\nabla u(t)|^2_{L^2} \le |\nabla u_0|^2_{L^2} + C_K\int_0^t |f(\tau)|^2_{L^2}\,d\tau + C_K\int_0^t \left(|u(\tau)|^4_{L^6_w}+1\right)|\nabla u(\tau)|^2_{L^2}\,d\tau$$

for $0\le t\le T$. Hence by virtue of the estimate (4.1) and Sobolev inequality (1.5), we have

$$|\nabla u(t)|^2_{L^2} \le C_K\left(\sup_{0<\tau<t}|\nabla u(\tau)|^4_{L^2}+1\right)\left(|\nabla u_0|^2_{L^2} + \int_0^\infty |f(\tau)|^2_{L^2}\,d\tau\right) \tag{4.2}$$

for $0 < t \le T$. Let us choose a positive constant ε such that

$$0 < \varepsilon < 1 \qquad \text{and} \qquad 6C_K\varepsilon^2 < 1.$$

We now prove the global existence of a strong solution under the assumption that $|\nabla u_0|_{L^2} \le \varepsilon$ and $\int_0^\infty |f(\tau)|_{L^2}^2\, d\tau \le \varepsilon^2$. The local existence of a unique solution (ρ, u) was already proved in [6] (see also [7]). To prove the global existence, we argue by contradiction. Assume that (ρ, u) blows up at some finite time T^*, $0 < T^* < \infty$. Then since (ρ, u) satisfies the regularity (1.8) for any $T < T^*$, it follows from (4.2) that

$$|\nabla u(t)|_{L^2}^2 \le \frac{1}{3}\left(\sup_{0<\tau<t} |\nabla u(\tau)|_{L^2}^4 + 1\right) \quad \text{for} \quad 0 < t < T^*. \tag{4.3}$$

Note that $|\nabla u_0|_{L^2}^2 < 1$ and $u \in C([0,T^*); D_0^1)$. Hence from (4.3), we easily deduce that $|\nabla u(t)|_{L^2}^2 < 1$ for any $0 \le t < T^*$. Therefore, in view of Sobolev embedding again, we conclude that

$$\int_0^{T^*} |u(t)|_{L^6_w}^4\, dt \le C \int_0^{T^*} |\nabla u(t)|_{L^2}^4\, dt < \infty,$$

which contradicts Theorem 1.3. This completes the proof of Theorem 1.5.

Acknowledgments. The author would like to express his deep gratitude to Professor Hideo Kozono for very helpful discussion on the weak spaces. The author is also grateful to the referees for several valuable comments on the paper.

REFERENCES

[1] S. A. Antontsev and A. V. Kazhikov, *Mathematical Study of Flows of Nonhomogeneous Fluids; Lecture Notes*, Novosibirsk State University, Novosibirsk, U.S.S.R., 1973 (in Russian).

[2] H. Beirao da Veiga, *A new regularity class for the Navier-Stokes equations in R^n*, Chinese Ann. Math. Ser. B, 16 (1995), pp. 407–412.

[3] L. C. Berselli, *On a regularity criterion for the solutions to the 3D Navier-Stokes equations*, Differential Integral Equations, 15 (2002), pp. 1129–1137.

[4] D. Chae and H. J. Choe, *Regularity of solutions to the Navier-Stokes equations*, Electron. J. Differential Equations, (1999), no. 5.

[5] Y. Cho and H. Kim, *Unique solvability for the density-dependent Navier-Stokes equations*, Nonlinear Anal. Ser. A: Theory Methods, 59 (2004), pp. 465–489.

[6] H. J. Choe and H. Kim, *Strong solutions of the Navier-Stokes equations for nonhomogeneous incompressible fluids*, Comm. Partial Differential Equations, 28 (2003), pp. 1183–1201.

[7] Y. Cho, H. J. Choe, and H. Kim, *Unique solvability of the initial boundary value problems for compressible viscous fluids*, J. Math. Pures Appl., 83 (2004), pp. 243–275 .

[8] S. Dubois, *Uniqueness for some Leray-Hopf solutions to the Navier-Stokes equations*, J. Differential Equations, 189 (2003), pp. 99–147.

[9] E. B. Fabes, B. F. Jones, and N. M. Riviere, *The initial value problem for the Navier-Stokes equations with data in L^p*, Arch. Rational Mech. Anal., 45 (1972), pp. 222–240.

[10] E. Fernandez-Cara and F. Guillen, *The existence of nonhomogeneous, viscous and incompressible flow in unbounded domains*, Comm. Partial Differential Equations, 17 (1992), pp. 1253–1265.

[11] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations*, Vol. 1, Springer Tracts in Natural Philosophy 38, Springer-Verlag, New York, 1994.

[12] Y. Giga, *Solutions for semilinear parabolic equations in L^p and regularity of weak solutions of the Navier-Stokes system*, J. Differential Equations, 62 (1986), pp. 186–212.

[13] A. V. Kazhikov, *Resolution of boundary value problems for nonhomogeneous viscous fluids*, Dokl. Akad. Nauk, 216 (1974), pp. 1008–1010 (in Russian).

[14] H. Kim and H. Kozono, *Interior regularity criteria in weak spaces for the Navier-Stokes equations*, Manuscripta Math., 115 (2004), pp. 85–100.

[15] J. U. Kim, *Weak solutions of an initial boundary value problem for an incompressible viscous fluid with nonnegative density*, SIAM J. Math. Anal., 18 (1987), pp. 89–96.

[16] H. Kozono, *Weak solutions of the Navier-Stokes equations with test functions in the weak L^n-space*, Tohoku Math. J., 53 (2001), pp. 55–79.

[17] H. Kozono and Y. Taniuchi, *Bilinear estimates in BMO and the Navier-Stokes equations*, Math. Z., 235 (2000), pp. 173–194.

[18] H. Kozono and M. Yamazaki, *Uniqueness criterion of weak solutions to the stationary Navier-Stokes equations in exterior domains*, Nonlinear Anal. Ser. A: Theory Methods, 38 (1999), pp. 959–970.

[19] O. Ladyzhenskaya and V. A. Solonnikov, *Unique solvability of an initial and boundary value problem for viscous incompressible non-homogeneous fluids*, J. Soviet. Math., 9 (1978), pp. 697–749.

[20] J. Leray, *Essai sur le mouvement d'un liquide visqueux emplissant l'espace*, Acta Math., 63 (1934), pp. 193–248.

[21] P.-L. Lions, *Mathematical Topics in Fluid Mechanics. Volume* 1: *Imcompressible Models*, Oxford Lecture Series in Math. and Its Appl. 3, Oxford University Press, New York, 1996.

[22] H. Okamoto, *On the equation of nonstationary stratified fluid motion: Uniqueness and existence of solutions*, J. Fac. Sci. Univ. Tokyo Sect. 1A Math., 30 (1984), pp. 615–643.

[23] M. Padula, *On the existence and uniqueness of non-homogeneous motions in exterior domains*, Math. Z., 203 (1990), pp. 581–604.

[24] R. Salvi, *The equations of viscous incompressible non-homogeneous fluids: On the existence and regularity*, J. Austral. Math. Soc. Ser. B, 33 (1991), pp. 94–110.

[25] J. Serrin, *On the interior regularity of weak solutions of the Navier-Stokes equations*, Arch. Rational Mech. Anal., 9 (1962) pp. 187–195.

[26] J. Simon, *Ecoulement d'um fluide non homogène avec une densité initiale s'annulant*, C.R. Acad. Sci. Paris Ser. A, 15 (1978), pp. 1009–1012.

[27] J. Simon, *Nonhomogeneous viscous incompressible fluids: Existence of velocity, density, and pressure*, SIAM J. Math. Anal., 21 (1990), pp. 1093–1117.

[28] H. Sohr, *A regularity class for the Navier-Stokes equations in Lorentz spaces*, J. Evol. Equ., 1 (2001), pp. 441–467.

[29] M. Struwe, *On partial regularity results for the Navier-Stokes equations*, Comm. Pure Appl. Math., 41 (1988), pp. 437–458.

[30] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, 2nd ed., Johann Ambrosius Barth, Heidelberg, 1995.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1435–1453

HOMOGENIZATION OF QUASICONVEX INTEGRALS VIA THE PERIODIC UNFOLDING METHOD*

DOINA CIORANESCU†, ALAIN DAMLAMIAN‡, AND RICCARDO DE ARCANGELIS§

Abstract. The homogenization problem in the general case of quasiconvex integral energies with polynomial growth, defined on vector-valued configurations, was studied by the Γ-convergence methods in [A. Braides, *Rend. Accad. Naz. Sci.* XL *Mem. Mat.* (5), 9 (1985), pp. 313–321] and [S. Müller, *Arch. Rational Mech. Anal.*, 99 (1987), pp. 189–212]. This paper presents a new proof by means of the periodic unfolding method introduced in [D. Cioranescu, A. Damlamian, and G. Griso, *C. R. Math. Acad. Sci. Paris*, 335 (2002), pp. 99–104]. It is an elementary proof since it reduces the homogenization process to a weak convergence problem in an L^p-type space.

Key words. homogenization, quasiqconvexity, periodic unfolding method

AMS subject classifications. 49J45, 35B27, 74Q05

DOI. 10.1137/040620898

Introduction. The homogenization of periodic structures was carried out in the last 30 years for various kinds of problems involving differential equations and systems, as well as integral energies. Starting from the basic works [DGS], [BLP], [MT], several methods and techniques were developed to approach the analytic study of the asymptotic behavior of such structures, thus generating a great range of results in different settings and a wide bibliography; cf., for example, the volumes [At], [BP], [BD], [CDA], [CD], [DM], [JKO], [SP]. In particular, various kinds of problems concerning both scalar-valued and vector-valued configurations were considered also for convex and nonconvex energies. Among the various tools, we recall the two-scale method introduced in [N] (and used in [A]) and the "dilation" operation introduced in [ADH] (to study the homogenization of periodic media with double porosity).

The periodic unfolding method, proposed for the study of the homogenization of multiscale periodic problems in [CDG] combines the dilation technique with ideas from finite elements approximations. This approach reduces two-scale convergence to a mere weak convergence in an appropriate space.

The present paper is part of a series of ongoing work concerning the applications of the periodic unfolding method to homogenization. The first part in the series, [CDDA], considers the periodic homogenization of nonlinear integrals with convex densities and polynomial growth. It recovers the well-known homogenization results already established in [M1] and [CS], and various types of limit formulas, via weak convergence in a space of type L^p.

*Received by the editors December 15, 2004; accepted for publication (in revised form) June 7, 2005; published electronically January 10, 2006. This work is part of the European Research Training Network "Homogenization and Multiple Scales" (HMS 2000), under contract HPRN-2000-00109. It is also part of the 2004-G.N.A.M.P.A. Project "Metodi Variazionali per Strutture Sottili, Frontiere Oscillanti ed Energie Vincolate."

http://www.siam.org/journals/sima/37-5/62089.html

†Université Pierre et Marie Curie (Paris VI), Laboratoire Jacques-Louis Lions, 175 rue du Chevaleret, 75013 Paris Cedex 05, France (cioran@ann.jussieu.fr).

‡Laboratoire d'Analyse et de Mathématiques Appliquées, CNRS UMR 8050, Université Paris XII - Val de Marne, 61 avenue du Général de Gaulle, 94010 Créteil Cedex, France (damla@univ-paris12.fr).

§Università di Napoli "Federico II," Dipartimento di Matematica e Applicazioni "Renato Caccioppoli," via Cintia Complesso Monte S. Angelo, 80126 Napoli, Italy (dearcang@unina.it).

Here, we study the homogenization problem in the general case of quasiconvex integral energies defined on vector-valued configurations. For such energies, the homogenization result was originally obtained, under p-growth assumptions, by sophisticated Γ-convergence arguments (see [B] and [Mu]). Since then, many attempts at simplifying the proofs (for example, by using two-scale convergence) seem not to have borne fruit. Using the unfolding method, we propose a direct approach which, again, reduces to weak convergence in L^p spaces.

Homogenization processes for integral functionals defined on vector-valued configurations are particularly interesting in applications, since they describe various physical situations in nonlinear elasticity. For example, the study of the overall behavior of cellular elastic materials with very fine structure can be developed in the framework of the homogenization of the corresponding stored energy functionals, defined on the set of the admissible deformations and whose densities are quasiconvex. Indeed, as the structure of the material becomes finer and finer, it behaves more and more like a limit material, whose energy is just the one produced by the homogenization process.

Let m and n be positive integers. We denote by Y the unit reference cell $]0,1[^n$ and by $\mathcal{A}_0$ the class of all bounded open subsets of $\mathbb{R}^n$ with Lipschitz boundary.

Let f be a Carathéodory energy density, i.e., a function satisfying

$$\text{(0.1)}\qquad \begin{cases} f\colon (y,z)\in\mathbb{R}^n\times\mathbb{R}^{nm}\mapsto f(y,z)\in[0,+\infty[, \\ f(\cdot,z) \text{ Lebesgue measurable and } Y\text{-periodic for every } z\in\mathbb{R}^{nm}, \\ f(y,\cdot) \text{ continuous for a.e. } x\in\mathbb{R}^n. \end{cases}$$

Assume furthermore that

$$\text{(0.2)}\qquad f(y,\cdot) \text{ is quasiconvex for a.e. } y\in\mathbb{R}^n$$

(cf. section 1.2 below for the definition of quasiconvexity).

For $p\in[1,+\infty[$, $M>0$, and a Y-periodic $a\in L^1(Y)$, we consider the following growth conditions:

$$\text{(0.3)}\qquad f(y,z)\le a(y)+M|z|^p \quad \text{for a.e. } y\in\mathbb{R}^n \text{ and every } z\in\mathbb{R}^{nm},$$

$$\text{(0.4)}\qquad |z|^p\le f(y,z) \quad \text{for a.e. } y\in\mathbb{R}^n \text{ and every } z\in\mathbb{R}^{nm},$$

the first of which is slightly more general that the corresponding ones in [B] and [Mu].

Our purpose, for each $\Omega\in\mathcal{A}_0$ and every $\{\varepsilon_h\}\subseteq\,]0,+\infty[$ converging to 0, is to find the limit, as h goes to infinity, of the sequence of functionals

$$u\in W^{1,p}(\Omega;\mathbb{R}^m)\mapsto \int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u\Big)\,\mathrm{d}x.$$

We start in section 1 by recalling the result of Castaing on the existence of measurable selections of measurable multivalued functions (Theorem 1.1), together with some basic facts concerning the periodic unfolding method (Theorem 1.7), as well as an essential approximation result (Lemma 1.8) due to De Giorgi.

The convergence results are established in section 2. As in [B] and [Mu], we first prove that the limit below exists for every $z\in\mathbb{R}^{nm}$, thus defining the homogenized energy density f_{hom},

$$\text{(0.5)}\quad f_{\mathrm{hom}}\colon z\in\mathbb{R}^{nm}\mapsto \lim_{t\to+\infty}\frac{1}{t^n}\inf\Big\{\int_{tY} f(y,z+\nabla v)\,\mathrm{d}y : v\in W_0^{1,p}(tY;\mathbb{R}^m)\Big\}.$$

Then, in Theorem 2.5 we show that for every $\Omega \in \mathcal{A}_0$ and $u \in W^{1,p}(\Omega;\mathbb{R}^m)$,

$$\begin{aligned}
&\inf\left\{\liminf_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)\,\mathrm{d}x : \{u_h\}\subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ u_h\to u \text{ in } L^p(\Omega;\mathbb{R}^m)\right\}\\
&\quad=\inf\left\{\limsup_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)\,\mathrm{d}x :\right.\\
&\qquad\qquad\left.\{u_h\}\subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ u_h\to u \text{ in } L^p(\Omega;\mathbb{R}^m)\right\}\\
&\quad=\int_\Omega f_{\mathrm{hom}}(\nabla u)\,\mathrm{d}x.
\end{aligned}$$

Finally, Lemma 1.8 allows to complete the proof of Theorem 2.5, in the sense that for every Ω in $\mathcal{A}_0$ and u in $W^{1,p}(\Omega;\mathbb{R}^m)$,

$$\begin{aligned}
&\inf\left\{\liminf_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)\,\mathrm{d}x : \{u_h\}\subseteq u+W_0^{1,p}(\Omega;\mathbb{R}^m),\ u_h\to u \text{ in } L^p(\Omega;\mathbb{R}^m)\right\}\\
&\quad=\inf\left\{\limsup_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)\,\mathrm{d}x :\right.\\
&\qquad\qquad\left.\{u_h\}\subseteq u+W_0^{1,p}(\Omega;\mathbb{R}^m),\ u_h\to u \text{ in } L^p(\Omega;\mathbb{R}^m)\right\}\\
&\quad=\int_\Omega f_{\mathrm{hom}}(\nabla u)\,\mathrm{d}x.
\end{aligned}$$

In the convex case, i.e., when the condition

$$f(x,\cdot) \text{ is convex for a.e. } x\in\mathbb{R}^n \tag{0.6}$$

replaces hypothesis (0.2), the proof becomes even simpler. The result is stated in Theorem 3.1.

1. Preliminary results. For every $\Omega \in \mathcal{A}_0$, we denote by $\mathcal{L}(\Omega)$ the σ-algebra of the Lebesgue measurable subsets of Ω, and, for every $S \in \mathcal{L}(\Omega)$, by $|S|$ the n-dimensional Lebesgue measure of S.

By $\mathcal{L}(\Omega)\times\mathcal{L}(Y)$ we denote the product σ-algebra of $\mathcal{L}(\Omega)$ and $\mathcal{L}(Y)$. We recall that (cf., for example, Theorems III.11.17 and III.2.22 in [DS]), if $\Omega\in\mathcal{A}_0$ and $U \in L^1(\Omega;L^1(Y))$, then there exists an $\mathcal{L}(\Omega)\times\mathcal{L}(Y)$-measurable function $\widetilde{U}\colon\Omega\times Y\to\mathbb{R}$, uniquely determined up to a subset in $\mathcal{L}(\Omega)\times\mathcal{L}(Y)$ of zero measure, such that $\widetilde{U}(x,\cdot)=U(x)$ for a.e. $x\in\Omega$. Moreover, if $p\in[1,+\infty[$ and $U\in L^p(\Omega;L^p(Y))$, then $\widetilde{U}\in L^p(\Omega\times Y)$, and $\|\widetilde{U}\|_{L^p(\Omega\times Y)}=\|U\|_{L^p(\Omega;L^p(Y))}$. Consequently, we will not distinguish between U and $\widetilde{U}$ but rather denote them by U. With this convention in mind, $C_0^\infty(\Omega\times Y)$ turns out to be dense in $L^p(\Omega;L^p(Y))$, $L^p(\Omega\times Y)$, and in $L^p(Y;L^p(\Omega))$ with the respective norms.

In what follows, C will be used to denote various constants depending only upon n, m, p, and M.

1.1. Castaing's theorem on measurable selections. Let Ω, X be sets, and let Γ be a multifunction from Ω to X. A function $\sigma\colon\Omega\to X$ is said to be a selection of Γ if $\sigma(x)\in\Gamma(x)$ for every $x\in\Omega$. The measurable selection result below is proved in Theorem III.6 and Proposition III.11 in [CV].

THEOREM 1.1. *Let X be a separable metric space, $(\Omega, \mathcal{M})$ a measurable space, and Γ a multifunction from Ω to X. Assume that for every $x \in \Omega$, $\Gamma(x)$ is nonempty and complete in X. Assume, moreover, that for every closed subset F of X, $\{x \in \Omega : \Gamma(x) \cap F \neq \emptyset\}$ belongs to $\mathcal{M}$. Then Γ admits an $\mathcal{M}$-measurable selection.*

1.2. Quasiconvexity.

DEFINITION 1.2. *A continuous function $g\colon \mathbb{R}^{nm} \to [0, +\infty[$ is said to be quasiconvex whenever*

$$g(z) \leq \frac{1}{|A|} \int_A g(z + \nabla\varphi(x))\,\mathrm{d}x$$

for every $A \in \mathcal{A}_0$ and every $(z, \varphi) \in \mathbb{R}^{nm} \times C_0^1(A; \mathbb{R}^m)$.

Jensen's inequality implies that a convex function is also quasiconvex. It is also well known that if $n = 1$ or $m = 1$, quasiconvexity reduces to convexity. Moreover, it is easy to verify that g is quasiconvex if the above inequality holds for some $A \in \mathcal{A}_0$ and for all $(z, \varphi) \in \mathbb{R}^{nm} \times C_0^1(A; \mathbb{R}^m)$.

The following lower semicontinuity result for convex Carathéodory functions is well known (a direct consequence of Fatou's lemma).

PROPOSITION 1.3. *Let f satisfy* (0.1) *and* (0.6). *Let $p \in [1, +\infty[$ and $\Omega \in \mathcal{A}_0$. Then, the functional*

$$w \in L^p(\Omega; \mathbb{R}^{nm}) \mapsto \int_\Omega f(x, w(x))\,\mathrm{d}x$$

is sequentially weakly $L^p(\Omega; \mathbb{R}^{nm})$-lower semicontinuous.

In the quasiconvex case, it turns out that a similar result holds (cf., for example, [Mo], [Ba], [D], [Bu], [AF], [M2]).

PROPOSITION 1.4. *Assume that f satisfies* (0.1), (0.2), *and* (0.3) *for some $p \in [1, +\infty[$. Then, for every $\Omega \in \mathcal{A}_0$ the functional*

$$w \in W^{1,p}(\Omega; \mathbb{R}^m) \mapsto \int_\Omega f(x, \nabla w(x))\,\mathrm{d}x$$

is sequentially weakly $W^{1,p}(\Omega; \mathbb{R}^m)$-lower semicontinuous.

Let f be as in (0.1). Consider the following local Lipschitz condition for $p \in [1, +\infty[$, $C \geq 0$ and a Y-periodic $a \in L^1(Y)$:

$$\begin{aligned} &|f(x, z_1) - f(x, z_2)| \leq C(a(x)^{1-\frac{1}{p}} + |z_1|^{p-1} + |z_2|^{p-1})|z_1 - z_2| \\ &\qquad \text{for a.e. } x \in \mathbb{R}^n \text{ and every } z_1, z_2 \in \mathbb{R}^{nm}. \end{aligned} \tag{1.1}$$

The result below is well known (cf., for example, [Mo], [F]).

PROPOSITION 1.5. *Assume that f satisfies* (0.1), (0.2), *and* (0.3) *for some $p \in [1, +\infty[$ and $a \in L^1(Y)$. Then,* (1.1) *holds for some constant $C \geq 0$.*

1.3. The unfolding operator and its main properties. For every z in $\mathbb{R}^n$ we denote by $[z]$ the vector whose coordinates are the integer parts of the corresponding coordinates of z.

DEFINITION 1.6. *Let $\Omega \in \mathcal{A}_0$. For $\varepsilon > 0$, the unfolding operator $\mathcal{T}_\varepsilon\colon L^1(\Omega) \to L^1(\mathbb{R}^n \times Y)$ is defined as follows:*

$$\mathcal{T}_\varepsilon(v)(x, y) = \widetilde{v}\Big(\varepsilon\Big[\frac{x}{\varepsilon}\Big] + \varepsilon y\Big)$$

for every $v \in L^1(\Omega)$ and a.e. $(x,y) \in \mathbb{R}^n \times Y$, where $\widetilde{v}$ is the extension of v by zero outside Ω.

Let $\Omega \in \mathcal{A}_0$. For every $\varepsilon > 0$ we set

$$\Omega_\varepsilon \doteq \bigcup \{\varepsilon(\xi + \overline{Y}) : \xi \in \mathbb{Z}^n,\ \varepsilon(\xi + \overline{Y}) \cap \overline{\Omega} \neq \emptyset\}.$$

Then one has (cf. [CDG])

$$\int_{\Omega_\varepsilon \times Y} \mathcal{T}_\varepsilon(v)(x,y)\,\mathrm{d}x\,\mathrm{d}y = \int_\Omega v(x)\,\mathrm{d}x \quad \text{for every } \varepsilon > 0 \text{ and for every } v \text{ in } L^1(\Omega). \tag{1.2}$$

Furthermore

$$\mathcal{T}_\varepsilon(v) \to \widetilde{v} \text{ in } L^p(\mathbb{R}^n \times Y) \text{ as } \varepsilon \to 0 \quad \text{for every } p \in [1,+\infty[\text{ and } v \in L^p(\Omega). \tag{1.3}$$

If d is a positive integer, $w = (w_1, \ldots, w_d)$ belongs to $L^1(\Omega;\mathbb{R}^d)$, and $\varepsilon > 0$, we set

$$\mathcal{T}_\varepsilon(w) = (\mathcal{T}_\varepsilon(w_1), \ldots, \mathcal{T}_\varepsilon(w_d)).$$

Concerning the unfolding operator, the main result is the following (cf. Theorem 1 in [CDG]).

THEOREM 1.7. *Let $\Omega \in \mathcal{A}_0$, $p \in]1,+\infty[$, and let $\{\varepsilon_h\}$ be a sequence of positive numbers converging to 0. Let $\{v_h\}$ be a sequence converging weakly in $W^{1,p}(\Omega)$ to some v. Then, there exist a subsequence $\{h_k\}$ and $V \in L^p(\Omega; W^{1,p}_{\mathrm{per}}(Y))$ such that, as $k \to +\infty$,*

$$\mathcal{T}_{\varepsilon_{h_k}}(\nabla v_{h_k}) \to \nabla v + \nabla_y V \ \textit{weakly in } L^p(\Omega \times Y;\mathbb{R}^n).$$

1.4. A result of De Giorgi. We recall De Giorgi's argument allowing us to fix boundary values when computing Γ-limits (see [DG] for the scalar case and [F] for an extension to the vector-valued case).

LEMMA 1.8. *Assume that f satisfies* (0.1), (0.3), *and* (0.4) *for some $p \in [1,+\infty[$. Let $\{\nu_h\} \subseteq]0,+\infty[$ be increasing and diverging. Then, for every $\Omega \in \mathcal{A}_0$, and every $u \in W^{1,p}(\Omega;\mathbb{R}^m)$,*

$$\begin{aligned}
&\inf\left\{ \liminf_{h\to+\infty} \int_\Omega f(\nu_h x, \nabla u_h)\,\mathrm{d}x : \{u_h\} \subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ u_h \to u \ \textit{in } L^p(\Omega;\mathbb{R}^m) \right\} \\
&\quad = \inf\Bigg\{ \liminf_{h\to+\infty} \int_\Omega f(\nu_h x, \nabla u_h)\,\mathrm{d}x : \\
&\qquad\qquad \{u_h\} \subseteq u + W^{1,p}_0(\Omega;\mathbb{R}^m),\ u_h \to u \ \textit{in } L^p(\Omega;\mathbb{R}^m) \Bigg\}.
\end{aligned}$$

2. The homogenization result. Let f satisfy hypothesis (0.1), and let p be in $[1,+\infty[$. For every $t > 0$, set

$$\begin{aligned}
f_t \colon z \in \mathbb{R}^{nm} &\mapsto \frac{1}{t^n} \inf\left\{ \int_{tY} f(y, z + \nabla v)\,\mathrm{d}y : v \in W^{1,p}_0(tY;\mathbb{R}^m) \right\} \\
&= \inf\left\{ \int_Y f(ty, z + \nabla v)\,\mathrm{d}y : v \in W^{1,p}_0(Y;\mathbb{R}^m) \right\}.
\end{aligned} \tag{2.1}$$

For every $t > 0$, the function f_t is upper semicontinuous as the infimum of a family of continuous functions. If, in addition, hypothesis (0.3) holds, then clearly

$$f_t(z) \le \int_Y a(ty)\,\mathrm{d}y + M|z|^p \quad \text{for every } t > 0 \text{ and every } z \in \mathbb{R}^{nm}. \tag{2.2}$$

LEMMA 2.1. *Assume that f satisfies (0.1) and that $f(\cdot, z) \in L^1(Y)$ for every $z \in \mathbb{R}^{nm}$. For every $t > 0$, let f_t be defined by (2.1). Then, for every z in $\mathbb{R}^{nm}$, the limit $\lim_{t\to+\infty} f_t(z)$ exists and*

$$\lim_{t\to+\infty} f_t(z) = \inf_{h\in\mathbb{N}} f_h(z).$$

Proof. Fix z in $\mathbb{R}^{nm}$. By the periodicity of $f(\cdot, z)$, one has

$$f_{k\nu}(z) \le f_\nu(z) \quad \text{for every } \nu, k \in \mathbb{N}. \tag{2.3}$$

Let s, $t \in]0, +\infty[$ with $s < t$. The periodicity of $f(\cdot, z)$ implies that

$$\begin{aligned} f_s(z) &\ge \frac{t^n}{s^n}\left(f_t(z) - \frac{1}{t^n}\int_{tY\setminus sY} f(y,z)\,\mathrm{d}y\right) \\ &\ge f_t(z) - \frac{1}{t^n}\int_{([t]+1)Y\setminus[s]Y} f(y,z)\,\mathrm{d}y \\ &= f_t(z) - \frac{([t]+1)^n - [s]^n}{t^n}\int_Y f(y,z)\,\mathrm{d}y. \end{aligned} \tag{2.4}$$

Combining (2.4) and (2.3), we get

$$\begin{aligned} f_t(z) &\le f_{[\frac{t}{[s]}][s]}(z) + \frac{([t]+1)^n - [\frac{t}{[s]}]^n[s]^n}{t^n}\int_Y f(y,z)\,\mathrm{d}y \\ &\le f_{[s]}(z) + \frac{([t]+1)^n - [\frac{t}{[s]}]^n[s]^n}{t^n}\int_Y f(y,z)\,\mathrm{d}y. \end{aligned} \tag{2.5}$$

Hence, applying again (2.4),

$$f_t(z) \le f_{s-1}(z) + \left(\frac{([s]+1)^n - [s-1]^n}{[s]^n} + \frac{([t]+1)^n - [\frac{t}{[s]}]^n[s]^n}{t^n}\right)\int_Y f(y,z)\,\mathrm{d}y. \tag{2.6}$$

By the summability assumption on $f(\cdot, z)$, letting first t and then s go to infinity in (2.6), we conclude that $\limsup_{t\to+\infty} f_t(z) \le \liminf_{s\to+\infty} f_s(z)$, namely, that the limit $\lim_{t\to+\infty} f_t(z)$ exists.

Now choose $h \in \mathbb{N}$ and $t \in]h, +\infty[$. Then (2.5), together with (2.3), implies

$$\begin{aligned} f_t(z) &\le f_{[\frac{t}{h}]h}(z) + \frac{([t]+1)^n - [\frac{t}{h}]^n h^n}{t^n}\int_Y f(y,z)\,\mathrm{d}y \\ &\le f_h(z) + \frac{([t]+1)^n - [\frac{t}{h}]^n h^n}{t^n}\int_Y f(y,z)\,\mathrm{d}y. \end{aligned}$$

Letting t go to infinity, we obtain

$$\limsup_{t\to+\infty} f_t(z) \le f_h(z)$$

for every $h \in \mathbb{N}$. This inequality completes the proof of the lemma. $\square$

LEMMA 2.2. *Assume that f satisfies* (0.1), (0.2), (0.3), *and* (0.4) *for some $p \in]1,+\infty[$. For every $t > 0$, let f_t be defined in* (2.1). *Then, for every Ω in $\mathcal{A}_0$, $t > 0$, and every u in $W^{1,p}(\Omega;\mathbb{R}^m)$, the following inequality holds:*

$$\int_\Omega f_t(\nabla u)\,\mathrm{d}x$$
$$= \frac{1}{t^n}\inf\left\{\int_{\Omega\times tY} f(y,\nabla u(x)+\nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V \in L^p(\Omega;W_0^{1,p}(tY;\mathbb{R}^m))\right\}.$$

Proof. Let Ω, t, u be as above. Fix V in $W_0^{1,p}(tY;\mathbb{R}^m)$. Obviously, for a.e. $x \in \Omega$,

$$\begin{aligned}\frac{1}{t^n}\int_{tY} &f(y,\nabla u(x)+\nabla_y V(x,y))\,\mathrm{d}y\\ &\geq \frac{1}{t^n}\inf\left\{\int_{tY} f(y,\nabla u(x)+\nabla v(y))\,\mathrm{d}y : v \in W_0^{1,p}(tY;\mathbb{R}^m)\right\}\\ &= f_t(\nabla u(x)),\end{aligned}$$

from which follows

$$\begin{aligned}\frac{1}{t^n}\inf&\left\{\int_{\Omega\times tY} f(y,\nabla u(x)+\nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V \in L^p(\Omega;W_0^{1,p}(tY;\mathbb{R}^m))\right\}\\ &\geq \int_\Omega f_t(\nabla u(x))\,\mathrm{d}x.\end{aligned}$$

The reverse inequality is obvious if $\int_\Omega f_t(\nabla u(x))\,\mathrm{d}x = +\infty$. If $\int_\Omega f_t(\nabla u(x))\,\mathrm{d}x$ is finite, we make use of Castaing's selection theorem applied to the metric space $X = W_0^{1,p}(tY;\mathbb{R}^m)$.

Let Γ be the multifunction defined by

$$\Gamma\colon z \in \mathbb{R}^{nm} \mapsto \left\{v \in W_0^{1,p}(tY;\mathbb{R}^m) : \frac{1}{t^n}\int_{tY} f(y,z+\nabla v(y))\,\mathrm{d}y = f_t(z)\right\}.$$

By (0.4) and Proposition 1.4 applied to tY, the infimum defining $f_t(z)$ is achieved for every $z \in \mathbb{R}^{nm}$. So, $\Gamma(z)$ is nonempty and weakly closed, and hence strongly closed. We claim that Γ has a $\mathcal{B}(\mathbb{R}^{nm})$-measurable selection (where $\mathcal{B}(\mathbb{R}^{nm})$ denotes the σ-algebra of the Borel subsets of $\mathbb{R}^{nm}$). By Theorem 1.1, it is enough to show that for every strongly closed subset F of $W_0^{1,p}(tY;\mathbb{R}^m)$, one has

$$\Gamma^-(F) \doteq \{\zeta \in \mathbb{R}^{nm} : \Gamma(\zeta)\cap F \neq \emptyset\} \in \mathcal{B}(\mathbb{R}^{nm}).$$

We first assume that F is a closed ball in $W_0^{1,p}(tY;\mathbb{R}^m)$ and prove that $\Gamma^-(F)$ is closed. Let $\{z_h\} \subseteq \Gamma^-(F)$, $z \in \mathbb{R}^{nm}$, with $z_h \to z$. For every $h \in \mathbb{N}$, let $v_h \in \Gamma(z_h)\cap F$. Then, $\{v_h\}$ turns out to be bounded in $W_0^{1,p}(tY;\mathbb{R}^m)$. In fact, (0.4), the upper semicontinuity of f_t, (0.3), and (2.2) imply that

$$\begin{aligned}\limsup_{h\to+\infty}\frac{1}{t^n}\int_{tY}|z_h+\nabla v_h(y)|^p\,\mathrm{d}y &\leq \limsup_{h\to+\infty}\frac{1}{t^n}\int_{tY} f(y,z_h+\nabla v_h(y))\,\mathrm{d}y\\ &= \limsup_{h\to+\infty} f_t(z_h) \leq f_t(z) < +\infty.\end{aligned}$$

Therefore, there is a subsequence $\{v_{h_k}\}$ of $\{v_h\}$, and some v_∞ in F such that $v_{h_k} \rightharpoonup v_\infty$ in $W_0^{1,p}(tY;\mathbb{R}^m)$. Then, Proposition 1.4 (applied to tY) and the upper semicontinuity of f_t imply

$$f_t(z) \leq \frac{1}{t^n} \int_{tY} f(y, z + \nabla v_\infty(y))\,\mathrm{d}y \leq \liminf_{k\to+\infty} \frac{1}{t^n} \int_{tY} f(y, z_{h_k} + \nabla v_{h_k}(y))\,\mathrm{d}y$$
$$\leq \limsup_{k\to+\infty} f_t(z_{h_k}) \leq f_t(z)$$

so that $v_\infty \in \Gamma(z) \cap F$ and $z \in \Gamma^-(F)$.

By using the separability of $W_0^{1,p}(tY;\mathbb{R}^m)$, every strongly closed subset F of $W_0^{1,p}(tY;\mathbb{R}^m)$ can be written as a countable intersection of countable unions of balls. Consequently, $\Gamma^-(F)$ is itself a countable intersection of countable unions of closed sets, and hence $\Gamma^-(F) \in \mathcal{B}(\mathbb{R}^{nm})$. By Theorem 1.1, Γ admits a $\mathcal{B}(\mathbb{R}^{nm})$-measurable selection σ.

For a.e. $x \in \Omega$, set $U(x) = \sigma(\nabla u(x))$. Then, U is $\mathcal{L}(\Omega)$-measurable with values in $W_0^{1,p}(tY;\mathbb{R}^m)$. Moreover, for a.e. $x \in \Omega$,

$$f_t(\nabla u(x)) = \frac{1}{t^n} \int_{tY} f(y, \nabla u(x) + \nabla_y U(x)(y))\,\mathrm{d}y.$$

Since $f_t(\nabla u)$ is summable, the previous equality, together with (0.4), implies that $\nabla_y U \in L^p(\Omega\times tY;\mathbb{R}^{nm})$. Thus, by the Poincaré inequality, U is in $L^p(\Omega; W_0^{1,p}(tY;\mathbb{R}^m))$. Consequently,

$$\begin{aligned}
&\int_\Omega f_t(\nabla u(x))\,\mathrm{d}x \\
&\quad= \frac{1}{t^n} \int_{\Omega\times tY} f(y, \nabla u(x) + \nabla_y U(x)(y))\,\mathrm{d}x\,\mathrm{d}y \\
&\quad\geq \frac{1}{t^n} \inf\left\{ \int_{\Omega\times tY} f(y, \nabla u(x) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V \in L^p(\Omega; W_0^{1,p}(tY;\mathbb{R}^m)) \right\},
\end{aligned}$$

and this completes the proof of the lemma. ∎

PROPOSITION 2.3. *Assume that f satisfies* (0.1), (0.2), (0.3), *and* (0.4) *for some $p \in\,]1, +\infty[$. Let f_{hom} be defined by* (0.5). *Then, for every $\Omega \in \mathcal{A}_0$ and $u \in W^{1,p}(\Omega;\mathbb{R}^m)$, the limit below exists, and*

$$\begin{aligned}
&\lim_{t\to+\infty} \frac{1}{t^n} \inf\left\{ \int_{\Omega\times tY} f(y, \nabla u(x) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V \in L^p(\Omega; W_0^{1,p}(tY;\mathbb{R}^m)) \right\} \\
&\qquad= \int_\Omega f_{\mathrm{hom}}(\nabla u(x))\,\mathrm{d}x.
\end{aligned}$$

Proof. The conclusion follows from Lemma 2.2, Lemma 2.1, (2.2), and the Lebesgue dominated convergence theorem. ∎

Remark 2.4. Under the assumptions of Proposition 2.3, note the various formulations for the limit energy, the latter deriving from Lemma 2.1:

$$\begin{aligned}
&\int_\Omega f_{\mathrm{hom}}(\nabla u)\,\mathrm{d}x \\
&\quad= \lim_{t\to+\infty} \frac{1}{t^n} \inf\left\{ \int_{\Omega\times tY} f(y, \nabla u(x) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V \in L^p(\Omega; W_0^{1,p}(tY;\mathbb{R}^m)) \right\} \\
&\quad= \lim_{t\to+\infty} \inf\left\{ \int_{\Omega\times Y} f(ty, \nabla u(x) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V \in L^p(\Omega; W_0^{1,p}(Y;\mathbb{R}^m)) \right\} \\
&\quad= \inf_{h\in\mathbb{N}} \inf\left\{ \int_{\Omega\times Y} f(hy, \nabla u(x) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V \in L^p(\Omega; W_0^{1,p}(Y;\mathbb{R}^m)) \right\}.
\end{aligned}$$

We now state the main homogenization result for nonlinear energy integral functionals.

THEOREM 2.5. *Let f satisfy* (0.1) *and* (0.2). *Let $p \in]1, +\infty[$, and assume that* (0.3) *and* (0.4) *hold. Let f_{hom} be defined by* (0.5). *Then, for every $\{\varepsilon_h\} \subseteq]0, +\infty[$ converging to* 0, *Ω in $\mathcal{A}_0$, and u in $W^{1,p}(\Omega; \mathbb{R}^m)$,*

$$\begin{aligned}
&\int_\Omega f_{\mathrm{hom}}(\nabla u)\,\mathrm{d}x \\
&\quad = \inf\left\{\liminf_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h}, \nabla u_h\Big)\,\mathrm{d}x : \{u_h\} \subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ u_h \to u \ in\ L^p(\Omega;\mathbb{R}^m)\right\} \\
&\quad = \inf\left\{\limsup_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h}, \nabla u_h\Big)\,\mathrm{d}x : \{u_h\} \subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ u_h \to u \ in\ L^p(\Omega;\mathbb{R}^m)\right\} \\
&\quad = \inf\left\{\liminf_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h}, \nabla u_h\Big)\,\mathrm{d}x : \right. \\
&\qquad\qquad \left. \{u_h\} \subseteq u + W^{1,p}_0(\Omega;\mathbb{R}^m),\ u_h \to u \ in\ L^p(\Omega;\mathbb{R}^m)\right\} \\
&\quad = \inf\left\{\limsup_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h}, \nabla u_h\Big)\,\mathrm{d}x : \right. \\
&\qquad\qquad \left. \{u_h\} \subseteq u + W^{1,p}_0(\Omega;\mathbb{R}^m),\ u_h \to u \ in\ L^p(\Omega;\mathbb{R}^m)\right\}.
\end{aligned}$$

The proof uses the following lemmas and will be given at the end of this section.

LEMMA 2.6. *Assume that f satisfies* (0.1) *and* (0.3) *for some $p \in]1, +\infty[$. Let $\{\varepsilon_h\} \subseteq]0, +\infty[$ converge to* 0, *Ω in $\mathcal{A}_0$, and u in $W^{1,p}(\Omega; \mathbb{R}^m)$. Then*

$$\begin{aligned}
&\inf\left\{\limsup_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h}, \nabla u_h\Big)\,\mathrm{d}x : \{u_h\} \subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ u_h \to u \ in\ L^p(\Omega;\mathbb{R}^m)\right\} \\
&\quad \le \inf_{k\in\mathbb{N}} \frac{1}{k^n} \inf\left\{\int_{\Omega\times kY} f(y, \nabla u(x) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V \in L^p(\Omega; W^{1,p}_{\mathrm{per}}(kY;\mathbb{R}^m))\right\}.
\end{aligned}$$

Proof. Let $k \in \mathbb{N}$, and let $U \in C^1(\mathbb{R}^n \times \mathbb{R}^n; \mathbb{R}^m)$ with $U(x,\cdot)$ kY-periodic for every $x \in \Omega$. For every $h \in \mathbb{N}$ and $x \in \Omega$, set $u_h(x) = u(x) + \varepsilon_h U(x, \frac{x}{\varepsilon_h})$. Clearly, $\nabla u_h(x) = \nabla u(x) + \varepsilon_h \nabla_x U(x, \frac{x}{\varepsilon_h}) + \nabla_y U(x, \frac{x}{\varepsilon_h})$ for every $h \in \mathbb{N}$ and $x \in \Omega$.

Using the unfolding operator, (1.2), and the periodicity properties of f and U, one has

$$\begin{aligned}
\int_\Omega f\Big(\frac{x}{\varepsilon_h}, \nabla u_h\Big)\,\mathrm{d}x &= \int_{\Omega_{k\varepsilon_h}\times Y} \mathcal{T}_{k\varepsilon_h}\Big(f\Big(\frac{\cdot}{\varepsilon_h}, \nabla u_h(\cdot)\Big)\Big)(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{\Omega_{k\varepsilon_h}\times Y} f\Big(\frac{1}{\varepsilon_h}\Big(k\varepsilon_h\Big[\frac{x}{k\varepsilon_h}\Big] + k\varepsilon_h y\Big), \mathcal{T}_{k\varepsilon_h}(\nabla u_h)(x,y)\Big)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{\Omega_{k\varepsilon_h}\times Y} f\Big(ky, \mathcal{T}_{k\varepsilon_h}(\nabla u)(x,y) + \varepsilon_h \nabla_x U\Big(k\varepsilon_h\Big[\frac{x}{k\varepsilon_h}\Big] + k\varepsilon_h y, k\Big[\frac{x}{k\varepsilon_h}\Big] + ky\Big) \\
&\qquad + \nabla_y U\Big(k\varepsilon_h\Big[\frac{x}{k\varepsilon_h}\Big] + k\varepsilon_h y, k\Big[\frac{x}{k\varepsilon_h}\Big] + ky\Big)\Big)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{\Omega_{k\varepsilon_h}\times Y} f\Big(ky, \mathcal{T}_{k\varepsilon_h}(\nabla u)(x,y) + \varepsilon_h \nabla_x U\Big(k\varepsilon_h\Big[\frac{x}{k\varepsilon_h}\Big] + k\varepsilon_h y, ky\Big) \\
&\qquad + \nabla_y U\Big(k\varepsilon_h\Big[\frac{x}{k\varepsilon_h}\Big] + k\varepsilon_h y, ky\Big)\Big)\,\mathrm{d}x\,\mathrm{d}y.
\end{aligned}$$

Since $\{u_h\}$ converges uniformly to u in Ω, we have

$$\begin{aligned}\inf\Big\{ \limsup_{h\to+\infty} \int_\Omega f\Big(\frac{x}{\varepsilon_h}, \nabla u_h\Big)\,\mathrm{d}x : \{v_h\} \subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ v_h \to u \text{ in } L^p(\Omega;\mathbb{R}^m)\Big\}\\ \le \limsup_{h\to+\infty} \int_{\Omega_{k\varepsilon_h}\times Y} f\Big(ky, \mathcal{T}_{k\varepsilon_h}(\nabla u)(x,y) + \varepsilon_h \nabla_x U\Big(k\varepsilon_h\Big[\frac{x}{k\varepsilon_h}\Big] + k\varepsilon_h y, ky\Big)\\ + \nabla_y U\Big(k\varepsilon_h\Big[\frac{x}{k\varepsilon_h}\Big] + k\varepsilon_h y, ky\Big)\Big)\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

On the other hand, by the continuity properties $\nabla_x U$ and $\nabla_y U$, one has

$$\varepsilon_h \nabla_x U\Big(k\varepsilon_h\Big[\frac{\cdot}{k\varepsilon_h}\Big] + k\varepsilon_h\cdot,\cdot\Big) \to 0,\ \nabla_y U\Big(k\varepsilon_h\Big[\frac{\cdot}{k\varepsilon_h}\Big] + k\varepsilon_h\cdot,\cdot\Big) \to \nabla_y U \text{ uniformly in } \Omega.$$

By (1.3), $\mathcal{T}_{k\varepsilon_h}(\nabla u) \to \widetilde{\nabla u}$ in $L^p(\mathbb{R}^n\times Y;\mathbb{R}^{nm})$. Then, since $|\partial\Omega| = 0$, using (0.3) we conclude that

$$\begin{aligned}\lim_{h\to+\infty} \int_{\Omega_{k\varepsilon_h}\times Y} f\Big(ky, \mathcal{T}_{k\varepsilon_h}(\nabla u)(x,y) + \varepsilon_h \nabla_x U\Big(k\varepsilon_h\Big[\frac{x}{k\varepsilon_h}\Big] + k\varepsilon_h y, ky\Big)\\ + \nabla_y U\Big(k\varepsilon_h\Big[\frac{x}{k\varepsilon_h}\Big] + k\varepsilon_h y, ky\Big)\Big)\,\mathrm{d}x\,\mathrm{d}y\\ = \int_{\Omega\times Y} f(ky, \nabla u(x) + \nabla_y U(x,ky))\,\mathrm{d}x\,\mathrm{d}y\\ = \frac{1}{k^n}\int_{\Omega\times kY} f(y, \nabla u(x) + \nabla_y U(x,y))\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

By the above inequalities we get

$$\begin{aligned}\inf\Big\{ \limsup_{h\to+\infty} \int_\Omega f\Big(\frac{x}{\varepsilon_h}, \nabla u_h\Big)\,\mathrm{d}x :\\ \{v_h\} \subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ v_h \to u \text{ in } L^p(\Omega;\mathbb{R}^m)\Big\}\\ \le \frac{1}{k^n}\int_{\Omega\times kY} f(y, \nabla u(x) + \nabla_y U(x,y))\,\mathrm{d}x\,\mathrm{d}y\end{aligned} \tag{2.7}$$

for every $U \in C^1(\mathbb{R}^n\times\mathbb{R}^n;\mathbb{R}^m)$ with $U(x,\cdot)$ kY-periodic for every $x\in\Omega$.

By virtue of (0.1) and (0.3), we observe that the right-hand side of (2.7) is continuous with respect to $U \in L^p(\Omega; W^{1,p}_{\mathrm{per}}(kY;\mathbb{R}^m))$ and that the set of functions U in $C^1(\mathbb{R}^n\times\mathbb{R}^n;\mathbb{R}^m)$ with $U(x,\cdot)$ kY-periodic for every $x\in\Omega$ is dense in $L^p(\Omega; W^{1,p}_{\mathrm{per}}(kY;\mathbb{R}^m))$. Consequently, (2.7) holds for every $U \in L^p(\Omega; W^{1,p}_{\mathrm{per}}(kY;\mathbb{R}^m))$ as well. Taking first the infimum as U varies in $L^p(\Omega; W^{1,p}_{\mathrm{per}}(kY;\mathbb{R}^m))$, and then the infimum on k, completes the proof. ∎

Remark 2.7. Since, for every integer k, the space $W^{1,p}_0(kY;\mathbb{R}^m)$ is a subspace of $W^{1,p}_{\mathrm{per}}(kY;\mathbb{R}^m)$, Lemma 2.6 holds with the V in $L^p(\Omega; W^{1,p}_0(kY;\mathbb{R}^m))$.

LEMMA 2.8. *Assume that f satisfies* (0.1), (0.3), (0.4), *and* (1.1) *for some $p \in [1,+\infty[$. Let $\{\varepsilon_h\} \subseteq]0,+\infty[$ converge to 0, Ω in $\mathcal{A}_0$, and u in $W^{1,p}(\Omega;\mathbb{R}^m)$. Then*

$$\inf\Big\{\liminf_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u+\nabla u_h\Big)\,\mathrm{d}x : \{u_h\}\subseteq W_0^{1,p}(\Omega;\mathbb{R}^m),\ u_h\to 0 \text{ in } L^p(\Omega;\mathbb{R}^m)\Big\}$$

$$\geq \sup_{\nu\in\mathbb{N}}\inf\Big\{\liminf_{h\to+\infty}\int_\Omega f(\nu hx,\nabla u+\nabla v_h)\,\mathrm{d}x : \{v_h\}\subseteq W_0^{1,p}(\Omega;\mathbb{R}^m),\ v_h\to 0 \text{ in } L^p(\Omega;\mathbb{R}^m)\Big\}.$$

Proof. Let $\{u_h\}\subseteq W_0^{1,p}(\Omega;\mathbb{R}^m)$ be such that $u_h\to 0$ in $L^p(\Omega;\mathbb{R}^m)$. Clearly, we can assume

$$\liminf_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u+\nabla u_h\Big)\,\mathrm{d}x<+\infty.$$

Let Ω' in $\mathcal{A}_0$ satisfy $\overline{\Omega}\subseteq\Omega'$. For every $h\in\mathbb{N}$, we still denote u_h the zero extension of u_h from Ω to Ω'. We also select an extension of u to Ω' in $W^{1,p}(\Omega';\mathbb{R}^m)$ which, for simplicity, is denoted u. Then, clearly

$$\begin{aligned}&\int_{\Omega'} f\Big(\frac{x}{\varepsilon_h},\nabla u+\nabla u_h\Big)\,\mathrm{d}x\\ &\quad\leq\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u+\nabla u_h\Big)\,\mathrm{d}x+\int_{\Omega'\setminus\Omega}\Big(a\Big(\frac{x}{\varepsilon_h}\Big)+M|\nabla u(x)|^p\Big)\,\mathrm{d}x\\ &\quad\text{for every } h\in\mathbb{N}.\end{aligned}\tag{2.8}$$

Let $\{h_j\}\subseteq\mathbb{N}$ be such that $\{[\frac{1}{\nu\varepsilon_{h_j}}]\}$ is strictly increasing, and

$$\lim_{j\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_{h_j}},\nabla u+\nabla u_{h_j}\Big)\,\mathrm{d}x=\liminf_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u+\nabla u_h\Big)\,\mathrm{d}x.\tag{2.9}$$

Then (0.4) provides that $\sup_{j\in\mathbb{N}}\int_\Omega|\nabla u_{h_j}|^p\,\mathrm{d}x<+\infty$.

Now, let ν be a positive integer. For every $j\in\mathbb{N}$, set $\theta_j=\nu\varepsilon_{h_j}[\frac{1}{\nu\varepsilon_{h_j}}]$, and observe that $\theta_j\to 1^-$ as j goes to $+\infty$. For every $j\in\mathbb{N}$, define $v_j\doteq\frac{1}{\theta_j}u_{h_j}(\theta_j\cdot)$. Clearly, v_j has support in $\frac{1}{\theta_j}\overline{\Omega}$, which is included in Ω' for large values of j, so it is in $W_0^{1,p}(\Omega';\mathbb{R}^m)$ for the same values of j. From its definition, it is straightforward that $\{v_j\}$ goes to zero in $L^p(\Omega;\mathbb{R}^m)$ because $\{u_h\}$ does. Moreover,

$$\sup_{j\in\mathbb{N}}\int_{\Omega'}|\nabla v_j|^p\,\mathrm{d}x<+\infty.\tag{2.10}$$

By performing the change of variable $x=\theta_jx'$, for j large enough, we get successively

$$\begin{aligned}\int_{\Omega'} f\Big(\frac{x}{\varepsilon_{h_j}},\nabla u+\nabla u_{h_j}\Big)\,\mathrm{d}x&=\theta_j^n\int_{\theta_j^{-1}\Omega'} f\Big(\frac{\theta_jx'}{\varepsilon_{h_j}},(\nabla_xu+\nabla_xu_{h_j})(\theta_jx')\Big)\,\mathrm{d}x'\\ &=\theta_j^n\int_{\theta_j^{-1}\Omega'} f\Big(\frac{\theta_jx'}{\varepsilon_{h_j}},\nabla_xu(\theta_jx')+\nabla_{x'}v_j(x')\Big)\,\mathrm{d}x'\\ &\geq\theta_j^n\int_\Omega f\Big(\frac{\theta_jx'}{\varepsilon_{h_j}},\nabla_xu(\theta_jx')+\nabla_{x'}v_j(x')\Big)\,\mathrm{d}x'.\end{aligned}$$

Consequently, using (1.1) and Hölder's inequality,

$$\begin{aligned}
&\int_{\Omega'} f\Big(\frac{x}{\varepsilon_{h_j}}, \nabla u + \nabla u_{h_j}\Big)\,\mathrm{d}x \\
&\quad\geq \theta_j^n \int_\Omega f\Big(\frac{\theta_j x'}{\varepsilon_{h_j}}, \nabla_x u(x') + \nabla_{x'} v_j(x')\Big)\,\mathrm{d}x' \\
&\quad - \theta_j^n C\bigg\{\bigg(\int_\Omega a\Big(\frac{\theta_j x'}{\varepsilon_{h_j}}\Big)\,\mathrm{d}x'\bigg)^{1-\frac{1}{p}} + \bigg(\int_\Omega |\nabla_x u(\theta_j x')|^p\,\mathrm{d}x'\bigg)^{1-\frac{1}{p}} \\
&\qquad + \bigg(\int_\Omega |\nabla_x u(x')|^p\,\mathrm{d}x'\bigg)^{1-\frac{1}{p}} + \bigg(\int_\Omega |\nabla_{x'} v_j(x')|^p\,\mathrm{d}x'\bigg)^{1-\frac{1}{p}}\bigg\} \\
&\qquad \times \bigg(\int_\Omega |\nabla_x u(\theta_j x') - \nabla_x u(x')|^p\,\mathrm{d}x'\bigg)^{\frac{1}{p}} \quad \text{for } j \text{ large enough.}
\end{aligned} \tag{2.11}$$

Now, for every $j \in \mathbb{N}$, set $n_j = [\frac{1}{\nu\varepsilon_{h_j}}]$. Since $\frac{\theta_j}{\varepsilon_{h_j}} = \nu[\frac{1}{\nu\varepsilon_{h_j}}] = \nu n_j$, going to the limit in (2.11), exploiting the periodicity of a and (2.10), gives

$$\begin{aligned}
\liminf_{j\to+\infty} \int_\Omega f(\nu n_j x, \nabla u + \nabla v_j)\,\mathrm{d}x &= \liminf_{j\to+\infty} \theta_j^n \int_\Omega f\Big(\frac{\theta_j x}{\varepsilon_{h_j}}, \nabla u + \nabla v_j\Big)\,\mathrm{d}x \\
&\leq \lim_{j\to+\infty} \int_{\Omega'} f\Big(\frac{x}{\varepsilon_{h_j}}, \nabla u + \nabla u_{h_j}\Big)\,\mathrm{d}x,
\end{aligned} \tag{2.12}$$

where we used the fact that $\int_\Omega |\nabla_x u(\theta_j x') - \nabla_x u(x')|^p\,\mathrm{d}x'$ converges to 0 by the continuity as $\theta_j \to 1$ of the scaling operator in L^p.

Combining (2.12), (2.8), and (2.9), and using the periodicity of the function a, yields

$$\begin{aligned}
&\liminf_{j\to+\infty} \int_\Omega f(\nu n_j x, \nabla u + \nabla v_j)\,\mathrm{d}x \\
&\quad \leq \liminf_{h\to+\infty} \int_\Omega f\Big(\frac{x}{\varepsilon_h}, \nabla u + \nabla u_h\Big)\,\mathrm{d}x \\
&\qquad + |\Omega' \setminus \Omega| \int_Y a(y)\,\mathrm{d}y + M \int_{\Omega'\setminus\Omega} |\nabla u(x)|^p\,\mathrm{d}x.
\end{aligned} \tag{2.13}$$

Recall that $\{n_j\}$ is strictly increasing. Thus, for every $h \in \mathbb{N}$ we define $\{w_h\}$ as

$$w_h = \begin{cases} v_j & \text{if } h = n_j \text{ for some } j \in \mathbb{N}, \\ 0 & \text{otherwise.} \end{cases}$$

Observe that $\{w_h\} \subseteq W^{1,p}(\Omega;\mathbb{R}^m)$, and that $w_h \to 0$ in $L^p(\Omega;\mathbb{R}^m)$. Then, since $|\partial\Omega| = 0$ and $\{n_j\} \subseteq \mathbb{N}$, once Ω' shrinks to $\overline{\Omega}$, by (2.13) we conclude that

$$\begin{aligned}
&\inf\bigg\{\liminf_{h\to+\infty} \int_\Omega f(\nu h x, \nabla u + \nabla v_h)\,\mathrm{d}x : \{v_h\} \subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ v_h \to 0 \text{ in } L^p(\Omega;\mathbb{R}^m)\bigg\} \\
&\quad \leq \liminf_{h\to+\infty} \int_\Omega f(\nu h x, \nabla u + \nabla w_h)\,\mathrm{d}x \leq \liminf_{j\to+\infty} \int_\Omega f(\nu n_j x, \nabla u + \nabla v_j)\,\mathrm{d}x \\
&\quad \leq \liminf_{h\to+\infty} \int_\Omega f\Big(\frac{x}{\varepsilon_h}, \nabla u + \nabla u_h\Big)\,\mathrm{d}x.
\end{aligned}$$

The desired conclusion follows by using Lemma 1.8 and taking the supremum for ν in $\mathbb{N}$. ◻

LEMMA 2.9. *Assume that f satisfies* (0.1), (0.3), (0.4), *and* (1.1) *for some $p \in [1, +\infty[$. Let $\{\varepsilon_h\} \subseteq]0, +\infty[$ converge to 0, Ω in $\mathcal{A}_0$, and u in $W^{1,p}(\Omega; \mathbb{R}^m)$. Then*

$$\begin{aligned}
&\inf\left\{\liminf_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h}, \nabla u + \nabla u_h\Big)\,\mathrm{d}x : \{u_h\} \subseteq W_0^{1,p}(\Omega;\mathbb{R}^m),\ u_h \to 0 \textit{ in } L^p(\Omega;\mathbb{R}^m)\right\}\\
&\quad \geq \liminf_{h\to+\infty}\ \inf\left\{\int_{\Omega\times Y} f(hy, \nabla u(x) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V \in L^p(\Omega; W_0^{1,p}(Y;\mathbb{R}^m))\right\}.
\end{aligned}$$

Proof. Since Ω has Lipschitz boundary, u can be extended as an element of $W^{1,p}(\mathbb{R}^n;\mathbb{R}^m)$ without changing notation.

To prove the lemma, we estimate from below the right-hand side of the inequality in Lemma 2.8.

Let $\nu \in \mathbb{N}$, and let $\{v_h\} \subseteq W_0^{1,p}(\Omega;\mathbb{R}^m)$ be such that $v_h \to 0$ in $L^p(\Omega;\mathbb{R}^m)$. As before, each v_h can be considered as extended by zero outside of Ω. For every $j \in \mathbb{Z}^n$, set $Y_{\nu,j} = \frac{1}{\nu}(j + Y)$ and let $J_\nu = \{j \in \mathbb{Z}^n : \overline{Y_{\nu,j}} \cap \overline{\Omega} \neq \emptyset\}$. Then

$$\begin{aligned}
&\sum_{j\in J_\nu}\liminf_{h\to+\infty}\int_{Y_{\nu,j}} f(\nu h x, \nabla u + \nabla v_h)\,\mathrm{d}x\\
&\leq \liminf_{h\to+\infty}\int_\Omega f(\nu h x, \nabla u + \nabla v_h)\,\mathrm{d}x + |\Omega_{1/\nu}\setminus\Omega|\int_Y a\,\mathrm{d}y + M\int_{\Omega_{1/\nu}\setminus\Omega}|\nabla u|^p\,\mathrm{d}x.
\end{aligned} \tag{2.14}$$

For every $j \in J_\nu$, we apply Lemma 1.8 to $Y_{\nu,j}$, and get $\{w_{\nu,j,h}\} \subseteq W_0^{1,p}(Y_{\nu,j};\mathbb{R}^m)$ such that $w_{\nu,j,h} \to 0$ in $L^p(Y_{\nu,j};\mathbb{R}^m)$, and

$$\liminf_{h\to+\infty}\int_{Y_{\nu,j}} f(\nu h x, \nabla u + \nabla w_{\nu,j,h})\,\mathrm{d}x \leq \liminf_{h\to+\infty}\int_{Y_{\nu,j}} f(\nu h x, \nabla u + \nabla v_h)\,\mathrm{d}x. \tag{2.15}$$

For every $j \in J_\nu$ and $h \in \mathbb{N}$, using the unfolding operator $\mathcal{T}_{1/\nu}$ on $Y_{\nu,j}$, together with the same arguments as in Lemma 2.6, gives

$$\begin{aligned}
&\int_{Y_{\nu,j}} f(\nu h x, \nabla u + \nabla w_{\nu,j,h})\,\mathrm{d}x\\
&\qquad \geq \int_{Y_{\nu,j}\times Y} f(hy, \mathcal{T}_{1/\nu}(\nabla u)(x,y) + \mathcal{T}_{1/\nu}(\nabla w_{\nu,j,h})(x,y))\,\mathrm{d}x\,\mathrm{d}y.
\end{aligned}$$

Observe now that for every $h \in \mathbb{N}$ and $j \in J_\nu$,

$$\mathcal{T}_{1/\nu}(\nabla w_{\nu,j,h}) = \nu\nabla_y\mathcal{T}_{1/\nu}(w_{\nu,j,h}).$$

When x varies almost everywhere in some $Y_{\nu,j}$, the function $\mathcal{T}_{1/\nu}(w_{\nu,j,h})(x,\cdot)$ belongs to $W_0^{1,p}(Y;\mathbb{R}^m)$, which implies $\mathcal{T}_{1/\nu}(w_{\nu,j,h}) \in L^p(Y_{\nu,j}; W_0^{1,p}(Y;\mathbb{R}^m))$. Consequently,

$$\begin{aligned}
&\int_{Y_{\nu,j}} f(\nu h x, \nabla u + \nabla w_{\nu,j,h})\,\mathrm{d}x\\
&\qquad \geq \inf\bigg\{\int_{Y_{\nu,j}\times Y} f(hy, \mathcal{T}_{1/\nu}(\nabla u)(x,y) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y :\\
&\qquad\qquad V \in L^p(Y_{\nu,j}; W_0^{1,p}(Y;\mathbb{R}^m))\bigg\}.
\end{aligned}$$

Since the sets $\{Y_{\nu,j}\}_{j\in J_\nu}$ are pairwise disjoint and $\Omega \subseteq \Omega_{1/\nu} = \cup_{j\in J_\nu} Y_{\nu,j}$ up to a null set, it is easy to verify that for every $h \in \mathbb{N}$

$$\begin{aligned}
&\sum_{j\in J_\nu} \inf\bigg\{\int_{Y_{\nu,j}\times Y} f(hy, \mathcal{T}_{1/\nu}(\nabla u)(x,y) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : \\
&\qquad V \in L^p(Y_{\nu,j}; W_0^{1,p}(Y;\mathbb{R}^m))\bigg\} \\
&\quad = \inf\bigg\{\int_{\Omega_{1/\nu}\times Y} f(hy, \mathcal{T}_{1/\nu}(\nabla u)(x,y) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : \\
&\qquad V \in L^p(\Omega; W_0^{1,p}(Y;\mathbb{R}^m))\bigg\} \\
&\quad \geq \inf\bigg\{\int_{\Omega\times Y} f(hy, \mathcal{T}_{1/\nu}(\nabla u)(x,y) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : \\
&\qquad V \in L^p(\Omega; W_0^{1,p}(Y;\mathbb{R}^m))\bigg\}.
\end{aligned}$$

Then (2.15), (2.15), and Proposition 2.3 provide

$$\begin{aligned}
&\liminf_{h\to+\infty} \int_\Omega f(\nu hx, \nabla u + \nabla v_h)\,\mathrm{d}x + |\Omega_{1/\nu} \setminus \Omega| \int_Y a\,\mathrm{d}y + M \int_{\Omega_{1/\nu}\setminus\Omega} |\nabla u|^p\,\mathrm{d}x \\
&\quad \geq \sum_{j\in J_\nu} \liminf_{h\to+\infty} \inf\bigg\{\int_{Y_{\nu,j}\times Y} f(hy, \mathcal{T}_{1/\nu}(\nabla u)(x,y) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : \\
&\qquad V \in L^p(Y_{\nu,j}; W_0^{1,p}(Y;\mathbb{R}^m))\bigg\} \\
&\quad = \lim_{h\to+\infty} \sum_{j\in J_\nu} \inf\bigg\{\int_{Y_{\nu,j}\times Y} f(hy, \mathcal{T}_{1/\nu}(\nabla u)(x,y) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : \\
&\qquad V \in L^p(Y_{\nu,j}; W_0^{1,p}(Y;\mathbb{R}^m))\bigg\} \\
&\quad \geq \lim_{h\to+\infty} \inf\bigg\{\int_{\Omega\times Y} f(hy, \mathcal{T}_{1/\nu}(\nabla u)(x,y) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : \\
&\qquad V \in L^p(\Omega; W_0^{1,p}(Y;\mathbb{R}^m))\bigg\}.
\end{aligned}$$

As ν goes to $+\infty$, $|\Omega_{1/\nu} \setminus \Omega| \to 0$. Therefore, to complete the proof, we just need to verify that

$$\begin{aligned}
&\liminf_{\nu\to+\infty} \lim_{h\to+\infty} \inf\bigg\{\int_{\Omega\times Y} f(hy, \mathcal{T}_{1/\nu}(\nabla u(x)) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : \\
&\qquad V \in L^p(\Omega; W_0^{1,p}(Y;\mathbb{R}^m))\bigg\} \\
&\quad \geq \liminf_{h\to+\infty} \inf\bigg\{\int_{\Omega\times Y} f(hy, \nabla u(x) + \nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : \\
&\qquad V \in L^p(\Omega; W_0^{1,p}(Y;\mathbb{R}^m))\bigg\}.
\end{aligned} \tag{2.16}$$

To do this, we first recall that, by (1.3), $\{\mathcal{T}_{1/\nu}(\nabla u)\}$ converges to ∇u in $L^p(\Omega\times Y;\mathbb{R}^{mn})$ as ν goes to $+\infty$. For ν and h in $\mathbb{N}$, using (0.3), (0.4), and the boundedness of $\{\int_Y a(hy)\,\mathrm{d}y\}$, it is straightforward to find a $K(u)\in\,]0,+\infty[$, not depending on ν and h, such that

$$\begin{aligned}&\inf\Big\{\int_{\Omega\times Y} f(hy,\mathcal{T}_{1/\nu}(\nabla u)(x,y)+\nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y:\\&\qquad V\in L^p(\Omega;W_0^{1,p}(Y;\mathbb{R}^m))\Big\}\\&\quad=\inf\Big\{\int_{\Omega\times Y} f(hy,\mathcal{T}_{1/\nu}(\nabla u)(x,y)+\nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y:\\&\qquad V\in L^p(\Omega;W_0^{1,p}(Y;\mathbb{R}^m)),\ \|\mathcal{T}_{1/\nu}(\nabla u)+\nabla_y V\|_{L^p(\Omega\times Y;\mathbb{R}^m)}\le K(u)\Big\}.\end{aligned}\tag{2.17}$$

If $V\in L^p(\Omega;W_0^{1,p}(Y;\mathbb{R}^m))$ satisfies $\|\mathcal{T}_{1/\nu}(\nabla u)+\nabla_y V\|_{L^p(\Omega\times Y;\mathbb{R}^m)}\le K(u)$, using (1.1) and Hölder's inequality, we have

$$\begin{aligned}&\int_{\Omega\times Y} f(hy,\mathcal{T}_{1/\nu}(\nabla u)(x,y)+\nabla_y V(x,y))\,\mathrm{d}x\\&\quad\ge\int_{\Omega\times Y} f(hy,\nabla u(x)+\nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y\\&\qquad-C\bigg\{\Big(|\Omega|\int_Y a(hy)\,\mathrm{d}y\Big)^{1-\frac{1}{p}}+\|\mathcal{T}_{1/\nu}(\nabla u)+\nabla_y V\|^{p-1}_{L^p(\Omega\times Y;\mathbb{R}^m)}\\&\qquad\quad+\|\nabla u+\nabla_y V\|^{p-1}_{L^p(\Omega\times Y;\mathbb{R}^m)}\bigg\}\|\mathcal{T}_{1/\nu}(\nabla u)-\nabla u\|_{L^p(\Omega\times Y;\mathbb{R}^m)}\\&\quad\ge\int_{\Omega\times Y} f(hy,\nabla u(x)+\nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y-C\bigg\{\Big(|\Omega|\int_Y a(hy)\,\mathrm{d}y\Big)^{1-\frac{1}{p}}+K(u)^{p-1}\\&\qquad+(\|\nabla u-\mathcal{T}_{1/\nu}(\nabla u)\|_{L^p(\Omega\times Y;\mathbb{R}^m)}+K(u))^{p-1}\bigg\}\|\mathcal{T}_{1/\nu}(\nabla u)-\nabla u\|_{L^p(\Omega\times Y;\mathbb{R}^m)}.\end{aligned}$$

This, together with (2.17), implies for every ν, h in $\mathbb{N}$ that

$$\begin{aligned}&\inf\Big\{\int_{\Omega\times Y} f(hy,\mathcal{T}_{1/\nu}(\nabla u)(x,y)+\nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y:\\&\qquad V\in L^p(\Omega;W_0^{1,p}(Y;\mathbb{R}^m))\Big\}\\&\quad\ge\inf\Big\{\int_{\Omega\times Y} f(hy,\nabla u(x)+\nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y:\\&\qquad V\in L^p(\Omega;W_0^{1,p}(Y;\mathbb{R}^m))\Big\}\\&\qquad-C\bigg\{\Big(|\Omega|\int_Y a(hy)\,\mathrm{d}y\Big)^{1-\frac{1}{p}}+K(u)^{p-1}\\&\qquad\quad+\|\nabla u-\mathcal{T}_{1/\nu}(\nabla u)\|^{p-1}_{L^p(\Omega\times Y;\mathbb{R}^m)}\bigg\}\|\mathcal{T}_{1/\nu}(\nabla u)-\nabla u\|_{L^p(\Omega\times Y;\mathbb{R}^m)}.\end{aligned}\tag{2.18}$$

From the boundedness of $\{\int_Y a(hy)\,\mathrm{d}y\}$, and the convergence of $\{\mathcal{T}_{1/\nu}(\nabla u)\}$ to ∇u in $L^p(\Omega\times Y;\mathbb{R}^{mn})$, (2.18) implies equality (2.16) by letting h and ν successively go to $+\infty$. ∎

Proof of Theorem 2.5. First, (1.1) follows by Proposition 1.5. Then the theorem is a direct consequence of Lemma 2.6, Remark 2.7, Lemma 2.9, Lemma 1.8, and Proposition 2.3. ∎

3. The convex case. Let $W^{1,p}_{\mathrm{per}}(Y;\mathbb{R}^m)$ denote the Banach space of Y-periodic functions in $W^{1,p}_{\mathrm{loc}}(\mathbb{R}^n;\mathbb{R}^m)$ endowed with the $W^{1,p}(Y;\mathbb{R}^m)$-norm.

Consider the case where hypothesis (0.2) is replaced by (0.6). We show below how the proof of the homogenization result in simplified by the periodic unfolding method.

In the convex case, the homogenized density is classically defined as

$$f_{\#}\colon z\in\mathbb{R}^{nm}\mapsto \inf\left\{\int_Y f(y,z+\nabla v)\,\mathrm{d}y : v\in W^{1,p}_{\mathrm{per}}(Y;\mathbb{R}^m)\right\}. \tag{3.1}$$

THEOREM 3.1. *Let f satisfy* (0.1) *and* (0.6). *Let $p\in\,]1,+\infty[$, and assume that* (0.3) *and* (0.4) *hold. Let $f_{\#}$ be defined by* (3.1). *Then, for every $\{\varepsilon_h\}\subseteq\,]0,+\infty[$ converging to* 0, *Ω in $\mathcal{A}_0$, and u in $W^{1,p}(\Omega;\mathbb{R}^m)$,*

$$\begin{aligned}
&\int_\Omega f_{\#}(\nabla u)\,\mathrm{d}x\\
&\quad=\inf\left\{\liminf_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)\,\mathrm{d}x : \{u_h\}\subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ u_h\to u \textit{ in } L^p(\Omega;\mathbb{R}^m)\right\}\\
&\quad=\inf\left\{\limsup_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)\,\mathrm{d}x : \{u_h\}\subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ u_h\to u \textit{ in } L^p(\Omega;\mathbb{R}^m)\right\}\\
&\quad=\inf\Bigg\{\liminf_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)\,\mathrm{d}x :\\
&\qquad\qquad \{u_h\}\subseteq u+W^{1,p}_0(\Omega;\mathbb{R}^m),\ u_h\to u \textit{ in } L^p(\Omega;\mathbb{R}^m)\Bigg\}\\
&\quad=\inf\Bigg\{\limsup_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)\,\mathrm{d}x :\\
&\qquad\qquad \{u_h\}\subseteq u+W^{1,p}_0(\Omega;\mathbb{R}^m),\ u_h\to u \textit{ in } L^p(\Omega;\mathbb{R}^m)\Bigg\}.
\end{aligned}$$

Proof. First of all, we recall that (0.6) trivially implies (0.2) so that, by Proposition 1.5, (1.1) holds.

Let $\{\varepsilon_h\}$, Ω, and u be as above. We claim that

$$\begin{aligned}
&\inf\Bigg\{\liminf_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)\,\mathrm{d}x :\\
&\qquad \{u_h\}\subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ u_h\rightharpoonup u \text{ in } W^{1,p}(\Omega;\mathbb{R}^m)\Bigg\}\\
&\qquad\quad\geq \inf\left\{\int_{\Omega\times Y} f(y,\nabla u(x)+\nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V\in L^p(\Omega;W^{1,p}_{\mathrm{per}}(Y;\mathbb{R}^m))\right\}.
\end{aligned} \tag{3.2}$$

Indeed, let $\{u_h\}\subseteq W^{1,p}(\Omega;\mathbb{R}^m)$ weakly converge to u in $W^{1,p}(\Omega;\mathbb{R}^m)$, and assume for simplicity that $\lim_{h\to+\infty}\int_\Omega f(\frac{x}{\varepsilon_h},\nabla u_h)\,\mathrm{d}x$ exists and is finite. Then, Theorem 1.7

implies that there exist a strictly increasing sequence of integers $\{h_k\}$ and some U in $L^p(\Omega; W^{1,p}_{\mathrm{per}}(Y;\mathbb{R}^m))$ with

$$\mathcal{T}_{\varepsilon_{h_k}}(\nabla u_{h_k}) \rightharpoonup \nabla u + \nabla_y U \text{ in } L^p(\Omega\times Y;\mathbb{R}^{nm}). \tag{3.3}$$

By (1.2) and the periodicity properties of f, for every h in $\mathbb{N}$, we have

$$\begin{aligned}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)\,\mathrm{d}x &\geq \int_{\Omega\times Y} f\Big(\frac{1}{\varepsilon_h}\Big(\varepsilon_h\Big[\frac{x}{\varepsilon_h}\Big]+\varepsilon_h y\Big),\mathcal{T}_{\varepsilon_h}(\nabla u_h)(x,y)\Big)\,\mathrm{d}x\,\mathrm{d}y\\ &= \int_{\Omega\times Y} f(y,\mathcal{T}_{\varepsilon_h}(\nabla u_h)(x,y))\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

Hence, using (3.3), (0.6), and Proposition 1.3 applied to $\Omega\times Y$, we conclude that

$$\begin{aligned}\lim_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)\,\mathrm{d}x &= \lim_{k\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_{h_k}},\nabla u_{h_k}\Big)\,\mathrm{d}x\\ &\geq \int_{\Omega\times Y} f(y,\nabla u(x)+\nabla_y U(x,y))\,\mathrm{d}x\,\mathrm{d}y\\ &\geq \inf\Big\{\int_{\Omega\times Y} f(y,\nabla u(x)+\nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V\in L^p(\Omega;W^{1,p}_{\mathrm{per}}(Y;\mathbb{R}^m))\Big\},\end{aligned}$$

from which (3.2) follows.

By (3.2), (0.4), and Rellich's theorem, it then follows that

$$\begin{aligned}&\inf\Big\{\liminf_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)\,\mathrm{d}x :\\ &\qquad \{u_h\}\subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ u_h\to u \text{ in } L^p(\Omega;\mathbb{R}^m)\Big\}\\ &\quad\geq \inf\Big\{\int_{\Omega\times Y} f(y,\nabla u(x)+\nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V\in L^p(\Omega;W^{1,p}_{\mathrm{per}}(Y;\mathbb{R}^m))\Big\}.\end{aligned}\tag{3.4}$$

On the other hand, Lemma 2.6 yields

$$\begin{aligned}&\inf\Big\{\limsup_{h\to+\infty}\int_\Omega f\Big(\frac{x}{\varepsilon_h},\nabla u_h\Big)\,\mathrm{d}x :\\ &\qquad \{u_h\}\subseteq W^{1,p}(\Omega;\mathbb{R}^m),\ u_h\to u \text{ in } L^p(\Omega;\mathbb{R}^m)\Big\}\\ &\quad\leq \inf\Big\{\int_{\Omega\times Y} f(y,\nabla u(x)+\nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V\in L^p(\Omega;W^{1,p}_{\mathrm{per}}(Y;\mathbb{R}^m))\Big\}.\end{aligned}\tag{3.5}$$

Eventually, we observe that the very same proof of Lemma 2.2 for $t=1$, with $W^{1,p}_0(Y;\mathbb{R}^m)$ replaced by $W^{1,p}_{\mathrm{per}}(Y;\mathbb{R}^m)$, and f_1 by $f_\#$, provides

$$\begin{aligned}&\inf\Big\{\int_{\Omega\times Y} f(y,\nabla u(x)+\nabla_y V(x,y))\,\mathrm{d}x\,\mathrm{d}y : V\in L^p(\Omega;W^{1,p}_{\mathrm{per}}(Y;\mathbb{R}^m))\Big\}\\ &\quad= \int_\Omega f_\#(\nabla u)\,\mathrm{d}x.\end{aligned}\tag{3.6}$$

Consequently, (3.4), (3.5), (3.6), and Lemma 1.8 prove the theorem. $\square$

REFERENCES

[AF] E. Acerbi and N. Fusco, *Semicontinuity problems in the calculus of variations*, Arch. Rational Mech. Anal., 86 (1984), pp. 125–145.

[A] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[ADH] T. Arbogast, J. Douglas, and U. Hornung, *Derivation of the double porosity model of single phase flow via homogenization theory*, SIAM J. Math. Anal., 21 (1990), pp. 823–836.

[At] H. Attouch, *Variational Convergence for Functions and Operators*, Pitman, London, 1984.

[BP] N. S. Bakhvalov and G. P. Panasenko, *Homogenization: Averaging Processes in Periodic Media*, Math. Appl. (Soviet Ser.) 36, Kluwer Academic Publishers, Dordrecht, 1989.

[Ba] J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1977), pp. 337–403.

[BLP] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, Stud. Math. Appl. 5, North-Holland, Amsterdam, 1978.

[B] A. Braides, *Homogenization of some almost periodic coercive functional*, Rend. Accad. Naz. Sci. XL Mem. Mat. (5), 9 (1985), pp. 313–321.

[BD] A. Braides and A. Defranceschi, *Homogenization of Multiple Integrals*, Oxford Lecture Ser. Math. Appl. 12, Oxford University Press, Oxford, 1998.

[Bu] G. Buttazzo, *Semicontinuity, Relaxation and Integral Representation in the Calculus of Variations*, Pitman Res. Notes Math. Ser. 207, Longman Scientific & Technical, Harlow, 1989.

[CDA] L. Carbone and R. De Arcangelis, *Unbounded Functionals in the Calculus of Variations. Representation, Relaxation, and Homogenization*, Chapman & Hall/CRC Monogr. Surv. Pure Appl. Math. 125, Chapman & Hall/CRC, Boca Raton, FL, 2001.

[CS] L. Carbone and C. Sbordone, *Some properties of* Γ*-limits of integral functionals*, Ann. Mat. Pura Appl. (4), 122 (1979), pp. 1–60.

[CV] C. Castaing and M. Valadier, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer-Verlag, Berlin, 1977.

[CDDA] D. Cioranescu, A. Damlamian, and R. De Arcangelis, *Homogenization of nonlinear integrals via the periodic unfolding method*, C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 77–82.

[CDG] D. Cioranescu, A. Damlamian, and G. Griso, *Periodic unfolding and homogenization*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 99–104.

[CD] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford Lecture Ser. Math. Appl. 17, Oxford University Press, Oxford, 1999.

[D] B. Dacorogna, *Direct Methods in the Calculus of Variations*, Appl. Math. Sci. 78, Springer-Verlag, Berlin, 1989.

[DM] G. Dal Maso, *An Introduction to* Γ*-Convergence*, Progr. Nonlinear Differential Equations Appl. 8, Birkhäuser-Verlag, Boston, 1993.

[DG] E. De Giorgi, *Sulla convergenza di alcune successioni d'integrali del tipo dell'area*, Rend. Mat. (6), 8 (1975), pp. 277–294.

[DGS] E. De Giorgi and S. Spagnolo, *Sulla convergenza degli integrali dell'energia per operatori ellittici del secondo ordine*, Boll. Un. Mat. Ital. (4), 8 (1973), pp. 391–411.

[DS] N. Dunford and J. T. Schwartz, *Linear Operators* I: *General Theory*, Wiley Classics Lib., Wiley Interscience, New York, 1988.

[F] N. Fusco, *On the convergence of integral functionals depending on vector-valued functions*, Ricerche Mat., 32 (1983), pp. 321–339.

[JKO] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.

[M1] P. Marcellini, *Periodic solutions and homogenization of non linear variational problems*, Ann. Mat. Pura Appl. (4), 117 (1978), pp. 139–152.

[M2] P. Marcellini, *Approximation of quasiconvex functions, and lower semicontinuity of multiple integrals*, Manuscripta Math., 51 (1985), pp. 1–28.

[Mo] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Grundlehren Math. Wiss. 130, Springer-Verlag, Berlin, 1966.

[Mu] S. Müller, *Homogenization of nonconvex integral functionals and cellular elastic materials*, Arch. Rational Mech. Anal., 99 (1987), pp. 189–212.

[MT] F. Murat and L. Tartar, *H-convergence*, in Topics in the Mathematical Modelling of Composite Materials, A. Cherkaev and R. Kohn, eds., Progr. Nonlinear Differential Equations Appl. 31, Birkhäuser-Verlag, Boston, 1997, pp. 21–44.

[N] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal., 20 (1989), pp. 608–623.

[SP] E. Sanchez-Palencia, *Nonhomogeneous Media and Vibration Theory*, Lecture Notes in Phys. 127, Springer-Verlag, Berlin, 1980.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1454–1464

YOUNG MEASURES ASSOCIATED WITH HOMOGENIZATION*

PABLO PEDREGAL†

Abstract. We address the issue of determining the Young measure associated with the sequence of gradients of the solutions to a highly oscillatory boundary value problem as in a homogenization setting. After proving a general result, we focus on the standard periodic case providing explicit formulae for the situation of a first-order laminate.

Key words. homogenization, Γ-convergence, Young measure

AMS subject classifications. 49J45, 74Q05

DOI. 10.1137/040614578

1. Introduction.

In [10], a treatment was given by which Γ-convergence [4] can be understood and examined, in some situations, by means of Young measures. See also [6], [7]. Here we would like to provide some hint not only about Γ-limits of functionals in typical cases and their minimizers, but also about the full Young measure associated with minimizers of a sequence of functionals. From this perspective, this work can be considered as a second part of [10].

To be more precise we will focus on a typical situation in homogenization. Suppose that the sequence $\{u_j\}$ is the sequence of solutions for

$$\text{div}\,[A_j(x)\nabla u_j(x)] = 0 \text{ in } \Omega, \quad u_j = u_0 \text{ on } \partial\Omega, \tag{1.1}$$

where $u_0 \in H^1(\Omega)$, Ω is a regular bounded domain, and we have the uniform bounds

$$0 < a \leq A_j(x) \leq b.$$

We have considered the homogeneous case for simplicity. Any source right-hand-side term can similarly be examined. Suppose that for a given relevant quantity ϕ we know that $\{\phi(\nabla u_j)\}$ weakly converges in $L^1(\Omega)$. How can we determine the limit

$$\lim_{j\to\infty}\int_\Omega \phi\left(\nabla u_j(x)\right)\,dx,$$

or, more generally,

$$\lim_{j\to\infty}\int_\Omega \xi(x)\phi\left(\nabla u_j(x)\right)\,dx, \tag{1.2}$$

for an arbitrary, measurable, uniformly bounded function $\xi(x)$ and continuous ϕ? This is the issue of the identification of arbitrary macroscopic quantities (see [5]).

We know too well that when the density ϕ is linear the answer is given by the solution of the corresponding homogenized equation. What can be said for a general nonlinear quantity if the sequence of gradients $\{\nabla u_j\}$ does not converge strongly in

*Received by the editors September 7, 2004; accepted for publication (in revised form) June 7, 2005; published electronically January 10, 2006. This work is supported by research project MTM2004-07114 (Spain).

http://www.siam.org/journals/sima/37-5/61457.html

†ETSI Industriales, Universidad de Castilla-La Mancha, 13071, Ciudad Real, Spain (pablo.pedregal@uclm.es).

$H^1(\Omega)$, which is the case of interest? This issue is nothing more than determining the Young measure corresponding to the sequence of gradients $\{\nabla u_j\}$ [3], [8], [11], [12]. Whenever $\psi(x,\lambda)$ is a Carathéodory integrand (measurable in x and continuous in λ), the Young measure associated with $\{\nabla u_j\}$ furnishes a representation of all limits of $\{\psi(x,\nabla u_j)\}$ whenever this sequence is weakly convergent in $L^1(\Omega)$, namely,

$$\lim_{j\to\infty}\int_\Omega \psi(x,\nabla u_j(x))\,dx = \int_\Omega\int_{\mathbf{R}^N}\psi(x,\lambda)\,d\nu_x(\lambda)\,dx$$

if $\nu=\{\nu_x\}_{x\in\Omega}$ is the Young measure generated by $\{\nabla u_j\}$.

In the one-dimensional situation, for some simple examples the functions u_j can even be computed explicitly (see [9]) so that, in particular, we can determine the associated Young measure. Having a sequence of functions is much more than having its underlying Young measure. The whole point is to determine and understand the limits (1.2) without calculating explicitly the solutions. Obviously, this is always the case in higher dimensions. Specifically, we will see the role played by the homogenized equation in determining such families of probability measures.

Since we will describe our results from the horizon of Γ-convergence and convergence of functionals, an additional comparison with variational problems may be of help. When we face a nonconvex variational problem such as

$$\text{minimize in } u: \quad \int_\Omega W(\nabla u(x))\,dx$$

subject to $u\in W^{1,p}(\Omega)$, $u-u_0\in W_0^{1,p}(\Omega)$, we need to examine relaxation at two different levels. At first we examine the convexified problem

$$\text{minimize in } u: \quad \int_\Omega CW(\nabla u(x))\,dx,$$

where CW designates the (quasi-)convexification of W under appropriate technical assumptions. This new problem admits minimizers which are the weak limits of minimizing sequences of the initial nonconvex problem. At the second relaxation level we explore the generalized variational problem in terms of Young measures,

$$\text{minimize in } \nu=\{\nu_x\}_{x\in\Omega}: \quad \int_\Omega\int_{\mathbf{M}} W(X)\,d\nu_x(X)\,dx,$$

where the admissible families of probability measures ν need to verify some main structural assumptions [8], [11] so that this new variational problem is truly a different way of looking at the same underlying variational principle. The optimal solutions of this generalized problem are the Young measures associated with the gradients of minimizing sequences of the initial problem, and their first moments are the minimizers of the first level of relaxation. Indeed, the identification of an optimal Young measure starts by determining its first moment (mean field) as the solution of this first level of relaxation or of the corresponding homogenized equation. Once we have this mean field, by looking at the second level of relaxation and, based on the knowledge of the first moment, we try to identify the optimal underlying Young measure, the "fluctuations" around the mean field.

We would like to follow this same strategy in the context of homogenization, where solutions of homogenized equations provide first moments of associated Young measures. As indicated, this would be like a first level of relaxation. The second step would require the full Young measure which somehow needs to be recovered from its first moment. In fact, the identification of the Young measure is somehow encoded in the process of computing the integrand for the convexified functional or for the Γ-limit, so that it typically involves a careful re-examination of how this integrand was computed or defined.

Despite our strategy described above, a more general and abstract treatment may be possible in which relevant Young measures are sought directly without any reference to its first moment or knowledge of it. See [2]. Our ideas are obviously very closely connected and motivated by homogenization techniques. In particular, we would like to refer to [1] and [6].

In the context of Γ-convergence, and without being precise at this stage, if I_j Γ-converges to I and u is a given field, we say that $\nu = \{\nu_x\}_{x\in\Omega}$ is an optimal Young measure associated with u if there is a sequence $\{u_j\}$ generating ν such that $I_j(u_j) \to I(u)$. When u is a minimizer for I, ν will be the Young measure corresponding to a sequence of minimizers for I_j and, hence, it provides macroscopic quantities.

Although more general results are possible, our main contribution here focuses on the standard periodic homogenization setting

$$A_j(x) = A(jx), \quad 0 < a \le A(y) \le b, \ y \in Q, \quad A \text{ is } Q\text{-periodic},$$

where Q is the unit cube in $\mathbf{R}^2$. We restrict our attention to the two-dimensional situation. Here Q replaces the domain Ω above as we concentrate on a periodic situation. For a given, Q-periodic $u_0 \in H^1(Q)$, let $u_j \in H^1(Q)$ be the unique, Q-periodic, weak solution of (1.1). Let

$$\xi : \mathbf{R}^2 \times \mathbf{R}^2 \to \mathbf{R}, \quad \xi(\lambda, \rho),$$

Q-periodic in the λ-variable, be the unique, Q-periodic, weak solution in $H^1(Q)$ of the cell problem for fixed ρ,

$$\operatorname{div}_\lambda \left[A(\lambda)\left(\rho + \nabla_\lambda \xi(\lambda, \rho)\right)\right] = 0. \tag{1.3}$$

THEOREM 1.1. *Let $\nu = \{\nu_x\}_{x\in\Omega}$ be the Young measure associated with $\{\nabla u_j\}$. Then for any continuous ϕ we have*

$$\langle \phi, \nu_x \rangle = \int_Q \phi\left(\nabla u(x) + \nabla_\lambda \xi(\lambda, \nabla u(x))\right) d\lambda,$$

where u is the solution of the homogenized equation

$$\operatorname{div}\left[\frac{\partial \psi}{\partial \rho}\left(\nabla u(x)\right)\right] = 0 \text{ in } Q, \quad u = u_0 \text{ on } \partial Q,$$
$$\psi(\rho) = \int_Q \frac{A(\lambda)}{2} \left|\rho + \nabla_\lambda \xi(\lambda, \rho)\right|^2 d\lambda.$$

Notice that because of (1.3) and the periodicity of $\xi(\lambda, \rho)$, we also have

$$\frac{\partial \psi}{\partial \rho}(\rho) = \int_Q A(\lambda)(\rho + \nabla_\lambda \xi(\lambda, \rho))\, d\lambda.$$

When $\{\psi(x, \nabla u_j(x))\}$ converges weakly in $L^1(\Omega)$, Theorem 1.1 implies

$$\lim_{j\to\infty} \int_Q \psi(x, \nabla u_j(x))\, dx = \int_Q \int_Q \psi\left(x, \nabla u(x) + \nabla_\lambda \xi(\lambda, \nabla u(x))\right) d\lambda\, dx.$$

Theorem 1.1 is a direct consequence of results proved in sections 2 and 3, namely, the main result, Theorem 2.2, and Corollary 3.2.

More explicit formulae can be given for more specific examples. For a typical first-order laminate, where we take

$$A(y) = \chi(y \cdot n)a + (1 - \chi(y \cdot n))b, \quad 0 < a < b, \ |n| = 1,$$

χ being the characteristic function of $(0, t)$ in $(0, 1)$ extended by periodicity and n the unit normal to the layers, we have

$$\psi(\rho) = (ta + (1-t)b) |\rho|^2 - \frac{(b-a)^2 t(1-t)}{(1-t)a + tb} (\rho \cdot n)^2$$

and

$$\lim_{j\to\infty} \int_Q \phi(\nabla u_j(x))\, dx = \int_Q \left[t\phi\left(F(\nabla u(x))\right) + (1-t)\phi\left(G(\nabla u(x))\right)\right] dx,$$

where u is the solution of the homogenized equation

$$\operatorname{div} \left[\frac{\partial \psi}{\partial \rho}(\nabla u(x))\right] = 0 \text{ in } Q, \quad u = u_0 \text{ on } \partial Q,$$

and $F(\rho)$, $G(\rho)$ are the linear mappings with matrices

$$\mathbf{1} + \frac{(1-t)(b-a)}{(1-t)a + tb} n \otimes n, \quad \mathbf{1} - \frac{t(b-a)}{(1-t)a + tb} n \otimes n,$$

respectively, where $\mathbf{1}$ is the identity matrix.

Corollary 1.2. *For any Carathéodory integrand $\psi(x, \lambda)$ such that $\{\psi(x, \nabla u_j(x))\}$ weakly converges in $L^1(Q)$, we have*

$$\begin{aligned}
&\lim_{j\to\infty} \int_Q \psi(x, \nabla u_j(x))\, dx \\
&= \int_\Omega t\psi\left(x, \left(\mathbf{1} + \frac{(1-t)(b-a)}{(1-t)a + tb} n \otimes n\right) \nabla u(x)\right) \\
&\qquad + (1-t)\psi\left(x, \left(\mathbf{1} - \frac{t(b-a)}{(1-t)a + tb} n \otimes n\right) \nabla u(x)\right) dx.
\end{aligned}$$

As an even more specific example, if $\{|\nabla u_j|^2\}$ weakly converges in $L^1(Q)$, we should have

$$\lim_{j\to\infty} \int_Q |\nabla u_j(x)|^2\, dx = \int_Q \left(|\nabla u(x)|^2 + \frac{t(1-t)(b-a)^2}{((1-t)a + tb)^2} (\nabla u(x) \cdot n)^2\right) dx,$$

where u is the solution of the corresponding homogenized equation. Similar formulae are valid for higher moments.

In section 2, as a preliminary step we treat the nongradient case, while in section 3 we concentrate on the gradient case. An equivalent analysis for situations more general than the typical periodic case investigated here is possible, for instance for the almost periodic case. This, however, would depend on a preliminary study analogous to the one in [10] for more general situations.

2. Case without derivatives. Let us first consider the case without derivatives as a preliminary step. It can serve as a training ground for the main ideas we will have to implement later for the case with gradients,

$$I_j(v) = \int_\Omega W(A_j(x), v(x))\, dx,$$

where the assumptions on

$$W(\lambda, \rho) : \mathbf{R}^m \times \mathbf{R}^d \to \mathbf{R}$$

and the sequence $\{A_j\}$ are as follows:

1. $\{A_j\}$ is a weakly convergent sequence in $L^q(\Omega)$ for some $q > 1$;
2. W is uniformly coercive in ρ for all j with an exponent $p > 1$ and is uniformly bounded from above by the same power so that

$$c\,(|\rho|^p - 1) \leq W(A_j(x), \rho) \leq C\,(|\rho|^p + 1)$$

   for some $C \geq c > 0$, all j, and a.e. $x \in \Omega$. Under this hypothesis, every sequence $\{u_j\}$ such that $\{I_j(u_j)\}$ is bounded from above will be bounded in $L^p(\Omega)$ and thus, possibly for a subsequence, it will converge weakly to some $u \in L^p(\Omega)$.
3. W is uniformly continuous in λ as indicated earlier,

$$|W(\lambda_1, \rho) - W(\lambda_2, \rho)| \leq w(|\lambda_1 - \lambda_2|)\, |\rho|^p\,,$$

   where w is continuous and $w(0) = 0$.
4. W is strictly convex in ρ for every fixed λ. This assumption can be dispensed with as we will see in the proof of the theorem below, but we will retain it for simplicity.

Let $u \in L^p(\Omega)$ be given, and let $\sigma = \{\sigma_x\}_{x\in\Omega}$ be the Young measure generated by the sequence $\{A_j\}$.

DEFINITION 2.1. *A family of probability measures $\nu = \{\nu_x\}_{x\in\Omega}$ is an optimal Young measure associated with $u \in L^p(\Omega)$ if it is the Young measure associated with a sequence of functions $\{u_j\}$ such that the limit*

$$\lim_{j\to\infty} \int_\Omega W(A_j(x), u_j(x))\, dx$$

is precisely the value of the Γ-limit at u taken with respect to the weak topology in $L^p(\Omega)$.

Our main task consists of determining an optimal Young measure associated with any given $u \in L^p(\Omega)$.

THEOREM 2.2. *There exists a map*

$$\varphi_0 : \Omega \times \mathbf{R}^m \times \mathbf{R}^d \to \mathbf{R}^d$$

such that the family of measures supported in $\mathbf{R}^d$ and determined by

$$\langle \phi, \nu_x \rangle = \int_{\mathbf{R}^m} \phi\left(\varphi_0(x, \lambda, u(x))\right)\, d\sigma_x(\lambda)$$

for any continuous ϕ is an optimal Young measure associated with u.

Proof. The proof is nothing more than a careful analysis of how the Γ-limit of $\{I_j\}$ was obtained in [10] through Young measures. We recall here some of those main ideas for the convenience of the readers.

In the nongradient case, the integrand ψ for the Γ-limit is given by

$$\psi(x,\rho) = \min_{\varphi} \left\{ \int_{\mathbf{R}^m} CW(\lambda, \varphi(\lambda))\, d\sigma_x(\lambda) : \rho = \int_{\mathbf{R}^m} \varphi(\lambda)\, d\sigma_x(\lambda) \right\}.$$

CW stands for the convexification with respect to λ of W. Let us pretend we do not have the hypothesis of convexity on W. Since, under our assumptions, there are optimal solutions for this optimization problem, let us put $\varphi_0(x,\lambda,\rho)$ for one such optimal solution so that

$$\begin{aligned} \rho &= \int_{\mathbf{R}^m} \varphi_0(x,\lambda,\rho)\, d\sigma_x(\lambda), \\ \psi(x,\rho) &= \int_{\mathbf{R}^m} CW(\lambda, \varphi_0(x,\lambda,\rho))\, d\sigma_x(\lambda). \end{aligned} \tag{2.1}$$

Further, by convexity arguments, find a family of probability measures $\mu^{(0)} = \{\mu^{(0)}_{\lambda,x}\}$ such that

$$\begin{aligned} \varphi_0(x,\lambda,u(x)) &= \int_{\mathbf{R}^d} \rho\, d\mu^{(0)}_{\lambda,x}(\rho), \\ CW\left(\lambda, \varphi_0(x,\lambda,u(x))\right) &= \int_{\mathbf{R}^d} W(\lambda,\rho)\, d\mu^{(0)}_{\lambda,x}(\rho). \end{aligned} \tag{2.2}$$

Write

$$\Lambda = \{\Lambda_x\}_{x\in\Omega}, \quad \Lambda_x = \mu^{(0)}_{\lambda,x}(\rho) \otimes \sigma_x(\lambda).$$

Then [10] Λ is the joint Young measure generated by $\{(A_j, u_j)\}$ for some sequence $\{u_j\}$ so that

$$\lim_{j\to\infty} \int_\Omega W(A_j(x), u_j(x))\, dx = \int_\Omega \psi(x, u(x))\, dx.$$

The main idea now consists of projecting Λ_x onto the ρ-variable to obtain the Young measure $\nu = \{\nu_x\}_{x\in\Omega}$ corresponding to $\{u_j\}$,

$$\Lambda_x = \pi_{\rho,x}(\lambda) \otimes \nu_x(\rho),$$

for a certain family of probability measures $\pi_{\rho,x}$. Notice how Λ_x, for a.e. $x \in \Omega$, is a probability measure in the product space $\mathbf{R}^m \times \mathbf{R}^d$ that has been obtained and defined through the optimality property written in (2.1) and (2.2). Its projection onto $\mathbf{R}^m$ is $\sigma = \{\sigma_x\}_{x\in\Omega}$ which is the Young measure associated with $\{A_j\}$. Then, we are decomposing such probability measure Λ_x projecting onto the other component $\mathbf{R}^d$ to obtain a family $\nu = \{\nu_x\}_{x\in\Omega}$, which, due to that optimality property coming from $\mu^{(0)}_{\lambda,x}$ (see [10]), turns out to be an optimal Young measure associated with u. As indicated, this is the main idea of this paper and it completes the analysis in [10].

Indeed, by (2.1) and (2.2), we have

$$\begin{aligned}\psi(x,u(x)) &= \int_{\mathbf{R}^m} CW(x,\varphi_0(x,\lambda,u(x)))\,d\sigma_x(\lambda)\\ &= \int_{\mathbf{R}^m}\int_{\mathbf{R}^d} W(\lambda,\rho)\,d\mu^{(0)}_{\lambda,x}(\rho)\,d\sigma_x(\lambda)\\ &= \int_{\mathbf{R}^d}\int_{\mathbf{R}^m} W(\lambda,\rho)\,d\pi_{\rho,x}(\lambda)\,d\nu_x(\rho).\end{aligned}$$

Consequently,

$$\int_\Omega \psi(x,u(x))\,dx = \int_\Omega\int_{\mathbf{R}^d\times\mathbf{R}^m} W(\lambda,\rho)\,d\Lambda_x(\lambda,\rho)\,dx = \lim_{j\to\infty}\int_\Omega W(A_j(x),u_j(x))\,dx.$$

For any continuous ϕ

$$\langle\phi,\nu_x\rangle = \int_{\mathbf{R}^d\times\mathbf{R}^m}\phi(\rho)\,d\mu^{(0)}_{\lambda,x}(\rho)\,d\sigma_x(\lambda).$$

But notice that the strict convexity of W with respect to ρ implies that in fact

$$\mu^{(0)}_{\lambda,x}(\rho) = \delta_{\varphi_0(x,\lambda,u(x))}$$

and therefore

$$\langle\phi,\nu_x\rangle = \int_{\mathbf{R}^m}\phi\left(\varphi_0(x,\lambda,u(x))\right)\,d\sigma_x(\lambda).$$

This is the statement of the theorem. ∎

What is interesting is that this map φ_0 can be explicitly computed in some standard cases.

Take first for $0<a<b$, $p>1$, and $0<t<1$

$$\sigma_x = t\delta_a + (1-t)\delta_b, \quad W(\lambda,\rho) = \lambda\,|\rho|^p\,.$$

As it was shown in [10], the integrand for the Γ-limit is obtained from the mathematical programming problem

$$\psi(\rho) = \min_\varphi\left\{tW(a,\varphi(a)) + (1-t)W(b,\varphi(b)) : \rho = t\varphi(a) + (1-t)\varphi(b)\right\},$$

where if we set $F=\varphi(a)$, $G=\varphi(b)$, and we put $F(\rho)$ and $G(\rho)$ for the optimal choices for F and G so that

$$\psi(\rho) = tW(a,F(\rho)) + (1-t)W(b,G(\rho)), \quad \rho = tF(\rho) + (1-t)G(\rho),$$

then the optimal Young measure associated with a given $u\in L^p(\Omega)$ comes from the projection of

$$t\delta_{F(u(x))}\otimes\delta_a + (1-t)\delta_{G(u(x))}\otimes\delta_b$$

onto the first variable, i.e.,

$$t\delta_{F(u(x))} + (1-t)\delta_{G(u(x))}.$$

Explicit expressions for $F(\rho)$ and $G(\rho)$ are found in an elementary way.

Consider as a second example a sequence $\{A_j\}$ whose Young measure is

$$d\lambda|_{[1,2]}$$

independent of $x \in \Omega$. Then the integrand $\psi(\rho)$ is found through the minimum

$$\psi(\rho) = \min_{\varphi} \left\{ \int_1^2 W(\lambda, \varphi(\lambda))\, d\lambda : \rho = \int_1^2 \varphi(\lambda)\, d\lambda \right\}.$$

The optimal solution of this zero-order variational problem should be a solution of

$$\frac{\partial W}{\partial \rho}(\lambda, \varphi(\lambda)) = \gamma$$

if W is strictly convex with respect to ρ, where γ is a multiplier determined so that the integral constraint is fulfilled. As an explicit example take

$$W(\lambda, \rho) = \frac{1}{2(\lambda - 1)(2 - \lambda)} |\rho|^2 .$$

Then it is elementary to check that the optimal mapping φ_0 is given by

$$\varphi_0(\lambda, \rho) = \rho\, 6(\lambda - 1)(2 - \lambda),$$

and then the optimal Young measure associated with $u \in L^2(\Omega)$ is given by

$$\nu = \{\nu_x\}_{x \in \Omega}, \quad \langle \phi, \nu_x \rangle = \int_1^2 \phi\left(6(\lambda - 1)(2 - \lambda)\, u(x)\right) d\lambda.$$

We see that this time the projection is not one-to-one since φ_0 itself is not one-to-one. There is some overlapping effect on the oscillations from the viewpoint of the u variable.

3. Gradient case. We now consider the sequence of functionals

$$I_j(v) = \int_\Omega W(A_j(x), \nabla v(x))\, dx,$$

where W and the sequence $\{A_j\}$ are as in section 2. The difficulties attached to the same analysis have been emphasized in [10]. To be able to provide a complete analysis of the gradient situation, we need to add an important structural assumption on the sequence $\{A_j\}$. This is the AGP (average gradient property) which, in plain words, amounts to the fact that "averages of gradients over level sets of a_j are gradients themselves." See [10] for a much more precise formulation. Here we would like to show a general result similar to Theorem 2.2 for the gradient case based on this main assumption, and then apply it to the typical situation of periodic homogenization. The abstract theorem is formally the same as in the nongradient case. The difference is hidden in the much stronger AGP condition. In the case of periodic homogenization, under our technical assumptions, the AGP requirement has been seen to be fulfilled in [10] so that we do not need to bother here about it. To avoid unnecessary repetition, we refer our readers to this work for a complete discussion.

THEOREM 3.1. *Under the general assumptions on $\{A_j\}$ and W indicated in the preceding section and the AGP condition on the sequence $\{A_j\}$, there exists a map*

$$\varphi_0 : \Omega \times \mathbf{R}^m \times \mathbf{R}^d \to \mathbf{R}^d$$

such that the family of measures supported in $\mathbf{R}^d$ *and determined by*

$$\langle \phi, \nu_x \rangle = \int_{\mathbf{R}^m} \phi\left(\varphi_0(x, \lambda, \nabla u(x))\right) \, d\sigma_x(\lambda),$$

for any continuous ϕ, *is an optimal Young measure associated with* u.

The proof of this result is exactly the same as the proof of Theorem 2.2. The precise place where the gradient constraint enters is the definition of the integrand for the Γ-limit which again is

$$\psi(x, \rho) = \min_{\varphi} \left\{ \int_{\mathbf{R}^m} CW(\lambda, \varphi(\lambda)) \, d\sigma_x(\lambda) : \rho = \int_{\mathbf{R}^m} \varphi(\lambda) \, d\sigma_x(\lambda) \right\}.$$

In the gradient case, this is so because of the AGP condition assumed on the sequence $\{a_j\}$ (Lemma 9 in [10]). In the nongradient case, there is no need for this structural assumption, as there is "no gradient constraint."

We next focus on the periodic homogenization situation, where

$$A_j(x) = A(jx)$$

and A is Q-periodic. Recall that Q is the unit cube in $\mathbf{R}^2$. A direct corollary for this situation follows.

COROLLARY 3.2. *There exists a mapping*

$$\xi : \mathbf{R}^2 \times \mathbf{R}^2 \to \mathbf{R}, \quad \xi(\lambda, \rho)$$

Q*-periodic in the first variable* λ *such that an optimal Young measure associated with any given* $u \in W^{1,p}(\Omega)$ *is given by*

$$\nu = \{\nu_x\}_{x \in \Omega}, \quad \langle \phi, \nu_x \rangle = \int_Q \phi\left(\nabla u(x) + \nabla_\lambda \xi(\lambda, \nabla u(x))\right) \, d\lambda$$

for any continuous ϕ.

This mapping ξ is given as the solution of the typical cell problem in homogenization

$$\int_Q W\left(\lambda, \rho + \nabla_\lambda \xi(\lambda, \rho)\right) \, d\lambda = \min_{\zeta} \int_Q W\left(\lambda, \rho + \nabla_\lambda \zeta(\lambda, \rho)\right) \, d\lambda.$$

In the typical quadratic case, where we put

$$W(\lambda, \rho) = \frac{A(\lambda)}{2} |\rho|^2,$$

the integrand for the Γ-limit is given by

$$\psi(\rho) = \int_Q \frac{A(\lambda)}{2} |\rho + \nabla_\lambda \xi(\lambda, \rho)|^2 \, d\lambda.$$

In this way, if $u_j \in H^1(Q)$ is the solution of

$$\operatorname{div} [A(jx) \nabla u_j(x)] = 0 \text{ in } Q, \quad u = u_0 \text{ on } \partial Q,$$

then for any continuous ϕ for which $\{\phi(\nabla u_j)\}$ weakly converges in $L^1(Q)$, we have

$$\lim_{j \to \infty} \int_Q \phi(\nabla u_j(x)) \, dx = \int_Q \int_Q \phi\left(\nabla u(x) + \nabla_\lambda \xi(\lambda, \nabla u(x))\right) \, d\lambda \, dx,$$

where u is the solution of the homogenized equation

$$\operatorname{div}\left[\frac{\partial\psi}{\partial\rho}(\nabla u(x))\right] = 0 \text{ in } Q, \quad u = u_0 \text{ on } \partial Q.$$

In the case of a single, first-order laminate, more explicit calculations can be performed. Indeed if we put

$$A(x) = \chi(x\cdot n)a + (1-\chi(x\cdot n))\,b, \qquad 0<a<b, \quad |n|=1,$$

where χ is the characteristic function of the interval $(0,t)$ over $(0,1)$ extended by periodicity and the unit vector n indicates the normal direction to the layers, then the optimal mapping $\nabla_\lambda\xi(\lambda,\rho)$ can be computed explicitly. This calculation can be found in [10], namely,

$$\psi(\rho) = \min_{F,G\in\mathbf{R}^2}\left\{t\frac{a}{2}\,|F|^2 + (1-t)\frac{b}{2}\,|G|^2 : \rho = tF + (1-t)G, (F-G)\cdot Tn = 0\right\}.$$

T is the counterclockwise, $\pi/2$ rotation in the plane. Then

$$\psi(\rho) = (ta + (1-t)b)\,|\rho|^2 - \frac{(b-a)^2 t(1-t)}{(1-t)a + tb}\,(\rho\cdot n)^2$$

and the optimal pair $F(\rho)$ and $G(\rho)$ furnishing this value can also be written explicitly (see the introduction). Then if

$$\operatorname{div}\,[A(jx)\nabla u_j(x)] = 0 \text{ in } Q, \quad u_j = u_0 \text{ on } \partial Q,$$

and if for any given continuous ϕ we have that $\{\phi(\nabla u_j)\}$ converges in $L^1(Q)$, we have

$$\lim_{j\to\infty}\int_Q \phi(\nabla u_j(x))\,dx = \int_Q \left[t\phi\left(F(\nabla u(x))\right) + (1-t)\phi\left(G(\nabla u(x))\right)\right]dx,$$

where u is the solution of the homogenized equation

$$\operatorname{div}\left[\frac{\partial\psi}{\partial\rho}(\nabla u(x))\right] = 0 \text{ in } Q, \quad u = u_0 \text{ on } \partial Q.$$

More general examples can also be treated.

Acknowledgments. The author would like to express his gratitude to several referees for their valuable comments in improving the original manuscript.

REFERENCES

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[2] O. Anza Hafsa, J. P. Mandallena, and G. Michaille, *Homogenization of periodic nonconvex integral functionals in terms of Young measures*, ESAIM Control Optim. Calc. Var., to appear.

[3] E. J. Balder, *Lectures on Young Measures*, Cahiers de Mathématiques de la Décision No. 9517, CEREMADE, Université Paris IX, Paris, 1995.

[4] G. Dal Maso, *Introduction to* Γ*-convergence*, Birkhäuser Boston, Boston, MA, 1993.

[5] R. Lipton, *Assessment of the local stress state through macroscopic variables*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Engrg. Sci., 361 (2003), pp. 921–946.

[6] M. L. Mascarenhas and A. M. Toader, *Scale convergence in homogenization*, Numer. Funct. Anal. Optim., 22 (2001), pp. 127–158.

[7] G. Michaille and M. Valadier, *Young measures generated by a class of integrands: A narrow epicontinuity and applications to homogenization*, J. Math. Pures Appl., 81 (2002), pp. 1277–1312.

[8] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser, Basel, 1997.

[9] P. Pedregal, *Optimization, relaxation and Young measures*, Bull. AMS, 36 (1999), pp. 27–58.

[10] P. Pedregal, Γ*-convergence through Young measures*, SIAM J. Math. Anal., 36 (2004), pp. 423–440.

[11] T. Roubicek, *Relaxation in Optimization Theory and Variational Calculus*, W. De Gruyter, Berlin, 1997.

[12] M. Valadier, *Young measures*, in Methods of Nonconvex Analysis, Lecture Notes in Math. 1446, Springer, Berlin, 1990, pp. 152–188.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1465–1481

EXISTENCE OF EIGENVALUES OF A LINEAR OPERATOR PENCIL IN A CURVED WAVEGUIDE—LOCALIZED SHELF WAVES ON A CURVED COAST*

E. R. JOHNSON†, MICHAEL LEVITIN‡, AND LEONID PARNOVSKI†

Abstract. The question of the existence of nonpropagating, trapped continental shelf waves (CSWs) along curved coasts reduces mathematically to a spectral problem for a self-adjoint operator pencil in a curved strip. Using methods developed for the waveguide trapped mode problem, we show that such CSWs exist for a wide class of coast curvature and depth profiles.

Key words. continental shelf waves, curved coasts, trapped modes, essential spectrum, operator pencil

AMS subject classifications. 35P05, 35P15, 86A05, 76U05

DOI. 10.1137/040615936

1. Introduction. Measurements of velocity fields along the coasts of oceans throughout the world show that much of the fluid energy is contained in motions with periods of a few days or longer. The comparison of measurements at different places along the same coast shows that in general these low-frequency disturbances propagate along coasts with shallow water to the right in the northern hemisphere and to the left in the southern hemisphere. These waves have come to be known as continental shelf waves (CSWs). The purpose of the present paper is to demonstrate, using the most straightforward model possible, the possibility of nonpropagating, trapped CSWs along curved coasts. The existence of such nonpropagating modes would be significant as they would tend to be forced by atmospheric weather systems, which have similar periods of a few days, similar horizontal extent, and a reasonably broad spectrum in space and time. Areas where such modes were trapped would thus appear to be likely to show higher than normal energy in the low-frequency horizontal velocity field.

The simplest models for CSWs take the coastal oceans to be inviscid and of constant density. Both these assumptions might be expected to fail in various regions such as when strong currents pass sharp capes or when the coastal flow is strongly stratified. However, for small-amplitude CSWs in quiescent flow along smooth coasts, viscous separation is negligible. Similarly most disturbance energy is concentrated in the modes with the least vertical structure, which are well described by the constant density model [LBMy]. The governing equations are then simply the rotating incompressible Euler equations. Further, coastal flows are shallow in the sense that the ratio of depth to typical horizontal scale is small. Expanding the rotating incompressible Euler equations in powers of this ratio and retaining only the leading order terms

*Received by the editors September 29, 2004; accepted for publication (in revised form) June 7, 2005; published electronically January 12, 2006.
http://www.siam.org/journals/sima/37-5/61593.html

†Department of Mathematics, University College London, Gower Street, London WC1E 6BT, UK (e.johnson@ucl.ac.uk, leonid@math.ucl.ac.uk). The research of the third author was partially supported by the EPSRC Spectral Theory Network.

‡Maxwell Institute for Mathematical Sciences, Department of Mathematics, Heriot-Watt University, Riccarton, Edinburgh EH14 4AS, UK (m.levitin@ma.hw.ac.uk). The research of this author was partially supported by the EPSRC Spectral Theory Network.

gives the rotating shallow water equations [Pe]:

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \mathbf{grad}\,\mathbf{u} - 2\Omega\mathbf{k} \times \mathbf{u} = -g\,\mathbf{grad}\,\widetilde{H}, \tag{1.1}$$

$$\frac{\partial \widetilde{H}}{\partial t} + \mathbf{div}[(\widetilde{H} + H)\mathbf{u}] = 0. \tag{1.2}$$

Here **div** and **grad** are taken with respect to horizontal coordinates (x, y) in a frame fixed to the rotating Earth, $\mathbf{k}$ is a vertical unit vector, $\mathbf{u}(x, y, t)$ is the horizontal velocity (with components $\mathbf{u} = (u, v)$), Ω is the (locally constant) vertical component of the Earth's rotation, g is the gravitational acceleration, $\widetilde{H}(x, y, t)$ is the vertical displacement of the free surface, and $H(x, y)$ is the local undisturbed fluid depth.

System (1.1), (1.2) admits waves of two types, denoted Class 1 and Class 2 by [La]. Class 1 waves are fast high-frequency waves, the rotation-modified form of the usual free surface water waves, although here present only as long, nondispersive waves with speeds of order $\sqrt{gH}$. Class 2 waves are slower, low-frequency waves that vanish in the absence of depth change or in the absence of rotation. It is the Class 2 waves that give CSWs. They have little signature in the vertical height field $\widetilde{H}(x, y, t)$ and are observed through their associated horizontal velocity fields [Ha]. The Class 1 waves can be removed from (1.1), (1.2) by considering the "rigid-lid" limit, where the external Rossby radius $\sqrt{gH}/2\Omega$ (which gives the relaxation distance of the free surface) is large compared to the horizontal scale of the motion. This is perhaps the most accurate of the approximations noted here, causing the time-dependent term to vanish from (1.2) and the right side of (1.1) to become a simple pressure gradient.

For small amplitude waves the nonlinear terms in (1.1), (1.2) are negligible, and cross-differentiating gives

$$\frac{\partial \zeta}{\partial t} + 2\Omega\,\mathbf{div}\,\mathbf{u} = 0, \tag{1.3}$$

$$\mathbf{div}(H\mathbf{u}) = 0, \tag{1.4}$$

where $\zeta = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}$ is the vertical component of relative vorticity. Equation (1.4) is satisfied by introducing the volume flux streamfunction defined through

$$Hu = -\frac{\partial \psi}{\partial y}, \qquad Hv = \frac{\partial \psi}{\partial x}, \tag{1.5}$$

allowing (1.3) to be written as the single equation

$$\mathbf{div}\left(\frac{1}{H}\,\mathbf{grad}\,\frac{\partial \psi}{\partial t}\right) + 2\Omega\mathbf{k} \cdot \mathbf{grad}\,\psi \times \mathbf{grad}\left(\frac{1}{H}\right) = 0. \tag{1.6}$$

Equation (1.6) is generally described as the topographic Rossby wave equation or the equation for barotropic CSWs. Many solutions have been presented for straight coasts, where the coast lies along $y = 0$ (say) and the depth H is a function of y alone (described as rectilinear topography here) [LBMy]. These have shown excellent agreement with observations of CSWs, as in [Ha]. There has been far less discussion of nonrectilinear geometries, where either the coast or the depth profile or both are not functions of a single coordinate. Yet interesting results appear. The papers [StHu1], [StHu2] present extensive numerical integrations of a low-order spectral model of a rectangular lake with idealized topography. For their chosen depth profiles normal modes can be divided into two types: *basin-wide modes* which extend throughout the lake and localized *bay modes*. These bay modes correspond to the high-frequency modes found in a finite-element model of Lake Lugarno by [Tr] and observed by [StHuSaTrZa]. The papers [Jo2], [StJo1], [StJo2] give a variational formulation and describe simplified quasi-analytical models that admit localized trapped bay modes. However the geometry changes in these models are large, with the sloping lower boundary terminating abruptly where it strikes a coastal wall. Further [Jo1] notes that (1.6) is invariant under conformal mappings and so any geometry that can be mapped conformally to a rectilinear shelf cannot support trapped modes. The question thus arises as to whether it is only for the most extreme topographic changes that shelf waves can be trapped or whether trapping can occur on smoothly varying shelves. The purpose of this paper is to provide the answer: trapped modes can exist on smoothly curving coasts.

The geometry considered here is that of a shelf of finite width lying along an impermeable coast. Thus sufficiently far from the coast the undisturbed fluid depth becomes the constant depth of the open ocean. It is shown in [Jo3] that at the shelf-ocean boundary of finite-width rectilinear shelves the tangential velocity component u vanishes for waves sufficiently long compared to the shelf width. The wavelength of long propagating disturbances is proportional to their frequency which is in turn proportional to the slope of the shelf. Thus it appears that for sufficiently weakly sloping shelves the tangential velocity component, i.e., the normal derivative of the streamfunction, at the shelf-ocean boundary can be made arbitrarily small. Here this will be taken as also giving a close approximation to the boundary condition at the shelf-ocean boundary when this boundary is no longer straight. The unapproximated boundary condition is that the streamfunction and its normal derivative are continuous across the boundary where they match to the decaying solution of Laplace's equation (to which (1.6) reduces in regions of constant depth). This gives a linear integral condition along the boundary. The unapproximated problem will not be pursued further here. The boundary condition at the coast is simply one of impermeability and thus on both rectilinear and curving coasts is simply that the streamfunction vanishes. Now consider flows of the form

$$\psi(x,y,t) = \mathrm{Re}\{\Phi(x,y)\exp(-2\mathrm{i}\omega\Omega\mathrm{t})\}, \tag{1.7}$$

so $\Phi(x,y)$ gives the spatial structure of the flow and ω its nondimensional frequency. Then Φ satisfies

$$\frac{1}{H}\,\Delta\,\Phi + \mathbf{grad}\left(\frac{1}{H}\right)\cdot\mathbf{grad}\,\Phi + \frac{i}{\omega}\mathbf{k}\cdot\left(\mathbf{grad}\,\Phi\times\mathbf{grad}\left(\frac{1}{H}\right)\right) = 0, \tag{1.8}$$

$$\Phi = 0 \text{ at the coast}, \tag{1.9}$$

$$\hat{\mathbf{n}}\cdot\mathbf{grad}\,\Phi = 0 \text{ at the shelf-ocean boundary}, \tag{1.10}$$

where vector $\hat{\mathbf{n}}$ is normal to the shelf-ocean boundary.

Mathematically, we are going to study the existence of trapped modes (i.e., the eigenvalues either embedded into the essential spectrum or lying in the gap of the essential spectrum) for the problem (1.8)–(1.10) in a curved strip. Similar problems for the Laplace operator have been extensively studied in the literature—either in a curved strip, in a straight strip with an obstacle, or in a strip with compactly perturbed boundary. In the case of Laplacians with Dirichlet boundary conditions these problems are usually called "quantum waveguides"; the Neumann case is usually referred to as "acoustic waveguides." The important result concerning quantum waveguides was established in [ExSe], [DuEx]: in the curved waveguides there always exists a trapped mode. Later this result was extended to more general settings; in particular, in [DiKr] (see also [KrKr]) it was shown that in the case of mixed boundary conditions (i.e., Dirichlet conditions on one side of the strip and Neumann conditions on the other side) trapped modes exist if the strip is curved "in the direction of the Dirichlet boundary."

The case of acoustic waveguides is more complicated because any eventual eigenvalues are embedded into the essential spectrum and are, therefore, highly unstable. Therefore, it is believed that in general the existence of trapped modes in this case is due to some sort of the symmetry of the problem (see [EvLeVa], [DaPa], [AsPaVa]).

In the present paper we use an approach similar to the one used in [DuEx] and [EvLeVa]; however, we have to modify this approach substantially due to the fact that we are working with a spectral problem for an operator pencil rather than that for an ordinary operator.

The rest of the paper is organized in the following way. In section 2, we discuss the rigorous mathematical statement of the problem; in section 3, we study the essential spectrum; and in section 4, we state and prove the main result on the existence of a discrete spectrum (Theorem 4.1). In particular we show that a trapped mode always exists if all of the following conditions are satisfied: (a) the depth profile H does not depend upon the longitudinal coordinate and is monotone increasing and logarithmically concave in the direction perpendicular to the coast; (b) the channel is curved in the direction of the Dirichlet boundary; (c) the curvature is sufficiently small.

Similar results can be obtained in a straight strip if the depth profile H depends nontrivially upon the longitudinal coordinate; we however do not discuss this problem here.

2. Mathematical statement of the problem.

2.1. Geometry. The original geometry is a straight planar strip of width δ:

$$G_0 = \{(x, y) : x \in \mathbb{R},\ y \in (0, \delta)\}.$$

Deformed geometry G is assumed to be a curved planar strip of constant width δ. To describe it precisely, we introduce the curve $\Gamma = \{(x = X(\xi),\, y = Y(\xi))\}$, $\xi \in \mathbb{R}$, where ξ is a natural arc-length parameter, i.e., $X'(\xi)^2 + Y'(\xi)^2 \equiv 1$. By

$$\gamma(\xi) = X''(\xi)Y'(\xi) - X'(\xi)Y''(\xi) \tag{2.1}$$

we denote a (signed) curvature of Γ (see Figure 1 and Remark 2.2). Note that $|\gamma(\xi)|^2 = X''(\xi)^2 + Y''(\xi)^2$.

We additionally assume

$$\operatorname{supp} \gamma \Subset [-R, R] \qquad \text{for some } R > 0, \tag{2.2}$$

and set

$$\kappa^+ = \sup_{\xi \in [-R,R]} \gamma(\xi), \qquad \kappa^- = -\inf_{\xi \in [-R,R]} \gamma(\xi). \tag{2.3}$$

We shall assume throughout the paper the smoothness condition

$$\gamma \in C^\infty(\mathbb{R}), \tag{2.4}$$

which can be obviously softened.

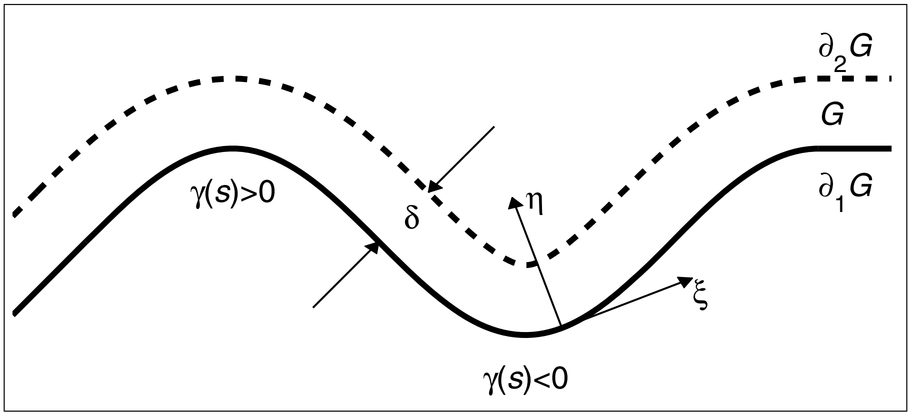


FIG. 1. *Domain G and curvilinear coordinates ξ, η. The solid line denotes the boundary $\partial_1 G$ with the Dirichlet boundary condition, and the dotted line denotes the boundary $\partial_2 G$ with the Neumann boundary condition.*

Now we can introduce, in a neighborhood of Γ, the curvilinear coordinates (ξ, η) as

$$x = X(\xi) - \eta Y'(\xi), \qquad y = Y(\xi) + \eta X'(\xi), \tag{2.5}$$

(where η is a distance from a point (x, y) to Γ) and describe the *deformed strip* G in these coordinates as

$$G = G_\gamma = \{(\xi, \eta) : \ \xi \in \mathbb{R}, \ \eta \in (0, \delta)\}. \tag{2.6}$$

REMARK 2.1. *As sets of points, $G_\gamma \equiv G_0$ for any γ, but the metrics are different, see below. We shall often omit the index γ if the metric is obvious from the context.*

REMARK 2.2. *Often one chooses the opposite sign in the definition of the signed curvature γ in* (2.1). *Our choice, though not canonical, is made to match the one in* [ExSe].

To avoid local self-intersections, we must restrict the width of the strip by natural conditions

$$\kappa^{\pm} \le A\delta^{-1}, \qquad A = \text{const} \in [0, 1). \tag{2.7}$$

We shall also assume throughout, without stating it explicitly, that G does not self-intersect globally, i.e., the mapping $(\xi, \eta) \mapsto (x, y)$ given by (2.5) is an injection on G.

Finally, it is an easy computation to show that the Euclidean metric in the curvilinear coordinates has a form $dx^2 + dy^2 = g d\xi^2 + d\eta^2$, where

$$g(\xi, \eta) = (1 + \eta\gamma(\xi))^2.$$

Later on, we shall widely use the notation

$$p(\xi, \eta) = (g(\xi, \eta))^{1/2} = 1 + \eta\gamma(\xi). \tag{2.8}$$

Note that in all the volume integrals,

$$\mathrm{d}G_\gamma = p(\xi, \eta)\, \mathrm{d}\xi\, \mathrm{d}\eta = (1 + \eta\gamma(\xi))\, \mathrm{d}\xi\, \mathrm{d}\eta = p(\xi, \eta)\, \mathrm{d}G_0.$$

2.2. Governing equations. For a given positive continuously differentiable function $H(\xi,\eta)$ (describing a depth profile), we are looking for a function $\Phi(\xi,\eta)$ satisfying (1.8) with spectral parameter ω.

By substituting

$$\beta(\xi,\eta) := \ln H(\xi,\eta), \tag{2.9}$$

and using explicit expressions for differential operators in curvilinear coordinates, we can rewrite (1.8) as

$$\begin{aligned}\omega\left(-\frac{1}{p^2}\frac{\partial^2\Phi}{\partial\xi^2}-\frac{\partial^2\Phi}{\partial\eta^2}+\left(\frac{1}{p^3}\frac{\partial p}{\partial\xi}+\frac{1}{p^2}\frac{\partial\beta}{\partial\xi}\right)\frac{\partial\Phi}{\partial\xi}+\left(\frac{\partial\beta}{\partial\eta}-\frac{1}{p}\frac{\partial p}{\partial\eta}\right)\frac{\partial\Phi}{\partial\eta}\right)\\ =\frac{i}{p}\left(\frac{\partial\beta}{\partial\xi}\frac{\partial\Phi}{\partial\eta}-\frac{\partial\beta}{\partial\eta}\frac{\partial\Phi}{\partial\xi}\right).\end{aligned} \tag{2.10}$$

Remark 2.3. *When deducing* (2.10)*, we have cancelled, on both sides, a common positive factor* $h(\xi,\eta) := \frac{1}{H(\xi,\eta)} = \mathrm{e}^{-\beta(\xi,\eta)}$*. However, we have to use this factor when considering corresponding variational equations, in order to keep the resulting forms symmetric. This leads to a special choice of weighted Hilbert spaces below.*

Further on, we consider only the case of a longitudinally uniform monotone depth profile,

$$\beta(\xi,\eta) \equiv \beta(\eta), \qquad \beta'(\eta) > 0, \tag{2.11}$$

in which case (2.10) simplifies to

$$\begin{aligned}\omega\left(-\frac{1}{p^2}\frac{\partial^2\Phi}{\partial\xi^2}-\frac{\partial^2\Phi}{\partial\eta^2}+\frac{1}{p^3}\frac{\partial p}{\partial\xi}\frac{\partial\Phi}{\partial\xi}+\left(\beta'-\frac{1}{p}\frac{\partial p}{\partial\eta}\right)\frac{\partial\Phi}{\partial\eta}\right)\\ =-\frac{i}{p}\beta'\frac{\partial\Phi}{\partial\xi},\end{aligned} \tag{2.12}$$

with $\beta' = \frac{\mathrm{d}\beta}{\mathrm{d}\eta}$.

2.3. Boundary conditions. Let $\partial_1 G = \{(\xi,0) : \xi\in\mathbb{R}\}$ and $\partial_2 G = \{(\xi,\delta) : \xi\in\mathbb{R}\}$ denote the lower and the upper boundary of the strip G, respectively. Boundary conditions (1.9), (1.10) then become

$$\Phi|_{\partial_1 G} = \left.\frac{\partial\Phi}{\partial\eta}\right|_{\partial_2 G} = 0. \tag{2.13}$$

Remark 2.4. *If the flow is confined to a channel, then the Dirichlet boundary condition* (1.9) *applies on both channel walls. This leads to a mathematically different problem which we do not consider in this paper.*

2.4. Function spaces and rigorous operator statement. We want to discuss the function spaces in which everything acts. Let us denote by $L_2(G;h)$ the Hilbert space of functions $\phi : G\to\mathbb{C}$ which are square-integrable on G with the weight $h(\eta)\equiv\frac{1}{H}=\exp(-\beta(\eta))$:

$$\|\phi\|^2_{L_2(G;h)} = \int_G |\phi(\xi,\eta)|^2\,h(\eta)\,\mathrm{d}G = \int_{\mathbb{R}}\int_0^\delta |\phi(\xi,\eta)|^2\,h(\eta)p(\xi,\eta)\,\mathrm{d}\eta\,\mathrm{d}\xi < \infty.$$

The corresponding inner product will be denoted $\langle\cdot,\cdot\rangle_{L_2(G;h)}$. Similarly we can define the space $L_2(F;h)$ for an arbitrary open subset F of G.

Let us formally introduce the operators

$$\mathcal{L}_\gamma : \Phi \mapsto -\frac{1}{p^2}\frac{\partial^2\Phi}{\partial\xi^2} - \frac{\partial^2\Phi}{\partial\eta^2} + \frac{1}{p^3}\frac{\partial p}{\partial\xi}\frac{\partial\Phi}{\partial\xi} + \left(\beta' - \frac{1}{p}\frac{\partial p}{\partial\eta}\right)\frac{\partial\Phi}{\partial\eta}$$

and

$$\mathcal{M}_\gamma : \Phi \mapsto -\frac{i}{p}\beta'\frac{\partial\Phi}{\partial\xi}.$$

(The dependence on γ is of course via p; see (2.8).) Then (2.12) can be formally rewritten as

$$\omega\mathcal{L}_\gamma\Phi = \mathcal{M}_\gamma\Phi, \tag{2.14}$$

or via an *operator pencil*

$$\mathcal{A}_\gamma \equiv \mathcal{A}_\gamma(\omega) = \omega\mathcal{L}_\gamma - \mathcal{M}_\gamma \tag{2.15}$$

as

$$\mathcal{A}_\gamma(\omega)\Phi = 0. \tag{2.16}$$

The *domain* of the pencil $\mathcal{A}_\gamma$ in the L_2-sense is naturally defined as

$$\mathrm{Dom}(\mathcal{A}_\gamma) = \{\Phi \in H^2(G),\ \Phi \text{ satisfies } (2.13)\}, \tag{2.17}$$

where H^2 denotes a standard Sobolev space.

On the domain (2.17), $\mathcal{M}_\gamma$ is symmetric, and $\mathcal{L}_\gamma$ is symmetric and positive in the sense of the scalar product $\langle\cdot,\cdot\rangle_{L_2(G;h)}$, with

$$\langle\mathcal{L}_\gamma\Phi,\Phi\rangle_{L_2(G;h)} = \int_{\mathbb{R}}\int_0^\delta \left(\frac{1}{p}\left|\frac{\partial\Phi}{\partial\xi}\right|^2 + p\left|\frac{\partial\Phi}{\partial\eta}\right|^2\right)\mathrm{e}^{-\beta}\,\mathrm{d}\eta\,\mathrm{d}\xi.$$

Later on, we shall use a *weak* (or *variational*) form of (2.16), and shall require some other function spaces described below. Let $F \subseteq G$, and suppose its boundary is decomposed into two disjoint parts: $\partial F = \partial_1 F \sqcup \partial_2 F$. We introduce the space

$$\begin{aligned}\widetilde{C}_0^\infty(F;\partial_1 F) = \{\phi \in C^\infty(F) : \overline{\operatorname{supp}\phi} \cap \partial_1 F = \emptyset,\\ \text{and there exists } r > 0 \text{ such that } \phi(\xi,\eta) = 0 \text{ for } (\xi,\eta)\in F,\ |\xi| \ge r\}\end{aligned}$$

consisting of smooth functions with compact support vanishing near $\partial_1 F$.

By $\widetilde{H}_0^1(F;\partial_1 F;h)$ we denote the closure of $\widetilde{C}_0^\infty(F;\partial_1 F)$ with respect to the scalar product

$$\langle\phi,\psi\rangle_{\widetilde{H}_0^1(F,\partial_1 F;h)} = \langle\phi,\psi\rangle_{L_2(F;h)} + \langle\mathbf{grad}\,\phi,\mathbf{grad}\,\psi\rangle_{L_2(F;h)}.$$

In what follows we shall study the operators $\mathcal{L}_\gamma$, $\mathcal{M}_\gamma$ and the pencil $\mathcal{A}_\gamma$ from a variational point of view. The details are given in the next section; here we note only that from now we understand the expression $\langle\mathcal{L}_\gamma\Psi,\Psi\rangle_{L_2(G_\gamma,h)}$ as the quadratic form for the operator $\mathcal{L}_\gamma$, with the quadratic form domain $\widetilde{H}_0^1(G;\partial_1 G;h)$.

The main purpose of this paper is to study the *spectral properties* of the operator pencil $\mathcal{A}_\gamma$. We recall the following definitions.

A number $\omega \in \mathbb{C}$ is said to belong to the *spectrum* of $\mathcal{A}_\gamma$ (denoted $\operatorname{spec}(\mathcal{A}_\gamma)$) if $\mathcal{A}_\gamma(\omega)$ is not invertible.

It is easily seen that in our case the spectrum of $\mathcal{A}_\gamma$ is real.

We say that $\omega \in \mathbb{R}$ belongs to the *essential spectrum* of the operator pencil $\mathcal{A}_\gamma$ (denoted $\omega \in \operatorname{spec}_{\mathrm{ess}}(\mathcal{A}_\gamma)$) if for this ω the operator $\mathcal{A}_\gamma(\omega)$ is non-Fredholm.

We say that $\omega \in \mathbb{C}$ belongs to the *point spectrum* of the operator pencil $\mathcal{A}_\gamma$ (denoted $\omega \in \operatorname{spec}_{\mathrm{pt}}(\mathcal{A}_\gamma)$), or, in other words, say that ω is an *eigenvalue*, if for this ω there exists a nontrivial solution $\Psi \in \operatorname{Dom}(\mathcal{A}_\gamma)$ of the problem $\mathcal{A}_\gamma(\omega)\Psi = 0$.

It is known that the essential spectrum is a closed subset of $\mathbb{R}$, and that any point of the spectrum outside the essential spectrum is an isolated eigenvalue of finite multiplicity. The set of all such points is called the *discrete spectrum*, and will be denoted $\operatorname{spec}_{\mathrm{dis}}(\mathcal{A}_\gamma)$. There may, however, exist the points of the spectrum which belong to both the essential spectrum and the point spectrum.

Our main result (Theorem 4.1 below) establishes some conditions on the curvature γ of the waveguide which guarantee the existence of eigenvalues of $\mathcal{A}_\gamma$.

It is more convenient to deal with problems of this type variationally, and we start the next section with an abstract variational scheme suitable for self-adjoint pencils with nonempty essential spectrum.

3. Essential spectrum.

3.1. Variational principle for the essential spectrum.

DEFINITION 3.1. *We set, for $j \in \mathbb{N}$,*

$$\mu_{\gamma,j} = \sup_{\substack{U \subset \widetilde{H}_0^1(G;\partial_1 G;h) \\ \dim U = j}} \inf_{\Psi \in U,\, \Psi \neq 0} \frac{\langle \mathcal{M}_\gamma \Psi, \Psi\rangle_{L_2(G_\gamma,h)}}{\langle \mathcal{L}_\gamma \Psi, \Psi\rangle_{L_2(G_\gamma,h)}}. \tag{3.1}$$

As $\langle \mathcal{L}_\gamma \Psi, \Psi\rangle_{L_2(G_\gamma,h)}$ is positive, the right-hand side of (3.1) is well defined, though the numbers $\mu_{\gamma,j}$ may a priori be finite or infinite.

Obviously, for any fixed curvature profile γ the numbers $\mu_{\gamma,j}$ form a nonincreasing sequence:

$$\mu_{\gamma,1} \geq \mu_{\gamma,2} \geq \cdots \geq \mu_{\gamma,j} \geq \mu_{\gamma,j+1} \geq \cdots .$$

DEFINITION 3.2. *Denote*

$$\underline{\mu}_\gamma = \lim_{j\to\infty} \mu_{\gamma,j}. \tag{3.2}$$

For general self-adjoint operator pencils the analogue of (3.2) may be finite or equal to $\pm\infty$; as we shall see below, in our case $\underline{\mu}_\gamma$ is finite.

The following result is a modification, to the case of an abstract self-adjoint linear pencil, of the general variational principle for a self-adjoint operator with an essential spectrum; see [Da, Prop. 4.5.2].

PROPOSITION 3.3. *Either*

(i) $\underline{\mu}_\gamma > -\infty$, *and then* $\sup \operatorname{spec}_{\mathrm{ess}}(\mathcal{A}_\gamma) = \underline{\mu}_\gamma$,

or

(ii) $\underline{\mu}_\gamma = -\infty$, *and then* $\operatorname{spec}_{\mathrm{ess}}(\mathcal{A}_\gamma) = \emptyset$.

Moreover, if $\mu_{\gamma,j} > \underline{\mu}_\gamma$, *then* $\mu_{\gamma,j} \in \operatorname{spec}_{\mathrm{dis}}(\mathcal{A}_\gamma)$.

Proposition 3.3 ensures that we can use the variational principle (3.1) in order to find the eigenvalues of the pencil $\mathcal{A}_\gamma$ lying *above* the supremum $\underline{\mu}_\gamma$ of the essential spectrum.

3.2. Essential spectrum for the straight strip. The spectral analysis in the case of a straight strip ($\gamma \equiv 0$) is rather straightforward as the problem admits in this case the separation of variables.

Let us seek the solutions of (2.12), (2.13) in the case of a straight strip ($\gamma \equiv 0$, and so $p \equiv 1$) in the form

$$\Phi(\xi, \eta) = \phi(\eta) \exp(i\alpha\xi); \tag{3.3}$$

it is sufficient to consider only real values of α.

After separation of variables, (2.12), (2.13) are written, for each α, as a one-dimensional transversal spectral problem

$$\omega(-\phi'' + \beta'\phi' + \alpha^2\phi) = \alpha\beta'\phi, \qquad \phi(0) = \phi'(\delta) = 0. \tag{3.4}$$

Alternatively, introduce operators

$$\mathfrak{l}_\alpha : \phi \mapsto -\phi'' + \beta'\phi' + \alpha^2\phi, \qquad \mathfrak{m}_\alpha : \phi \mapsto \alpha\beta'\phi,$$

and a pencil

$$\mathfrak{a}_\alpha(\omega) = \omega\mathfrak{l}_\alpha - \mathfrak{m}_\alpha,$$

(again understood in an $L_2((0,\delta); h)$ sense with the domain defined similarly to (2.17)), and consider a one-dimensional operator pencil spectral problem $\mathfrak{a}_\alpha(\omega)\phi = 0$.

For a fixed value of α, the one-dimensional linear operator pencil (3.4) has the essential spectrum $\{0\}$ and a discrete spectrum $\operatorname{spec}(\mathfrak{a}_\alpha)$; note that

$$\operatorname{spec}(\mathfrak{a}_{-\alpha}) = -\operatorname{spec}(\mathfrak{a}_\alpha). \tag{3.5}$$

Denote, for $\alpha > 0$, the top of the spectrum of this transversal problem by $\omega_\alpha = \sup \operatorname{spec}(\mathfrak{a}_\alpha)$.

Lemma 3.4. *Let $\alpha > 0$. Then, under condition* (2.11)*,*

(i) $\operatorname{spec}(\mathfrak{a}_\alpha) \subset [0, +\infty)$*;*
(ii) $0 < \omega_\alpha < +\infty$*;*
(iii) $\omega_\alpha \to +0$ *as* $\alpha \to \infty$.

Proof. By the variational principle analogous to Proposition 3.3(i),

$$\omega_\alpha = \sup_{\substack{\phi \in \widetilde{H}_0^1((0,\delta),0,h) \\ \phi \neq 0}} J_\alpha(\phi),$$

where we set

$$J_\alpha(\phi) = \frac{\langle \mathfrak{m}_\alpha\phi, \phi\rangle_{L_2((0,\delta),h)}}{\langle \mathfrak{l}_\alpha\phi, \phi\rangle_{L_2((0,\delta),h)}} = \frac{\displaystyle\int_0^\delta \alpha\beta'(\eta)\phi(\eta)^2 h(\eta)\,\mathrm{d}\eta}{\displaystyle\int_0^\delta (-\phi''(\eta) + \beta'(\eta)\phi'(\eta) + \alpha^2\phi(\eta))\phi(\eta)h(\eta)\,\mathrm{d}\eta}. \tag{3.6}$$

After integrating by parts using $h(\eta) = \mathrm{e}^{-\beta(\eta)}$ and inverting the quotient, we get

$$J_\alpha(\phi) = \left(\alpha J^{(1)}(\phi) + \frac{1}{\alpha} J^{(2)}(\phi)\right)^{-1}, \tag{3.7}$$

where we denote

$$J^{(1)}(\phi) = \frac{\displaystyle\int_0^\delta \mathrm{e}^{-\beta(\eta)}|\phi(\eta)|^2\,\mathrm{d}\eta}{\displaystyle\int_0^\delta \beta'(\eta)\mathrm{e}^{-\beta(\eta)}|\phi(\eta)|^2\,\mathrm{d}\eta}$$

and

$$J^{(2)}(\phi) = \frac{\displaystyle\int_0^\delta \mathrm{e}^{-\beta(\eta)}|\phi'(\eta)|^2\,\mathrm{d}\eta}{\displaystyle\int_0^\delta \beta'(\eta)\mathrm{e}^{-\beta(\eta)}|\phi(\eta)|^2\,\mathrm{d}\eta}.$$

The statements (ii) and (iii) of the lemma now follow immediately from the estimates

$$J^{(1)}(\phi) \ge \frac{\inf\limits_{\eta\in(0,\delta)} \mathrm{e}^{-\beta(\eta)}}{\sup\limits_{\eta\in(0,\delta)} (\beta'(\eta)\mathrm{e}^{-\beta(\eta)})}$$

and

$$J^{(2)}(\phi) \ge \frac{\pi^2}{4\delta^2}\,\frac{\inf\limits_{\eta\in(0,\delta)} \mathrm{e}^{-\beta(\eta)}}{\sup\limits_{\eta\in(0,\delta)} (\beta'(\eta)\mathrm{e}^{-\beta(\eta)})},$$

where the latter inequality uses the variational principle and the fact that the principal eigenvalue of the mixed Dirichlet–Neumann spectral problem for the operator $-\frac{\mathrm{d}^2}{\mathrm{d}\eta^2}$ on the interval $(0,\delta)$ is equal to $\frac{\pi^2}{4\delta^2}$. The statement (i) follows from the positivity of the right-hand side of (3.7). ∎

We are now able to find the essential spectrum of the problem in a straight strip.

Lemma 3.5. *Assume that conditions* (2.11) *hold. Then*

$$\mathrm{spec}_{\mathrm{ess}}(\mathcal{A}_0) = [-\Omega_*, \Omega_*], \tag{3.8}$$

where

$$\Omega_* = \sup_{\phi\in\widetilde{H}_0^1((0,\delta),0,h)} \frac{\displaystyle\frac{1}{2}\int_0^\delta \beta'(\eta)\mathrm{e}^{-\beta(\eta)}|\phi(\eta)|^2\,\mathrm{d}\eta}{\sqrt{\displaystyle\int_0^\delta \mathrm{e}^{-\beta(\eta)}|\phi'(\eta)|^2\,\mathrm{d}\eta\cdot\int_0^\delta \mathrm{e}^{-\beta(\eta)}|\phi(\eta)|^2\,\mathrm{d}\eta}} > 0. \tag{3.9}$$

Proof. It is standard that

$$\mathrm{spec}_{\mathrm{ess}}(\mathcal{A}_0) = \overline{\bigcup_{\alpha\in\mathbb{R}} \mathrm{spec}(\mathfrak{a}_\alpha)}.$$

Thus, by Lemma 3.4, and with account of the antisymmetry of the spectrum of $\mathfrak{a}_\alpha$ with respect to α and its positivity for $\alpha>0$, we have

$$\sup\mathrm{spec}_{\mathrm{ess}}(\mathcal{A}_0) = \sup_{\alpha>0}\omega_\alpha = \sup_{\alpha>0}\ \sup_{\phi\in\widetilde{H}_0^1((0,\delta),0,h)} J_\alpha(\phi).$$

By maximizing first with respect to α, we obtain, from (3.7),

$$J_\alpha(\phi) \le J_{\alpha_*(\phi)}(\phi)$$

with the maximizer

$$\alpha_*(\phi) = \sqrt{\frac{J^{(2)}(\phi)}{J^{(1)}(\phi)}}.$$

Maximizing now with respect to ϕ gives $\sup \operatorname{spec}_{\mathrm{ess}}(\mathcal{A}_0) = \Omega_*$, with Ω_* given by (3.9). Finally, we note that ω_α depends continuously on $\alpha > 0$. Since $\omega_\alpha \to +0$ as $\alpha \to +\infty$ by Lemma 3.4(iii), ω_α thus takes all the values in $(0, \Omega_*]$. Therefore, the closed interval $[0, \Omega_*]$ lies in $\operatorname{spec}_{\mathrm{ess}}(\mathcal{A}_0)$. By symmetry (3.5), we also have $[-\Omega_*, 0] \subset \operatorname{spec}_{\mathrm{ess}}(\mathcal{A}_0)$, which finishes the proof. $\square$

3.3. Essential spectrum for a curved strip. It is now a standard procedure to show that under our conditions the essential spectrum of the problem in a curved strip coincides with the essential spectrum of the problem in a straight strip given by Lemma 3.5. Namely, we have the following.

LEMMA 3.6. *Let us assume conditions* (2.2), (2.4), *and* (2.11) *hold. Then*

$$\operatorname{spec}_{\mathrm{ess}}(\mathcal{A}_\gamma) = \operatorname{spec}_{\mathrm{ess}}(\mathcal{A}_0) = [-\Omega_*, \Omega_*]$$

with Ω_* *given by* (3.9).

The proof is based on the fact that any solution of the problem (2.12), (2.13) with $\gamma \not\equiv 0$ (and thus $p \not\equiv 1$) should coincide in

$$G \cap \{|\xi| > R > \max(|\inf \operatorname{supp} \gamma|, |\sup \operatorname{supp} \gamma|)\}$$

with a solution of the same problem for $\gamma \equiv 0$. An analogous result has been proved in a number of similar situations elsewhere (see, e.g., [ExSe], [EvLeVa], [DaPa], [KrTA]), so we omit the details of the proof. We briefly note that the inclusion $\operatorname{spec}_{\mathrm{ess}}(\mathcal{A}_\gamma) \subseteq \operatorname{spec}_{\mathrm{ess}}(\mathcal{A}_0)$ is proved using the separation of variables as above and a construction of appropriate Weyl's sequences, and in order to prove the inclusion $\operatorname{spec}_{\mathrm{ess}}(\mathcal{A}_\gamma) \supseteq \operatorname{spec}_{\mathrm{ess}}(\mathcal{A}_0)$ one can use the Dirichlet–Neumann bracketing and the discreteness of the spectrum of the problem (2.12), (2.13) considered in $G \cap \{|\xi| < R\}$ with additional Dirichlet or Neumann boundary conditions imposed on the "cuts" $\{\xi = \pm R\}$.

4. Main result. Our main result consists in stating some sufficient conditions on the depth profile $\beta(\eta)$ and the curvature profile $\gamma(\xi)$ which guarantee the existence of an eigenvalue of the pencil $\mathcal{A}_\gamma$ lying outside the essential spectrum.

THEOREM 4.1. *Assume, as before, that condition* (2.11) *holds. Assume additionally that*

$$\beta''(\eta) < 0 \qquad \textit{for } \eta \in (0, \delta). \tag{4.1}$$

Then there exists a constant $C_\beta > 0$, *which depends only on the depth profile* β, *such that* $\operatorname{spec}_{\mathrm{dis}}(\mathcal{A}_\gamma) \neq \emptyset$ *whenever* γ *satisfies conditions* (2.2), (2.4), *and*

$$\int \gamma(\xi)\, \mathrm{d}\xi > C_\beta \int \gamma(\xi)^2\, \mathrm{d}\xi. \tag{4.2}$$

We give an explicit expression for C_β below; see (4.15).

An integral sufficient condition (4.2) may be replaced by a pointwise, although more restrictive, condition.

COROLLARY 4.2. *Assume that conditions* (2.11) *and* (4.1) *hold. Then there exists a constant* $c_{\beta,R} = \frac{C_\beta}{2R}$ *which depends only on the depth profile* β *and a given* $R > 0$ *such that* $\operatorname{spec}_{\mathrm{dis}}(\mathcal{A}_\gamma) \neq \emptyset$ *whenever* $\gamma \not\equiv 0$ *satisfies conditions* (2.2), (2.4), *and*

$$0 \le \gamma(\xi) < c_{\beta,R} \qquad \text{for } |\xi| \le R. \tag{4.3}$$

We prove Theorem 4.1 using a number of simple lemmas, the central of which is the following.

LEMMA 4.3. *Suppose there exists a function* $\widetilde{\Psi} \in \widetilde{H}_0^1(G_\gamma, h)$ *such that*

$$\frac{\langle \mathcal{M}_\gamma \widetilde{\Psi}, \widetilde{\Psi}\rangle_{L_2(G_\gamma,h)}}{\langle \mathcal{L}_\gamma \widetilde{\Psi}, \widetilde{\Psi}\rangle_{L_2(G_\gamma,h)}} > \Omega_*. \tag{4.4}$$

Then there exists $\omega > \Omega_*$ *which belongs to* $\operatorname{spec}_{\mathrm{dis}}(\mathcal{A}_\gamma)$.

Lemma 4.3 is just a restatement of the variational principle of Proposition 3.3. The main difficulty in its application is of course the choice of an appropriate test function $\widetilde{\Psi}$. However such choice becomes much easier if we use the following modification of this lemma which allows us to consider test functions which are not necessarily square-integrable on G_γ.

Denote, for brevity, $G_\gamma^r = G_\gamma \cap \{|\xi| < r\}$.

LEMMA 4.4. *Suppose there exist a function* Ψ *and a constant* D *such that, for any* $r > R$, *we have* $\Psi \in \widetilde{H}_0^1(G_\gamma^r, h)$ *and*

$$\langle \mathcal{M}_\gamma \Psi, \Psi\rangle_{L_2(G_\gamma^r,h)} - \Omega_* \langle \mathcal{L}_\gamma \Psi, \Psi\rangle_{L_2(G_\gamma^r,h)} \ge D > 0. \tag{4.5}$$

Then there exists $\omega > \Omega_*$ *which belongs to* $\operatorname{spec}_{\mathrm{dis}}(\mathcal{A}_\gamma)$.

The proof of Lemma 4.4 uses the construction of an appropriate cutoff function $\chi(\xi)$ such that $\widetilde{\Psi} = \chi\Psi$ satisfies the conditions of Lemma 4.3; cf. [DaPa, Prop. 1].

We now proceed as follows.

Let $\phi_*(\eta)$ be a maximizer in (3.9), and set

$$\Psi(\xi,\eta) = \phi_*(\eta) \mathrm{e}^{i\alpha_\bullet \xi},$$

where

$$\alpha_\bullet = \alpha_*(\phi_*) = \sqrt{\frac{J^{(2)}(\phi_*)}{J^{(1)}(\phi_*)}}. \tag{4.6}$$

It is important to note that Ψ is in fact an "eigenfunction" of the essential spectrum of $\mathcal{A}_\gamma$ corresponding to its highest positive point Ω_* and that ϕ_* is an eigenfunction of (3.4) with $\alpha = \alpha_\bullet$ (i.e., of the pencil $\mathfrak{a}_{\alpha_\bullet}$) again corresponding to the eigenvalue Ω_*, and so

$$\phi_*'' = \beta'\phi_*' + (\alpha_\bullet^2 - \Lambda_*\alpha_\bullet\beta')\phi_*, \qquad \phi_*(0) = \phi_*'(\delta) = 0 \tag{4.7}$$

with $\Lambda_* := \frac{1}{\Omega_*}$ (cf. (3.4)).

For future use, we summarize the relations obtained so far:

$$\begin{aligned}
\mathcal{L}_0\Psi &= (-\phi_*''(\eta) + \beta'(\eta)\phi_*'(\eta) + \alpha_\bullet\phi_*(\eta))\,\mathrm{e}^{i\alpha_\bullet\xi} = (\mathfrak{l}_{\alpha_\bullet}\phi_*)\,\mathrm{e}^{i\alpha_\bullet\xi},\\
\mathcal{M}_0\Psi &= \alpha_\bullet\beta'(\eta)\phi_*(\eta)\,\mathrm{e}^{i\alpha_\bullet\xi} = (\mathfrak{m}_{\alpha_\bullet}\phi_*)\,\mathrm{e}^{i\alpha_\bullet\xi},\\
\mathcal{L}_\gamma\Psi &= \Bigg(-\phi_*''(\eta) + \left(\beta'(\eta) - \frac{1}{p(\xi,\eta)}\,\frac{\partial p(\xi,\eta)}{\partial\eta}\right)\phi_*'(\eta)\\
&\qquad + \left(\frac{i\alpha_\bullet}{p(\xi,\eta)^3}\,\frac{\partial p(\xi,\eta)}{\partial\xi} + \frac{\alpha_\bullet^2}{p(\xi,\eta)^2}\right)\phi_*(\eta)\Bigg)\,\mathrm{e}^{i\alpha_\bullet\xi},\\
\mathcal{M}_\gamma\Psi &= \frac{\alpha_\bullet}{p(\xi,\eta)}\beta'(\eta)\phi_*(\eta)\,\mathrm{e}^{i\alpha_\bullet\xi},\\
p(\xi,\eta) &= 1 + \gamma(\xi)\eta
\end{aligned}$$

(with $'$ denoting differentiation with respect to η).

It is important to note that for any $r > 0$,

$$\frac{\langle\mathcal{M}_0\Psi,\Psi\rangle_{L_2(G_0^r,h)}}{\langle\mathcal{L}_0\Psi,\Psi\rangle_{L_2(G_0^r,h)}} = \frac{\langle\mathfrak{m}_{\alpha_\bullet}\phi_*,\phi_*\rangle_{L_2((0,\delta),h)}}{\langle\mathfrak{l}_{\alpha_\bullet}\phi_*,\phi_*\rangle_{L_2((0,\delta),h)}} = \Omega_* > 0$$

and, as explicit formulae above show,

$$\begin{aligned}
\langle\mathcal{M}_\gamma\Psi,\Psi\rangle_{L_2(G_\gamma^r,h)} &= \langle\mathcal{M}_0\Psi,\Psi\rangle_{L_2(G_0^r,h)} = \alpha_\bullet\int_{-r}^{r}\int_0^\delta \beta'(\eta)\mathrm{e}^{-\beta(\eta)}|\phi_*(\eta)|^2\,\mathrm{d}\eta\,\mathrm{d}\xi\\
&= 2r\alpha_\bullet\int_0^\delta \beta'(\eta)\mathrm{e}^{-\beta(\eta)}|\phi_*(\eta)|^2\,\mathrm{d}\eta > 0.
\end{aligned} \tag{4.8}$$

We want to show that under conditions of Theorem 4.1 and with the choice of Ψ as above, inequality (4.5) holds for any $r > R$.

In view of (4.8), it is enough to show that

$$D_\gamma := \langle\mathcal{L}_\gamma\Psi,\Psi\rangle_{L_2(G_\gamma^r,h)} - \langle\mathcal{L}_0\Psi,\Psi\rangle_{L_2(G_0^r,h)}$$

is *negative* for $r > R$.

Explicit substitution gives, after taking into account the formula

$$\int \frac{1}{p(\xi,\eta)^2}\,\frac{\partial p(\xi,\eta)}{\partial\xi}\,\mathrm{d}\xi = 0$$

(due to (2.2), with account of (2.8)), the following expression:

$$D_\gamma = \int_{-r}^{r}\int_0^\delta \gamma(\xi)\eta\mathrm{e}^{-\beta(\eta)}|\phi_*'(\eta)|^2\,\mathrm{d}\eta\,\mathrm{d}\xi - \int_{-r}^{r}\int_0^\delta \alpha_\bullet^2\frac{\eta\gamma(\xi)}{1+\eta\gamma(\xi)}\mathrm{e}^{-\beta(\eta)}|\phi_*(\eta)|^2\,\mathrm{d}\eta\,\mathrm{d}\xi.$$

This, in turn, can be rewritten, using the obvious identity

$$\frac{\eta\gamma(\xi)}{1+\eta\gamma(\xi)} = \eta\gamma(\xi) - \frac{\eta^2\gamma(\xi)^2}{1+\eta\gamma(\xi)},$$

as

$$\begin{aligned}
D_\gamma &= \int_{-r}^{r}\gamma(\xi)\int_0^\delta \eta\mathrm{e}^{-\beta(\eta)}\left(|\phi_*'(\eta)|^2 - \alpha_\bullet^2|\phi_*(\eta)|^2\right)\mathrm{d}\eta\,\mathrm{d}\xi\\
&\quad + \alpha_\bullet^2\int_{-r}^{r}\int_0^\delta \frac{\eta^2\gamma(\xi)^2}{1+\eta\gamma(\xi)}\mathrm{e}^{-\beta(\eta)}|\phi_*(\eta)|^2\,\mathrm{d}\eta\,\mathrm{d}\xi.
\end{aligned} \tag{4.9}$$

We shall deal with the two terms in (4.9) separately.

The first one is more difficult. As (4.6) yields explicitly

$$\alpha_\bullet^2 = \frac{\displaystyle\int_0^\delta \mathrm{e}^{-\beta(\eta)}|\phi_*'(\eta)|^2\,\mathrm{d}\eta}{\displaystyle\int_0^\delta \mathrm{e}^{-\beta(\eta)}|\phi_*(\eta)|^2\,\mathrm{d}\eta},$$

we get

$$\begin{aligned}
I_1 &:= \int_0^\delta \eta\mathrm{e}^{-\beta(\eta)}\left(|\phi_*'(\eta)|^2 - \alpha_\bullet^2|\phi_*(\eta)|^2\right)\mathrm{d}\eta \\
&= \frac{1}{\displaystyle\int_0^\delta \mathrm{e}^{-\beta(\eta)}|\phi_*(\eta)|^2\,\mathrm{d}\eta} \times \left(\int_0^\delta \eta\mathrm{e}^{-\beta(\eta)}|\phi_*'(\eta)|^2\,\mathrm{d}\eta \cdot \int_0^\delta \mathrm{e}^{-\beta(\eta)}|\phi_*(\eta)|^2\,\mathrm{d}\eta\right. \\
&\qquad\left. - \int_0^\delta \mathrm{e}^{-\beta(\eta)}|\phi_*'(\eta)|^2\,\mathrm{d}\eta \cdot \int_0^\delta \eta\mathrm{e}^{-\beta(\eta)}|\phi_*(\eta)|^2\,\mathrm{d}\eta\right).
\end{aligned} \tag{4.10}$$

We want to show that the term in brackets is negative under some reasonable assumptions.

Lemma 4.5. *Assume that the conditions of Theorem* 4.1 *hold. Then* $I_1 < 0$.

The proof of Lemma 4.5 uses the following simple fact.[1]

Lemma 4.6. *Let* $(a,b) \subset (0,+\infty)$ *be a finite interval, and let a function* $g : (a,b) \to \mathbb{R}$ *be nonincreasing. Then*

$$\left(\int_a^b xg(x)f(x)\,\mathrm{d}x\right)\cdot\left(\int_a^b f(x)\,\mathrm{d}x\right) - \left(\int_a^b g(x)f(x)\,\mathrm{d}x\right)\cdot\left(\int_a^b xf(x)\,\mathrm{d}x\right) \le 0$$

for any nonnegative function $f : (a,b) \to \mathbb{R}$.

Proof of Lemma 4.6. We have

$$\begin{aligned}
&\left(\int_a^b xg(x)f(x)\,\mathrm{d}x\right)\cdot\left(\int_a^b f(x)\,\mathrm{d}x\right) - \left(\int_a^b g(x)f(x)\,\mathrm{d}x\right)\cdot\left(\int_a^b xf(x)\,\mathrm{d}x\right) \\
&\quad = \int_a^b\int_a^b xg(x)f(x)f(y)\,\mathrm{d}x\,\mathrm{d}y - \int_a^b\int_a^b g(x)f(x)yf(y)\,\mathrm{d}x\,\mathrm{d}y \\
&\quad = \int_a^b\int_a^y (x-y)f(x)f(y)g(x)\,\mathrm{d}x\,\mathrm{d}y + \int_a^b\int_y^b (x-y)f(x)f(y)g(x)\,\mathrm{d}x\,\mathrm{d}y.
\end{aligned}$$

Interchanging the variables x and y in the last integral, we obtain that the whole expression is equal to

$$\int_a^b\int_a^y \underbrace{(x-y)}_{\text{nonpositive}}\ \underbrace{f(x)f(y)(g(x)-g(y))}_{\text{nonnegative}}\,\mathrm{d}x\,\mathrm{d}y$$

and is therefore nonpositive. ∎

[1]We are grateful to Daniel Elton for a useful suggestion that helped to prove this lemma.

We can now proceed with evaluating I_1.

Proof of Lemma 4.5. We act by doing a lot of integrations by parts. We shall also use (4.7).

We have (all integrals are over $[0,\delta]$ and with respect to η)

$$\begin{aligned}\int \eta \mathrm{e}^{-\beta}|\phi_*'|^2 &= -\int \phi_* \cdot (\eta \mathrm{e}^{-\beta}\phi_*')' \\ &= -\int \phi_* \cdot (\mathrm{e}^{-\beta}\phi_*' - \beta'\eta\mathrm{e}^{-\beta}\phi_*' + \eta\mathrm{e}^{-\beta}\phi_*'') \\ &= -\int \phi_* \cdot (\mathrm{e}^{-\beta}\phi_*' + (\alpha_\bullet^2 - \Lambda_*\alpha_\bullet\beta')\mathrm{e}^{-\beta}\eta\phi_*).\end{aligned}$$

Further,

$$-\int (\phi_*\mathrm{e}^{-\beta})\phi_*' = \left(\int \left(\phi_*'\mathrm{e}^{-\beta} - \beta'\phi_*\mathrm{e}^{-\beta}\right)\phi_*\right) - \mathrm{e}^{-\beta(\delta)}\phi_*^2(\delta),$$

thus producing

$$\begin{aligned}\int \eta\mathrm{e}^{-\beta}|\phi_*'|^2 &= -\frac{1}{2}\int \beta'\phi_*^2\mathrm{e}^{-\beta} \underbrace{-\frac{1}{2}\mathrm{e}^{-\beta(\delta)}\phi_*^2(\delta)}_{\text{negative constant}} \\ &\quad - \alpha_\bullet^2\int \eta\mathrm{e}^{-\beta}|\phi_*|^2 + \Lambda_*\alpha_\bullet\int \eta\beta'\mathrm{e}^{-\beta}|\phi_*|^2.\end{aligned} \tag{4.11}$$

Also,

$$\begin{aligned}\int \mathrm{e}^{-\beta}|\phi_*'|^2 &= -\int \mathrm{e}^{-\beta}\phi_*(-\beta'\phi_*' + \phi_*'') \\ &= -\int \mathrm{e}^{-\beta}\phi_*\left(-\beta'\phi_*' + \beta'\phi_*' + \alpha_\bullet^2\phi_* - \Lambda_*\alpha_\bullet\beta'\phi_*\right) \\ &= -\int \mathrm{e}^{-\beta}\phi_*^2\,(\alpha_\bullet^2 - \Lambda_*\alpha_\bullet\beta').\end{aligned} \tag{4.12}$$

Substituting (4.11) and (4.12) into (4.10), and simplifying, we get

(4.13)

$$\begin{aligned}I_1 \cdot \underbrace{\int \mathrm{e}^{-\beta(\eta)}|\phi_*(\eta)|^2}_{\text{positive integral}} &= \int \eta\mathrm{e}^{-\beta}|\phi_*'|^2 \cdot \int \mathrm{e}^{-\beta}|\phi_*|^2 - \int \mathrm{e}^{-\beta}|\phi_*'|^2 \cdot \int \eta\mathrm{e}^{-\beta}|\phi_*|^2 \\ &= \left(\underbrace{\left(-\frac{1}{2}\int \beta'\mathrm{e}^{-\beta}|\phi_*|^2\right)}_{\text{negative as } \beta'>0} + \underbrace{\left(-\frac{1}{2}\mathrm{e}^{-\beta(\delta)}\phi_*^2(\delta)\right)}_{\text{negative constant}}\right)\int \mathrm{e}^{-\beta}|\phi_*|^2 \\ &\quad + \underbrace{(\Lambda_*\alpha_\bullet)}_{\text{+ve constant}} \times \underbrace{\left(\int \eta\beta'\mathrm{e}^{-\beta}|\phi_*|^2 \cdot \int \mathrm{e}^{-\beta}|\phi_*|^2 - \int \beta'\mathrm{e}^{-\beta}|\phi_*|^2 \cdot \int \eta\mathrm{e}^{-\beta}|\phi_*|^2\right)}_{\text{nonpositive by Lemma 4.6 with } g=\beta',\ f=\mathrm{e}^{-\beta}|\phi_*|^2 \text{ as } g'=\beta''\le 0}.\end{aligned}$$

Thus $I_1 < 0$. $\square$

Let us now return to (4.9) and deal with the second term in the right-hand side. We have, with account of (2.3) and (2.7),

$$\frac{\eta^2\gamma(\xi)^2}{1+\eta\gamma(\xi)} \le \begin{cases} \eta^2\gamma(\xi)^2 & \text{if } \gamma(\xi)\ge 0,\\ \dfrac{1}{1-A}\eta^2\gamma(\xi)^2 & \text{if } \gamma(\xi)<0 \end{cases} \le \max\left\{1, \frac{1}{1-A}\right\}\eta^2\gamma(\xi)^2,$$

and so

$$\alpha_\bullet^2\int_{-r}^{r}\int_0^\delta \frac{\eta^2\gamma(\xi)^2}{1+\eta\gamma(\xi)}\mathrm{e}^{-\beta(\eta)}|\phi_*(\eta)|^2\,\mathrm{d}\eta\,\mathrm{d}\xi \le I_2\int_{-r}^{r}\gamma(\xi)^2\,\mathrm{d}\xi,$$

where

$$I_2 := \max\left\{1, \frac{1}{1-A}\right\}\alpha_\bullet^2\int_0^\delta \eta^2\mathrm{e}^{-\beta(\eta)}|\phi_*(\eta)|^2\,\mathrm{d}\eta \ge 0. \tag{4.14}$$

Thus, as $\gamma(\xi)$ vanishes for $|\xi|>R$, we have

$$D_\gamma = I_1\int\gamma(\xi)\,\mathrm{d}\xi + I_2\int\gamma(\xi)^2\,\mathrm{d}\xi = (-I_1)\left(C_\beta\int\gamma(\xi)^2\,\mathrm{d}\xi - \int\gamma(\xi)\,\mathrm{d}\xi\right),$$

where

$$C_\beta = \frac{I_2}{-I_1} = \frac{\max\left\{1, \dfrac{1}{1-A}\right\}\alpha_\bullet^2\displaystyle\int_0^\delta \eta^2\mathrm{e}^{-\beta(\eta)}|\phi_*(\eta)|^2\,\mathrm{d}\eta}{\displaystyle\int_0^\delta \eta\mathrm{e}^{-\beta(\eta)}\left(|\phi_*'(\eta)|^2-\alpha_\bullet^2|\phi_*(\eta)|^2\right)\mathrm{d}\eta} \tag{4.15}$$

is a positive constant.

As soon as (4.2) holds, D_γ is negative, and so (4.5) holds. This proves Theorem 4.1.

Finally, it is sufficient to note that (4.3) implies $\int\gamma(\xi)^2\,\mathrm{d}\xi < 2Rc_{\beta,R}\int\gamma(\xi)\,\mathrm{d}\xi$, which proves Corollary 4.2.

5. Conclusions. It has been shown that a trapped mode is possible in the model presented here. To increase confidence that such modes exist on real coasts further work is clearly required to demonstrate that this mode is not an artifact of the modelling assumptions. However these assumptions are the usual ones for the simple theory of CSWs and extensions to include stratification and more realistic boundary conditions have not in general contradicted them [LBMy]. The result here suggests that it would be of interest to compare low-frequency velocity records in the neighborhood of capes with those on nearby straight coasts to determine whether there is indeed enhanced energy at the cape. Both the above endeavors are being pursued.

REFERENCES

[AsPaVa] A. Aslanyan, L. Parnovski, and D. Vassiliev, *Complex resonances in acoustic waveguides*, Quart. J. Mech. Appl. Math., 53 (2000), pp. 429–447.

[Da] E. B. Davies, *Spectral Theory and Differential Operators*, Cambridge University Press, Cambridge, UK, 1995.

[DaPa] E. B. Davies and L. Parnovski, *Trapped modes in acoustic waveguides*, Quart. J. Mech. Appl. Math., 51 (1998), pp. 477–492.

[DiKr] J. Dittrich and J. Kříž, *Curved planar quantum wires with Dirichlet and Neumann boundary conditions*, J. Phys. A, 35 (2002), pp. L269–L275.

[DuEx] P. Duclos and P. Exner, *Curvature-induced bound states in quantum waveguides in two and three dimensions*, Rev. Math. Phys., 7 (1995), pp. 73–102.

[EvLeVa] D. V. Evans, M. Levitin, and D. Vassiliev, *Existence theorems for trapped modes*, J. Fluid Mech., 261 (1994), pp. 21–31.

[ExSe] P. Exner and P. Šeba, *Bound states in curved quantum waveguides*, J. Math. Phys., 30 (1989), pp. 2574–2580.

[Ha] B. V. Hamon, *Continental shelf waves and the effect of atmospheric pressure and wind stress on sea level*, J. Geophys. Res., 71 (1966), pp. 2883–2893.

[Jo1] E. R. Johnson, *A conformal-mapping technique for topographic-wave problems—semi-infinite channels and elongated basins*, J. Fluid Mech., 177 (1987), pp. 395–405.

[Jo2] E. R. Johnson, *Topographic waves in open domains.* I. *Boundary conditions and frequency estimates*, J. Fluid Mech., 200 (1989), pp. 69–76.

[Jo3] E. R. Johnson, *Low-frequency barotropic scattering on a shelf bordering an ocean*, J. Phys. Oceanogr., 21 (1991), pp. 720–727.

[KrKr] D. Krejčiřík and J. Kříž, *On the spectrum of curved planar waveguides*, Publ. Res. Inst. Math. Sci., 41 (2005), pp. 757–791.

[KrTA] D. Krejčiřík and R. Tiedra de Aldecoa, *The nature of the essential spectrum in curved quantum waveguides*, J. Phys. A, 37 (2004), pp. 5449–5466.

[La] H. Lamb, *Hydrodynamics*, 6th ed., Cambridge University Press, Cambridge, UK, 1932.

[LBMy] P. H. LeBlond and L. A. Mysak, *Waves in the Ocean*, Elsevier, Amsterdam, 1978.

[Pe] J. Pedlosky, *Geophysical Fluid Dynamics*, Springer, Berlin, 1986.

[StHu1] T. F. Stocker and K. Hutter, *One-dimensional models for topographic Rossby waves in elongated basins on the f-plane*, J. Fluid Mech., 170 (1986), pp. 435–459.

[StHu2] T. F. Stocker and K. Hutter, *Topographic waves in rectangular basins*, J. Fluid Mech., 185 (1987), pp. 107–120.

[StHuSaTrZa] K. Stocker, K. Hutter, G. Salvade, J. Trösch, and F. Zamboni, *Observations and analysis of internal seiches in the southern basin of Lake of Lugano*, Ann. Geophys. B, 5 (1987), pp. 553–568.

[StJo1] T. F. Stocker and E. R. Johnson, *Topographic waves in open domains.* II. *Bay modes and resonances*, J. Fluid Mech., 200 (1989), pp. 77–93.

[StJo2] T. F. Stocker and E. R. Johnson, *The trapping and scattering of topographic waves by estuaries and headlands*, J. Fluid Mech., 222 (1991), pp. 501–524.

[Tr] J. Trösch, *Finite element calculation of topographic waves in lakes*, in Proceedings of the 4th International Conference on Applied Numerical Modeling (Tainan, Taiwan, 1984), Science and Technology Series 63, H. M. Hsia, Y. L. Chou, S. Y. Wang, and S. J. Hsieh, eds., American Astronautical Society, 1986, pp. 307–311.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1482–1521

GLOBAL BIFURCATION THEORY OF DEEP-WATER WAVES WITH VORTICITY*

VERA MIKYOUNG HUR†

Abstract. The classical deep-water wave problem is to find a periodic traveling wave with a free surface of infinite depth. The main result is the construction of a global connected set of rotational solutions for a general class of vorticities. Each nontrivial solution on the continuum has a wave profile symmetric around the crests and monotone between crest and trough.

The problem is formulated as a nonlinear elliptic boundary value problem in an unbounded domain with a parameter. The analysis is based on generalized degree theory and the global theory of bifurcation. The unboundedness of the domain renders consideration of approximate problems with stronger compactness properties.

Key words. water waves, vorticity, nonlinear elliptic, Leray–Schauder degree, bifurcation

AMS subject classifications. 76B15, 35J60, 47J15, 76B03

DOI. 10.1137/040621168

1. Introduction.

The classical deep-water wave problem is to find a periodic traveling wave with a free surface of infinite depth. A two-dimensional incompressible inviscid fluid, at rest at great depths, occupies the domain bounded by a free surface, where the pressure balances out with the constant atmospheric pressure. The main result is to construct a global connected set of nontrivial solutions to the full Euler equations for positive monotone vorticities. The methods involved are the global theory of bifurcation which has its foundation in topological degree theory, maximum principles and the Schauder theory for nonlinear second-order elliptic partial differential equations.

Irrotational solutions to the deep-water wave problem are referred to as *Stokes waves*, for which existence theory has been studied in a number of works. The first rigorous mathematical treatment is the construction, following Levi-Civita [26] and Nekrasov [30], of small-amplitude waves by a series expansion method. The global theory of existence began with Krasovskii [24] who used a degree-theoretic argument to obtain finite-amplitude waves. Subsequently, with the invention by Rabinowitz [33] of abstract global bifurcation theory, Keady and Norbury [22] inferred that these solutions form a global connected set. Amick [3], Amick, Fraenkel, and Toland [4], and McLeod [29] refined this result further to prove that this continuum contains in its closure Stokes waves of extreme form with stagnation at their crests. A survey on the existence theory for Stokes waves is to be found in [39].

The existence theory for *rotational* waves is, on the other hand, much less complete. Gerstner [14, 6] found an explicit formula for a family of trochoidal deep-water waves for a particular nonzero vorticity early in 1802. However, an extensive existence theory for a general class of vorticities appeared much later in the work of Dubreil-Jacotin [12], who constructed small-amplitude waves by a series expansion method. This result was later improved by Goyon [16] and Zeidler [42]. Recently, Constantin

*Received by the editors December 20, 2004; accepted for publication (in revised form) June 20, 2005; published electronically January 12, 2006.
http://www.siam.org/journals/sima/37-5/62116.html
†Department of Mathematics, Brown University, Providence, RI 02912 (verahur@math.brown.edu, http://www.math.brown.edu/~verahur).

and Strauss [9, 10] employed generalized degree theory [18, 23] as adapted for nonlinear elliptic boundary value problems and global bifurcation theory [33] to obtain a global connected set of rotational waves of finite depth. The purpose of the present work is to obtain an analogous result for the case of infinite depth. An important difference from [10] lies in the lack of compact embedding properties of Ascoli type due to the unboundedness of the physical domain. Under the additional restriction that the vorticity distribution is positive and monotone, the set of nontrivial solutions acquires the desired compactness with the exploitation of maximum principles.

At the beginning of section 3.1, the unknown fluid region is conveniently mapped onto a fixed semi-infinite strip in the plane by exchanging the roles of the stream function and the vertical coordinate. The depth below the free surface is regarded as a solution to a quasi-linear elliptic partial differential equation with a nonlinear oblique boundary condition. The problem is successively formulated as an operator equation of the form $F(\lambda, w) = 0$; λ is a real parameter and the "nontrivial perturbation" w from the trivial shear current belongs to the Banach space, denoted by X, of $C^{3+\alpha}$ Hölder continuous functions which vanish asymptotically at the infinite bottom. The equation is singular in the sense that the operator is not Fredholm, and thus topological degree theory/global bifurcation theory may not be applied directly. This difficulty is overcome by studying a sequence of approximate problems $F^\epsilon(\lambda, w) = 0$ for $\epsilon > 0$, where F^ϵ is a nonlinear Fredholm operator.

Section 4 gives a detailed global bifurcation analysis for the approximate problems. The first step is to linearize F^ϵ around the trivial solution and to find a sequence of bifurcation points through each of which a local curve of nontrivial solutions emanates. This is an application of the local bifurcation theorem due to Crandall and Rabinowitz [11]. Next, several properties of F^ϵ including the properness and the Fredholm property of the linearized operator are established. The generalized degree theory developed in [18, 23] and the global bifurcation theorem [33] apply at once, and after further detailed analysis based on the maximum principle and its sharp form at corner points [35], a global continuum of nontrivial solutions is obtained to each approximate problem.

The central part of this work is section 5.1 on the existence theory for the singular problem. We employ the development by Rabinowitz of the global theory of bifurcation to conclude that the continuum of the nontrivial solutions to $F(\lambda, w) = 0$ either is unbounded in $\mathbb{R} \times X$ or intersects the boundary of the admissible set where the boundary value problem becomes degenerate. An important observation here is that the positiveness and monotonicity of the vorticity enable us to obtain the uniform decay of solutions at infinity and hence the compactness of the solution set. In the following subsection, with the use of the regularity estimates for quasi-linear elliptic partial differential equations, the alternative that the size of w increases unboundedly in the $C^{3+\alpha}$ norm along the continuum reduces to the unboundedness of either w or its first-order derivatives in the maximum norm.

Finally, section 5.3 is the examination of the remaining alternatives to prove that our global continuum contains a sequence of nontrivial solutions to the deep-water wave problem such that along a sequence on the continuum either the relative flow speed at the wave troughs becomes arbitrarily large or the relative flow speed somewhere on the free surface becomes arbitrarily close to zero (stagnation). In the case of irrotational flows, the existence of a limiting wave with stagnation at its crests, where the wave profile has sharp cusps with the containing angle of $2\pi/3$, was conjectured by Stokes [37] and proved by Amick, Fraenkel, and Toland [4].

2. Formulation and the main result.

2.1. The deep-water wave problem in physical variables. The gravity water wave problem concerns the motion of an incompressible inviscid fluid with a free surface acted on only by gravity. Consider a two-dimensional flow which at time t occupies the region in the (x, y)-plane bounded by a free surface $y = \eta(t; x)$. In the fluid region $\{(x, y) : y < \eta(t; x)\}$, the velocity field $(u(t; x, y), v(t; x, y))$ and the pressure $P(t; x, y)$ satisfy the following Euler equations:

$$\begin{aligned} u_t + uu_x + vu_y &= -P_x, \\ v_t + uv_x + vv_y &= -P_y - g, \\ u_x + v_y &= 0. \end{aligned}$$

Here g denotes the gravitational constant of acceleration. The flow is supposed to be rotational and characterized by the vorticity $\omega = v_x - u_y$.

The dynamic and kinematic boundary conditions hold on the free surface $y = \eta(t; x)$:

$$P = P_0 \quad \text{and} \quad v = \eta_t + u\eta_x,$$

where P_0 is the constant atmospheric pressure. The boundary condition at the infinite bottom

$$(u, v) \to (0, 0) \quad \text{as } y \to -\infty \text{ uniformly for } x \in \mathbb{R}$$

expresses the fact [19] that the fluid at great depths is practically at rest. Our bottom boundary condition precludes wave trains on water of infinite depth, numerically calculated in [36, 40], which exhibit a nonphysical asymptotic behavior at the infinite bottom.

The steady water wave problem is then to find solutions for which the wave profile, the velocity, and the pressure have space-time dependence $(x - ct, y)$, where $c > 0$ is the speed of wave propagation. In the frame of reference moving with speed c, the velocity field $(u(x, y), v(x, y))$ and the pressure $P(x, y)$ of the steady flow satisfy

$$(u - c)u_x + vu_y = -P_x, \tag{2.1}$$

$$(u - c)v_x + vv_y = -P_y - g, \tag{2.2}$$

$$u_x + v_y = 0 \tag{2.3}$$

in the stationary region $\{(x, y) : y < \eta(x)\}$. The boundary conditions are

$$P = P_0 \quad \text{and} \quad v = (u - c)\eta_x \qquad \text{on } y = \eta(x), \tag{2.4}$$

$$(u, v) \to (-c, 0) \quad \text{as } y \to -\infty \text{ uniformly for } x \in \mathbb{R}. \tag{2.5}$$

Equations (2.1)–(2.5) are supplemented with the following periodicity and symmetry conditions:

$$\begin{aligned} \eta(-x) = \ \eta(x) \quad &= \eta(x + L), \\ u(-x, y) = u(x, y) &= u(x + L, y), \\ -v(-x, y) = v(x, y) &= v(x + L, y), \end{aligned} \tag{2.6}$$

where $L > 0$ is the wavelength.

Our objective here is to find nontrivial solutions to (2.1)–(2.6) for a general class of vorticities. If the flow is *irrotational*, i.e., $\omega \equiv 0$, the solutions are referred to as *Stokes waves*, for which existence theory has been studied in a number of works. A survey on the existence theory for Stokes waves is to be found in [39].

In this setting, L and c are considered as parameters whose values form part of the solution. The wavelength L of the deep-water wave problem is independent of other physical parameters, and hence is taken to be 2π for simplicity in what follows.

Two distinct features of problem (2.1)–(2.6) are that the free surface is *not* known a priori and that the physical domain is *unbounded*.

2.2. The vorticity-stream formulation. We define the (relative) *stream function* $\psi(x,y)$ by

$$\psi_x = -v, \qquad \psi_y = u - c. \tag{2.7}$$

This reduces (2.1)–(2.6) to a stationary elliptic boundary value problem.

The Poisson equation

$$-\Delta\psi = v_x - u_y = \omega$$

follows from (2.7). Note that ψ is 2π periodic in the x-variable. Indeed, $\int_0^{2\pi} v(s,y)ds$ is independent of y and therefore is zero by (2.5). The kinematic boundary condition allows us to normalize ψ so that $\psi = 0$ on the free surface. The boundary condition at the infinite bottom becomes $\nabla\psi(x,y) = (\psi_x, \psi_y) \to (0, -c)$ as $y \to -\infty$ uniformly for x.

The vorticity equation $(u-c)\omega_x + v\omega_y = 0$, compared to (2.7), states that ω is a function of ψ at least locally away from a stagnation point, a point where $u = c$ and $v = 0$. Under the restriction that the vorticity is nonnegative and nonincreasing with the depth, i.e., $\omega(x,y) \geq 0$ and $\omega_y(x,y) \geq 0$, one can show that $u - c < 0$ everywhere in the fluid; see Lemma 2.1. Experimental evidence [28] indicates that for wave patterns which are not near the spilling or breaking state, the speed of wave propagation is in general considerably larger than the horizontal velocity of any water particle.

We shall show later in section 3.1 that $u - c < 0$ guarantees that

$$-\Delta\psi = \gamma(\psi)$$

for some function γ throughout the fluid. The *vorticity function* γ measures the strength of the vorticity.

Assumptions on γ. For some $\alpha \in (0,1)$ a constant, $\gamma \in C^{1+\alpha}([0,\infty))$ is nonincreasing along with depth, $\gamma'(s) \leq 0$ for $s \in [0,\infty)$, and $\gamma(s) \in O(s^{-2-2l})$ as $s \to \infty$ for some $l > 0$.

Our assumptions imply that the vorticity is nonnegative, $\gamma(s) \geq 0$ for $s \in [0,\infty)$, and that ω vanishes in the limit as $y \to -\infty$.

Let us define

$$\Gamma(p) = \int_p^0 \gamma(-s)ds \tag{2.8}$$

and

$$\Gamma_\infty = \int_{-\infty}^0 \gamma(-s)ds.$$

From the equations of motion follows Bernoulli's law, which states that the quantity

$$E = \tfrac{1}{2}(\psi_x^2 + \psi_y^2) + gy + P + \Gamma(-\psi)$$

is constant throughout the fluid. The sum of the first four terms in the expression of E is the total mechanical energy of the flow; $\frac{1}{2}(\psi_x^2 + \psi_y^2)$ is the kinetic energy, gy is the gravitational potential energy, and P is the energy of fluid pressures. In view of Bernoulli's law, the dynamic boundary condition takes the form that

$$Q = \psi_x^2(x, \eta(x)) + \psi_y^2(x, \eta(x)) + 2g\eta(x)$$

is independent of x. The quantity $\frac{1}{2}Q$, i.e., E evaluated at the free surface, is called the *hydraulic head*. We remark that E and Q may vary along continua of solutions.

In summary, there results the following formulation of the deep-water wave problem, equivalent to (2.1)–(2.6). Let

$$\mathcal{D}_\eta = \{(x, y) : x \in \mathbb{R},\ \ y < \eta(x)\} \quad \text{and} \quad \mathcal{S}_\eta = \{(x, \eta(x)) : x \in \mathbb{R}\}.$$

For a parameter $c > 0$, there exist $\eta(x)$ and $\psi(x, y)$ such that η is even and 2π periodic, ψ is even and 2π periodic in the x-variable, and

$$-\Delta\psi = \gamma(\psi) \quad \text{in } \mathcal{D}_\eta, \tag{2.9a}$$

$$\psi = 0 \quad \text{on } \mathcal{S}_\eta, \qquad \psi > 0 \quad \text{in } \mathcal{D}_\eta, \tag{2.9b}$$

$$|\nabla\psi(x, \eta(x))|^2 + 2g\eta(x) = Q \quad \text{for } x \in \mathbb{R}, \tag{2.9c}$$

$$\nabla\psi(x, y) \to (0, -c) \quad \text{as } y \to -\infty \text{ uniformly for } x \in \mathbb{R}. \tag{2.9d}$$

We locate the origin of Cartesian coordinates at one of the wave crests (where η attains its maximum over one period), i.e., $\eta(0) = 0$. With this convention, $Q = (c - u(0, 0))^2 > 0$ is the square of the relative flow speed at the crest.

Under the assumptions on the vorticity, the exponential decay estimates of solutions follow as applications of the maximum principle.

LEMMA 2.1. *Assume that $\omega(x, y) \geq 0$ and $\omega_y(x, y) \geq 0$. Let $\eta(x)$ and $\psi(x, y)$ be a solution pair of* (2.9) *corresponding to the solution triple $(u - c, v, \eta)$ of* (2.1)–(2.6) *via the definition* (2.7).

(a) *$\psi_y = u - c < 0$ on $\overline{\mathcal{D}}_\eta$. Therefore, $\omega = \gamma(\psi)$ throughout the fluid region for some function γ. Note that γ satisfies the assumptions described above.*

(b) *Assume further that $\psi_x(x, \eta(x)) > 0$ for $x \in (-\pi, 0)$. Then $\psi_x(x, y) > 0$ on $\{(x, y) \in \overline{\mathcal{D}}_\eta : x \in (-\pi, 0)\}$, and $\psi_x(x, y)$ and $\psi_{xx}(0, y)$ decay exponentially as $y \to -\infty$ uniformly for x. More precisely,*

$$|\psi_x(x, y)| < Ae^y \qquad \text{on } (-\pi, \pi) \times (-\infty, \eta(\pm\pi)), \tag{2.10}$$

$$|\psi_{xx}(0, y)| < Be^y \qquad \text{for } y \in (-\infty, \eta(\pm\pi)), \tag{2.11}$$

where $A, B > 0$ depend only on $\eta(\pm\pi)$.

Remark 2.2. Assertion (a) justifies the existence of the vorticity function γ and supports a detailed global bifurcation analysis; see sections 3 and 4, respectively. Assertion (b) is important for compactness reasons later in Lemmas 4.6 and 5.1.

An easy consequence of the monotonicity of ψ_x and ψ_y is that the wave profile decreases monotonically from crest to trough. Constantin and Escher [7] proved that the symmetry of the wave profile is not a hypothesis but rather a conclusion when the wave profile is monotone between crest and trough and the vorticity is positive and decreases with depth.

Proof. (a) Since $\psi = 0$ on $\mathcal{S}_\eta$ and $\psi \to \infty$ as $y \to -\infty$, the subharmonic function ψ is positive in $\mathcal{D}_\eta$. It therefore attains its minimum on $\mathcal{S}_\eta$, and by the Hopf boundary point lemma $\psi_y < 0$ on $\mathcal{S}_\eta$.

Note that

$$\Delta\psi_y + \omega_y(x, y) = 0$$

and that ψ_y tends to $-c < 0$ as $y \to -\infty$. The maximum principle then implies that $\psi_y < 0$ on $\overline{\mathcal{D}}_\eta$.

(b) Note that

$$\Delta\psi_x + \gamma'(\psi)\psi_x = 0$$

and that $\psi_x \to 0$ as $y \to -\infty$ uniformly for x. Since $\psi_x(x, \eta(x)) > 0$ for $x \in (-\pi, 0)$, by the maximum principle, $\psi_x(x, y) > 0$ for $(x, y) \in \overline{\mathcal{D}}_\eta$ and $x \in (-\pi, 0)$.

For the second assertion, let

$$D = (-\pi, \pi) \times (-\infty, \eta(\pm\pi)) \quad \text{and} \quad D^- = \{(x, y) \in D : x \in (-\pi, 0)\}.$$

For $A > 0$ we define a C^2-function on D^- by

$$W(x, y) = \psi_x(x, y) - Ae^y \sin x.$$

Choose A sufficiently large so that $W(x, y) < 0$ on $(0, \pi) \times \{\eta(\pm\pi)\}$. Such an A depends only on $\eta(\pm\pi)$; see [7]. A simple calculation yields

$$\Delta W + \gamma'(\psi)W = -A\gamma'(\psi)e^y \sin x \geq 0 \quad \text{in } D^-.$$

By oddness and periodicity of ψ_x, we infer that $W(-\pi, y) = 0 = W(0, y)$ for $y \leq \eta(-\pi)$. Since $W(x, y) \to 0$ as $y \to -\infty$ uniformly for x, the maximum principle ensures $W(x, y) < 0$ in D^-. A similar consideration leads to an analogous inequality on $D \cap \{x \in (0, \pi)\}$, and (2.10) follows at once.

Finally, from the classical gradient estimate [15, page 37] for Poisson's equation, it follows that

$$\begin{aligned} |\psi_{xx}(0, y)| &\leq \frac{2}{\pi} \sup_D |\psi_x(x, y)| + \frac{\pi}{2} \sup_D |\gamma'(\psi)||\psi_x(x, y)| \\ &\leq Be^y \end{aligned}$$

for $y < \eta(\pm\pi) - \pi$, where $B > 0$ depends only on $\eta(\pm\pi)$. This proves (2.11). $\square$

2.3. The main result. For a positive integer k and a constant $\alpha \in (0, 1)$, a domain $D \subset \mathbb{R}^2$ is called a $C^{k+\alpha}$ domain if each point on its boundary ∂D has a neighborhood in which ∂D is the graph of a $C^{k+\alpha}$ function. Given a $C^{k+\alpha}$ domain $D \subset \mathbb{R}^2$ (not necessarily bounded), we define

$$C^{k+\alpha}_{per}(\overline{D}) = \{\, w \in C^{k+\alpha}(\overline{D}) : w \text{ is even and } 2\pi \text{ periodic in the } x\text{-variable} \,\}, \tag{2.12}$$

where $C^{k+\alpha}(\overline{D})$ is a Hölder space with the norm as in [2, Chapter II]:

$$\|w\|_{C^{k+\alpha}(\overline{D})} = \sum_{|\beta|=0}^{k} \max_{\overline{D}} |\partial^\beta w(x, y)| + \sup_{|\beta|=k} \sup_{\substack{(x,y)\neq(\tilde{x},\tilde{y}) \\ \overline{D}}} \frac{|\partial^\beta w(x, y) - \partial^\beta w(\tilde{x}, \tilde{y})|}{\sqrt{(x - \tilde{x})^2 + (y - \tilde{y})^2}^{\,\alpha}}.$$

This notation is extended in an obvious way to the case when $\alpha = 0$ and to functions of a single variable.

Our main result is the following theorem on the existence of nontrivial deep-water waves with vorticity.

THEOREM 2.3 (main theorem). *Suppose that the vorticity function $\gamma \in C^{1+\alpha}([0,\infty))$ for some $\alpha \in (0,1)$ satisfies $\gamma'(s) \leq 0$ for $s \in [0,\infty)$ and $\gamma \in O(s^{-2-2l})$ as $s \to \infty$ for some $l > 0$. Suppose furthermore that*

$$\Gamma_\infty = \int_{-\infty}^{0} \gamma(-s)ds < \frac{1}{2} \min \left(\left(\frac{g}{2}\right)^{2/3}, \left(\frac{g}{2}\right)^2 \right). \tag{2.13}$$

Consider the deep-water wave problem (2.1)–(2.6). There exists a connected set $\mathcal{C}$ of solution triples $(u-c, v, \eta)$ in the space $C^{2+\alpha}_{per}(\overline{\mathcal{D}}_\eta) \times C^{2+\alpha}_{per}(\overline{\mathcal{D}}_\eta) \times C^{3+\alpha}_{per}(\mathbb{R})$ such that

(i) *the continuum $\mathcal{C}$ contains a trivial shear flow with $v \equiv 0$ (under the flat surface $\eta \equiv 0$); and*

(ii) *there is a sequence of solution triples $\{(u_j - c_j, v_j, \eta_j)\} \subset \mathcal{C}$, for which*

$$\text{either} \quad \lim_{j\to\infty} (c_j - u_j(\pm\pi, \eta_j(\pm\pi))) = \infty \quad \text{or} \quad \lim_{j\to\infty} \min_{\mathcal{S}_{\eta_j}} (c_j - u_j) = 0.$$

Each nontrivial solution triple $(u-c, v, \eta) \in \mathcal{C}$ enjoys the following properties:

(iii) *η has a single maximum (crest) and a single minimum (trough) per wavelength; we say the crest occurs at $x = 0$;*

(iv) *the wave profile decreases monotonically from crest to trough, i.e., $\eta'(x) < 0$ for $x \in (0,\pi)$;*

(v) *the relative horizontal velocity is negative, $u - c < 0$, throughout the fluid; and*

(vi) *a water particle located at (x,y) with $x \in (0,\pi)$ has positive vertical velocity $v > 0$.*

In addition, if $\gamma(0)$ is sufficiently small, (ii) *can be replaced by*

(ii′) *there is a sequence of solution triples $\{(u_j - c_j, v_j, \eta_j)\} \subset \mathcal{C}$, for which*

$$\text{either} \quad \lim_{j\to\infty} (c_j - u_j(\pm\pi, \eta_j(\pm\pi))) = \infty \quad \text{or} \quad \lim_{j\to\infty} (c_j - u_j(0,0)) = 0.$$

The smallness condition (2.13) guarantees the local bifurcation of approximate problems; see Lemma 4.1.

Theorem 2.3 presents two alternatives in (ii). When the first alternative holds, either the speed of wave propagation or the wave amplitude becomes unboundedly large; see Remark 5.9. The second alternative means that the continuum contains waves with horizontal particle velocity somewhere on the free surface arbitrarily close to the speed of wave propagation. In other words, there is a region of almost stagnant fluid on the free surface, a region which is carried along by the traveling wave.

If $\gamma(0)$ is small enough so that the relative flow speed $c-u$ is nondecreasing from crest to trough (and it hence attains its minimum on the surface at the crest), then the crest is the point of almost stagnation. A precise condition is given in (5.23).

Our conclusion, in the case of zero vorticity, partly recovers the well-known result for Stokes waves [39] that the connected continuum contains a limiting wave with stagnation at the wave crests.

3. Reformulation: Reduction to an operator equation. We now turn to establishing the existence of solutions to (2.1)–(2.6). On account of the unknown surface, we use the coordinate transformation devised by Dubreil-Jacotin [12] to reformulate the problem as a boundary value problem in a fixed domain.

3.1. Change of independent variables. We begin by the observation that ψ is constant on the free surface and decreases along with y as a function of y; i.e., the y-coordinate is a single-valued function of ψ for each fixed x. This suggests the introduction of new independent variables

$$q = x \quad \text{and} \quad p = -\psi(x, y),$$

which map the fluid region of one period $\{(x, y) \in \mathcal{D}_\eta : x \in (-\pi, \pi)\}$ to a fixed semi-infinite strip $(-\pi, \pi) \times (-\infty, 0)$ in the (q, p)-plane, the free surface $\{(x, \eta(x)) : x \in (-\pi, \pi)\}$ to $(-\pi, \pi) \times \{0\}$. Let

$$R = \{(q, p) : -\pi < q < \pi,\, p < 0\} \quad \text{and} \quad T = \{(q, 0) : -\pi < q < \pi\}.$$

The depth function $h(q, p) = y - Q/2g$ replaces the dependent variables. An explicit calculation shows that

$$h_q = \frac{v}{u - c}, \qquad h_p = \frac{1}{c - u}. \tag{3.1}$$

From $u - c < 0$ throughout the fluid it follows that ω is a single-valued function of p. Indeed,

$$\partial_q \omega = \left(\partial_x + \frac{h_q}{h_p}\partial_p\right)\omega = \left(\partial_x - \frac{v}{c - u}\partial_y\right)\omega = 0.$$

We say $\omega = \gamma(-p)$.

The vorticity-stream formulation (2.9) is accordingly reformulated as an elliptic boundary value problem in a fixed domain:

$$(1 + h_q^2)h_{pp} - 2h_p h_q h_{pq} + h_p^2 h_{qq} = -\gamma(-p)h_p^3 \qquad \text{in } R, \tag{3.2a}$$

$$1 + 2ghh_p^2 + h_q^2 = 0 \qquad \text{on } T, \tag{3.2b}$$

$$\nabla h = (h_q, h_p) \to (0, \tfrac{1}{c}) \quad \text{as } p \to -\infty \qquad \text{uniformly for } q, \tag{3.2c}$$

with h even and 2π periodic in the q-variable.

The above formulation is equivalent to (2.1)–(2.6), as is presented in the next lemma.

LEMMA 3.1. *Suppose that $h \in C^{3+\alpha}_{per}(\overline{R})$ is a solution of* (3.2). *There corresponds to h a solution triple $(u - c, v, \eta)$ of* (2.1)–(2.6) *in the space $C^{2+\alpha}_{per}(\overline{\mathcal{D}}_\eta) \times C^{2+\alpha}_{per}(\overline{\mathcal{D}}_\eta) \times C^{3+\alpha}_{per}(\mathbb{R})$.*

The subscript "*per*" for a domain in the (q, p)-plane refers to evenness and periodicity in the q-variable.

Proof. The proof is similar to that of [10, Lemma 2.1], and therefore we will not carry out every detail. First, it is observed from the partial differential equation (3.2a) that $\gamma \in C^{1+\alpha}([0, \infty))$.

Suppose for the moment that $h \in C^2_{per}(\overline{R})$ is a solution of (3.2). We define $C^1_{per}(\overline{R})$ functions

$$F(q, p) = \frac{1}{h_p(q, p)}, \qquad G(q, p) = -\frac{h_q(q, p)}{h_p(q, p)}. \tag{3.3}$$

The wave profile is given by $\eta(x) = h(x, 0) + Q/2g$ for $x \in [-\pi, \pi]$ and is extended to the entire real line as an even and 2π periodic function. Clearly, $\eta \in C^2_{per}(\mathbb{R})$.

In order to recover the relative stream function ψ, we consider the ordinary differential equation

$$\begin{cases} \frac{d}{dy}\psi(x,y) = -F(x, -\psi(x,y)), \\ \psi(x, \eta(x)) = 0, \end{cases} \tag{3.4}$$

where $x \in [-\pi, \pi]$ is fixed. Since F is of C^1, from the standard theory of ordinary differential equations it follows that there exists a unique local solution $\psi(x, y)$ to (3.4). Furthermore, the solution exists for all $(-\infty, \eta(x)]$. Indeed, $F \geq \delta > 0$ for some δ over $\overline{R}$, and thus $F(q, p)$ defines a nondegenerate complete vector field on $p \in (-\infty, 0)$. Clearly, $\psi \in C^2_{per}(\overline{\mathcal{D}}_\eta)$.

It is straightforward to show that $\eta(x)$ and $\psi(x, y)$ constructed above are solutions to (2.9), completely analogous to the finite-depth case. One can repeat the calculation in [10, Lemma 2.1] to show that $\psi_x(x, y) = -G(x, -\psi(x, y))$. The boundary condition at infinity

$$\nabla\psi(x,y) \to (0, -c) \quad \text{as } y \to -\infty \text{ uniformly for } x \in \mathbb{R}$$

is fulfilled by (3.2c) and (3.3).

The relative velocity components are successively defined by

$$u(x,y) - c = -F(x, -\psi(x,y)), \qquad v(x,y) = G(x, -\psi(x,y)),$$

which are in $C^2_{per}(\overline{\mathcal{D}}_\eta)$ and solve (2.1)–(2.6). Finally, if h is $C^{3+\alpha}_{per}(\overline{R})$ for $\alpha \in (0, 1)$, then, by construction, $(u - c, v, \eta)$ belongs to the function space $C^{2+\alpha}_{per}(\overline{\mathcal{D}}_\eta) \times C^{2+\alpha}_{per}(\overline{\mathcal{D}}_\eta) \times C^{3+\alpha}_{per}(\mathbb{R})$. $\square$

As in section 2, let

$$\Gamma(p) = \int_p^0 \gamma(-s)ds, \qquad \Gamma_\infty = \int_{-\infty}^0 \gamma(-s)ds.$$

LEMMA 3.2 (trivial flows). *For each $\lambda \in (2\Gamma_\infty, \infty)$ the system* (3.2) *has a solution*

$$H(p) = H(p; \lambda) = \int_0^p \frac{ds}{\sqrt{\lambda - 2\Gamma(s)}} - \frac{\lambda}{2g}, \tag{3.5}$$

which corresponds to a parallel shear flow in the horizontal direction under the flat surface $\eta \equiv 0$.

Proof. These solutions do not depend on q, and thus (3.2a) reduces to $H'' = -\gamma(-p)(H')^3$. Here and elsewhere the prime denotes differentiation with respect to p. Solutions to this ordinary differential equation are

$$H'(p) = (\lambda - 2\Gamma(p))^{-1/2},$$

which are defined for $\lambda > 2\Gamma_\infty$. The formula (3.5) then follows from the boundary condition on top, $1 + 2gH(0)(H'(0))^2 = 0$.

The boundary condition at infinity, $H'(p) \to \frac{1}{c}$ as $p \to -\infty$, relates the bifurcation parameter λ with the wave speed c:

$$\lambda = c^2 + 2\Gamma_\infty. \qquad \square$$

Note 3.3. Let us denote

$$a(\lambda) = a(p;\lambda) = \sqrt{\lambda - 2\Gamma(p)}.$$

The derivatives of H can be expressed in terms of a:

$$H'(p) = a^{-1}(p;\lambda), \qquad H''(p) = -\gamma(-p)a^{-3}(p;\lambda).$$

Note that $a(\lambda)$ is bounded for each $\lambda > 2\Gamma_\infty$.

In order to establish the existence of nontrivial solutions to (3.2) via bifurcation theory, we need to formulate the problem as an abstract operator equation $F(\lambda, w) = 0$, where w belongs to a Banach space. However, a set of functions with the condition (3.2c) at infinity does not form a linear space. For this reason, we introduce the "nontrivial perturbation" $w(q,p)$ of the depth function $h(q,p)$ from that of a trivial flow $H(p)$ defined as

$$h(q,p) = H(p) + w(q,p). \tag{3.6}$$

Problem (2.1)–(2.6) is ultimately formulated as follows:

$$\begin{aligned}(1+w_q^2)w_{pp} - 2(a^{-1}(\lambda)+w_p)w_q w_{pq} + (a^{-1}(\lambda)+w_p)^2 w_{qq}&\\ + \gamma(-p)(a^{-1}(\lambda)+w_p)^3 - \gamma(-p)a^{-3}(\lambda)(1+w_q^2) = 0 &\qquad \text{in } R,\end{aligned} \tag{3.7a}$$

$$1 + (2gw-\lambda)(\lambda^{-1/2}+w_p)^2 + w_q^2 = 0 \qquad \text{on } T. \tag{3.7b}$$

Here w is required to be even and 2π periodic in the q-variable. The principal point is that ∇w vanishes asymptotically:

$$\nabla w = (w_q, w_p) \to 0 \quad \text{as } p \to -\infty \text{ uniformly for } q. \tag{3.7c}$$

3.2. The operator equation and its approximation. We introduce the function spaces. Let

$$\begin{aligned} X &= \{\, w \in C^{3+\alpha}_{per}(\overline{R}) : \partial^\beta w \in o(1) \text{ as } p \to -\infty,\ |\beta| \le 3 \text{ uniformly for } q \,\},\\ Y_1 &= \{\, w \in C^{1+\alpha}_{per}(\overline{R}) : \partial^\beta w \in o(1) \text{ as } p \to -\infty,\ |\beta| \le 1 \text{ uniformly for } q \,\},\end{aligned}$$

and $Y_2 = C^{2+\alpha}_{per}(T)$. Let $Y = Y_1 \times Y_2$ with the product topology. We equip X and Y with the Hölder norms (thus rendering them Banach spaces):

$$\begin{aligned} \|\cdot\|_X &:= \|\cdot\|_{C^{3+\alpha}(\overline{R})}\,,\\ \|\cdot\|_Y &:= \|\cdot\|_{Y_1} + \|\cdot\|_{Y_2}\,,\end{aligned}$$

where $\|\cdot\|_{Y_1} = \|\cdot\|_{C^{1+\alpha}(\overline{R})}$ and $\|\cdot\|_{Y_2} = \|\cdot\|_{C^{2+\alpha}(T)}$. Here the Hölder norms of functions on the unbounded domain $\overline{R}$ take supremum values over the entire domain (see [2, Chapter II]):

$$\|w\|_{C^{k+\alpha}(\overline{R})} = \sum_{|\beta|=0}^{k} \max_{\overline{R}} |\partial^\beta w(q,p)| + \sup_{|\beta|=k}\ \sup_{\substack{(q,p)\neq(\tilde q,\tilde p)\\ \overline{R}}} \frac{|\partial^\beta w(q,p) - \partial^\beta w(\tilde q,\tilde p)|}{\sqrt{(q-\tilde q)^2 + (p-\tilde p)^2}^{\,\alpha}}.$$

Let $Z = C^0_{per}(\overline{R})$ have the usual maximum norm $\|\cdot\|_Z = \|\cdot\|_{C^0(\overline{R})}$.

A nonlinear differential operator

$$F(\lambda, w) = (F_1(\lambda, w), F_2(\lambda, w)) : \mathbb{R} \times X \to Y$$

is defined by

$$\begin{aligned} F_1(\lambda, w) = {}& (1 + w_q^2)w_{pp} - 2(a^{-1}(\lambda) + w_p)w_q w_{pq} + (a^{-1}(\lambda) + w_p)^2 w_{qq} \\ & + \gamma(-p)(a^{-1}(\lambda) + w_p)^3 - \gamma(-p)a^{-3}(\lambda)(1 + w_q^2), \end{aligned} \tag{3.8}$$

$$F_2(\lambda, w) = 1 + (2gw - \lambda)(\lambda^{-1/2} + w_p)^2 + w_q^2\big|_T. \tag{3.9}$$

The operator form of the deep-water wave problem is then to find a nontrivial solution $(\lambda, w) \in \mathbb{R} \times X$ to

$$F(\lambda, w) = 0. \tag{3.10}$$

The existence theory "in the large" for the finite-depth case [9, 10] follows as an application to F of generalized degree theory [18, 23] and abstract global bifurcation theory [33]. For the infinite-depth case, however, the presence of the continuous spectrum due to the unboundedness of the physical domain renders this application much less than routine.

Denoted by $F_w(\lambda, w)$ is the Fréchet derivative of F in its second argument at $(\lambda, w) \in \mathbb{R} \times X$. A straightforward calculation yields

$$F_w(\lambda, w) = (A(\lambda, w), B(\lambda, w)),$$

where

$$\begin{aligned} A(\lambda, w)[\phi] = {}& (1 + w_q^2)\phi_{pp} - 2(a^{-1}(\lambda) + w_p)w_q\phi_{pq} + (a^{-1}(\lambda) + w_p)^2\phi_{qq} \\ & + \left(-2w_q w_{pq} + 2(a^{-1}(\lambda) + w_p)w_{qq} + 3\gamma(-p)(a^{-1}(\lambda) + w_p)^2\right)\phi_p \\ & + \left(2w_q w_{pp} - 2(a^{-1}(\lambda) + w_p)w_{pq} - 2\gamma(-p)a^{-3}(\lambda)w_q\right)\phi_q, \\ B(\lambda, w)[\phi] = {}& 2(2gw - \lambda)(\lambda^{-1/2} + w_p)\phi_p + 2w_q\phi_q + 2g(\lambda^{-1/2} + w_p)^2\phi\big|_T \end{aligned}$$

for $\phi \in X$.

The linear operator $(A(\lambda, w), B(\lambda, w))$ defines the "limiting" operator

$$\begin{cases} A_\infty(\lambda)[\phi] = \phi_{pp} + (\lambda - 2\Gamma_\infty)^{-1}\phi_{qq}, \\ B_\infty(\lambda)[\phi] = -2\lambda^{1/2}\phi_p + 2g\lambda^{-1}\phi\big|_T \end{cases}$$

for $\phi \in X$, which is obtained by substituting each coefficient function by its limit as $p \to -\infty$. Unfortunately, this operator is *not* semi-Fredholm as its spectrum consists of only the continuous spectrum $(-\infty, 0]$. In particular, $F_w(\lambda, w) : X \to Y$ is *not* a Fredholm operator of index zero, and thus topological degree theory may not be directly applied.

This difficulty can be overcome by studying a sequence of approximate problems:

$$F^\epsilon(\lambda, w) := \left(F_1(\lambda, w) - \epsilon(a^{-1}(\lambda) + w_p)^3 w, F_2(\lambda, w)\right) = 0, \tag{3.11}$$

where $\epsilon > 0$. The properness of F^ϵ and the Fredholm property of its linearization will follow from the fact that the limiting problem

$$\left(A_\infty(\lambda) - \epsilon(\lambda - 2\Gamma_\infty)^{-3/2}I, B_\infty(\lambda)\right)[\phi] = 0$$

admits only the zero solution; see Lemmas 4.4 and 4.6.

Remark 3.4. Construction of a topological degree for general nonlinear elliptic operators has been studied in a number of works (see [18, 23, 41] and the references therein) under the condition that the nonlinear elliptic operator is proper and the linearization is a Fredholm operator of index zero. The Fredholm property of a linear elliptic operator in an unbounded domain is related to the uniform decay of the solutions at infinity. Indeed, a linear elliptic operator in an unbounded domain is semi-Fredholm if and only if the limiting problem does not have nonzero solutions (see [41, Theorem 2.15]).

A simple calculation yields that

$$F^\epsilon_w(\lambda, w) = (A^\epsilon(\lambda, w), B(\lambda, w)),$$

where

$$\begin{aligned} A(\lambda,w)[\phi] &= (1+w_q^2)\phi_{pp} - 2(a^{-1}(\lambda)+w_p)w_q\phi_{pq} + (a^{-1}(\lambda)+w_p)^2\phi_{qq} \\ &\quad + \big(-2w_q w_{pq} + 2(a^{-1}(\lambda)+w_p)w_{qq} \\ &\qquad + 3\gamma(-p)(a^{-1}(\lambda)+w_p)^2 + 3\epsilon(a^{-1}(\lambda)+w_p)^2 w\big)\phi_p \\ &\quad + \left(2w_q w_{pp} - 2(a^{-1}(\lambda)+w_p)w_{pq} - 2\gamma(-p)a^{-3}(\lambda)w_q\right)\phi_q \\ &\quad - \epsilon(a^{-1}(\lambda)+w_p)^3\phi, \end{aligned} \tag{3.12}$$

$$B(\lambda,w)[\phi] = 2(2gw-\lambda)(\lambda^{-1/2}+w_p)\phi_p + 2w_q\phi_q + 2g(\lambda^{-1/2}+w_p)^2\phi\big|_T \tag{3.13}$$

for $\phi \in X$. Similarly,

$$F^\epsilon_\lambda(\lambda, w) = (F_{1\lambda}(\lambda, w), F_{2\lambda}(\lambda, w)),$$

where

$$\begin{aligned} F_{1\lambda}(\lambda,w)[\mu] &= \mu a^{-3}(\lambda)\big(w_q w_{pq} - (a^{-1}(\lambda)+w_p)w_{qq} \\ &\qquad - \tfrac{3}{2}\gamma(-p)(a^{-1}(\lambda)+w_p) \\ &\qquad + \tfrac{3}{2}\gamma(-p)a^{-2}(\lambda)(1+w_q^2) - 3\epsilon(a^{-1}(\lambda)+w_p)^2 w\big), \\ F_{2\lambda}(\lambda,w)[\mu] &= -\mu(\lambda^{-1/2}+w_p)\left(\lambda^{-1/2}+w_p+(2gw-\lambda)\lambda^{3/2}\right)\big|_T. \end{aligned}$$

Since F^ϵ_w and F^ϵ_λ are both continuous, $F^\epsilon : \mathbb{R}\times X \to Y$ is continuously differentiable. Indeed, the smoothness of γ ensures that F^ϵ is at least twice continuously Fréchet differentiable.

4. Existence theory for approximate problems. Our goal in this section is to construct for each ϵ a global connected set of nontrivial solutions to (3.11) via generalized degree theory and abstract global bifurcation theory.

4.1. The linearized approximate problem. The linearization of F^ϵ about the trivial solution $(\lambda, 0)$ is $F^\epsilon_w(\lambda, 0) = (A^\epsilon(\lambda, 0), B(\lambda, 0))$, where

$$\begin{aligned} A^\epsilon(\lambda,0)[\phi] &= \phi_{pp} + a^{-2}(\lambda)\phi_{qq} + 3\gamma(-p)a^{-2}(\lambda)\phi_p - \epsilon a^{-3}(\lambda)\phi \\ &= a^{-3}(\lambda)(a^3(\lambda)\phi_p)_p + a^{-2}(\lambda)\phi_{qq} - \epsilon a^{-3}(\lambda)\phi, \\ B(\lambda,0)[\phi] &= -2\lambda^{1/2}\phi_p + 2g\lambda^{-1}\phi\big|_T \end{aligned}$$

for $\phi \in X$. A necessary condition for bifurcation at a trivial solution $(\lambda, 0)$ is that

$$F^\epsilon_w(\lambda, 0) : X \to Y \text{ is not injective;}$$

equivalently, the following problem of self-adjoint form

$$(a^3(\lambda)\phi_p)_p + (a(\lambda)\phi_q)_q - \epsilon\phi = 0 \quad \text{in } R, \tag{4.1a}$$

$$\lambda^{3/2}\phi_p = g\phi \qquad \text{on } T \tag{4.1b}$$

admits a nontrivial solution in X.

LEMMA 4.1. *Suppose that $\gamma \in C^{1+\alpha}([0,\infty))$ for $\alpha \in (0,1)$ satisfies the smallness condition* (2.13).

(a) *For each $0 \le \epsilon < \epsilon_0$ and $0 < \epsilon_0 \le g$ fixed, there exist $\lambda^\epsilon \in (2\Gamma_\infty, g + 2\Gamma_\infty]$ and a nontrivial solution $\phi^\epsilon \in X$ to* (4.1).

(b) *For a sequence $\epsilon_j \to 0$ as $j \to \infty$, $\{\lambda^{\epsilon_j}\}$ converges to $\lambda^0 \in (2\Gamma_\infty, g+2\Gamma_\infty]$ as $j \to \infty$.*

Proof. (a) For each $0 \le \epsilon < \epsilon_0$ we will look for a solution of the form $\phi^\epsilon(q,p) = \Phi^\epsilon(p)\cos kq$ with $k \ge 0$ an integer; Φ^ϵ will solve the ordinary differential equation $-(a^3(\lambda)\Phi')' + \epsilon\Phi = -k^2 a(\lambda)\Phi$.

Let us define for $\lambda \in (2\Gamma_\infty, \infty)$ an ordinary differential operator $L^\epsilon v = -(a^3(\lambda)v')' + \epsilon v$. We consider the (singular) Sturm–Liouville problem

$$\begin{cases} L^\epsilon v = \mu(\lambda)a(\lambda)v & \text{for } p \in (-\infty, 0), \\ \lambda^{3/2}v'(0) = gv(0), \\ v, v' \to 0 & \text{as } p \to -\infty. \end{cases} \tag{4.2}$$

It is known that L^ϵ with the boundary conditions above has the essential spectrum $[\epsilon(\lambda - 2\Gamma_\infty)^3, \infty)$. We define for $0 \le \epsilon < \epsilon_0$

$$\begin{aligned} R^\epsilon(\lambda) &= R^\epsilon(v;\lambda) \\ &= \frac{-gv^2(0) + \int_{-\infty}^0 a^3(\lambda)(v')^2dp + \epsilon\int_{-\infty}^0 v^2dp}{\int_{-\infty}^0 a(\lambda)v^2dp} \end{aligned} \tag{4.3}$$

and

$$\Lambda^\epsilon(\lambda) = \inf\{\, R^\epsilon(v;\lambda) : v \in H^1((-\infty,0)) \text{ and } v \not\equiv 0\,\}. \tag{4.4}$$

A straightforward calculation shows that $R^\epsilon(\lambda)$ is bounded from below for each λ. More precisely, $R^\epsilon(\lambda) > -g^2(\lambda - 2\Gamma_\infty)^{-2}$. The Rayleigh principle then asserts that the complementary interval $(-\infty, \epsilon)$ contains a generalized eigenvalue μ (such that $L^\epsilon v = \mu a(\lambda)v$ for some $v \not\equiv 0$) if and only if $\Lambda^\epsilon(\lambda) < 0$. Furthermore, such $\Lambda^\epsilon(\lambda)$ is a simple eigenvalue.

Our aim is to find a λ^ϵ such that $\Lambda^\epsilon(\lambda^\epsilon) = -k^2$. There may be multiple solutions, for instance, corresponding to different values of k. Rather, here we restrict ourselves to finding one for $k = 1$. Note that Λ^ϵ is a C^1-function of λ.

First, for $\lambda \in [g + 2\Gamma_\infty, \infty)$ the inequality

$$\begin{aligned} \int_{-\infty}^0 \big(a(\lambda)v^2 + a^3(\lambda)(v')^2 + \epsilon v^2\big)dp &\ge \sqrt{g}\int_{-\infty}^0 \big(v^2 + g(v')^2\big)dp \\ &\ge 2g\int_{-\infty}^0 vv'dp = gv^2(0) \end{aligned}$$

holds for every $v \in H^1((-\infty,0))$. This in turn implies $R^\epsilon(\lambda) \geq -1$, and therefore $\Lambda^\epsilon(\lambda) \geq -1$. Next, provided that (2.13) holds, one can show that

$$\begin{aligned}\Lambda^\epsilon(2\Gamma_\infty) &\leq R^\epsilon(e^p;2\Gamma_\infty)\\ &= \frac{-g+\int_{-\infty}^0 a^3(2\Gamma_\infty)e^{2p}\,dp+\epsilon\int_{-\infty}^0 e^{2p}\,dp}{\int_{-\infty}^0 a(2\Gamma_\infty)e^{2p}\,dp}\\ &< \frac{-g+\frac{1}{2}\int_{-\infty}^0\sqrt{2\Gamma_\infty-2\Gamma(p)}\,dp+\frac{\epsilon}{2}}{\int_{-\infty}^0 e^{2p}\sqrt{2\Gamma_\infty-2\Gamma(p)}\,dp} < -1.\end{aligned}$$

By continuity, there exists $\lambda^\epsilon \in (2\Gamma_\infty, g+2\Gamma_\infty]$ such that $\Lambda^\epsilon(\lambda^\epsilon) = -1$.

Now we consider an eigenfunction $\Phi^\epsilon \in H^1((-\infty,0))$ to (4.2) with $\lambda = \lambda^\epsilon$ (and $k=1$). It follows from standard regularity theory that Φ^ϵ is indeed smooth. Note that $\sqrt{\lambda^\epsilon - 2\Gamma_\infty} \leq a(\lambda^\epsilon) \leq \sqrt{\lambda^\epsilon}$. The comparison theorem for second-order ordinary differential equations [8, Chapter 8] then asserts that $\Phi^\epsilon > 0$ and decays exponentially:

$$|\Phi^\epsilon(p)| \leq A\exp\bigl(p(\lambda-2\Gamma_\infty)^{-1/2}\bigr)$$

for some constant $A>0$. Therefore, $\phi^\epsilon = \Phi^\epsilon(p)\cos q$ is in X.

(b) Since $\{\lambda^\epsilon\}$ forms a bounded sequence in $\mathbb{R}$, there are a sequence $\epsilon_j \to 0$ as $j\to\infty$ and a subsequence $\{\lambda^{\epsilon_j}\}$ which converges to λ^0 in $\mathbb{R}$ as $j\to\infty$. By continuity, $\Lambda^0(\lambda^0) = -1$.

Since $R^0(\lambda) \geq -1$ for $\lambda \in [g+2\Gamma_\infty,\infty)$, it must hold that $\lambda^0 \leq g+2\Gamma_\infty$. On the other hand, $\lambda^0 > 2\Gamma_\infty$ since

$$\Lambda^0(2\Gamma_\infty) \leq R^0(e^p;2\Gamma_\infty) < R^\epsilon(e^p;2\Gamma_\infty) < -1.$$

Therefore, $\lambda^0 \in (2\Gamma_\infty, g+2\Gamma_\infty]$. This completes the proof. ∎

It follows as an application of local bifurcation theorem from a simple eigenvalue due to Crandall and Rabinowitz [11] that for each $0<\epsilon<\epsilon_0$ there emanates from $(\lambda^\epsilon,0)$ a local curve in $\mathbb{R}\times X$ of solutions to (3.11). The detailed analysis is carried out in Appendix A.

PROPOSITION 4.2 (local bifurcation for approximate problems). *For each $0<\epsilon<\epsilon_0$, there exist $s_0>0$ and a C^1-curve $\mathcal{C}^\epsilon_{loc}$ in $\mathbb{R}\times X$ of the form $(\lambda(s),w(s))$ such that each $(\lambda(s),w(s))$ for $|s|<s_0$ is a solution to (3.11) with $(\lambda(0),w(0)) = (\lambda^\epsilon,0)$.*

Recorded in the next lemma is the Fredholm property of $F^\epsilon_w(\lambda^\epsilon,0)$, which is a special case of Lemma 4.4.

LEMMA 4.3 (Fredholm property at the bifurcation point). *For each $0<\epsilon<\epsilon_0$, the linear operator $F^\epsilon_w(\lambda^\epsilon,0) = (A^\epsilon(\lambda^\epsilon,0),B(\lambda^\epsilon,0)) : X\to Y$ is a Fredholm operator of index zero.*

Proof. Let us denote the limiting operator of $(A^\epsilon(\lambda^\epsilon,0),B(\lambda^\epsilon,0))$ of constant coefficients by

$$\begin{aligned}A_0[\phi] &= \phi_{pp} + (\lambda^\epsilon-2\Gamma_\infty)^{-1}\phi_{qq} - \epsilon(\lambda^\epsilon-2\Gamma_\infty)^{-3/2}\phi,\\ B_0[\phi] &= B(\lambda^\epsilon,0) = -2(\lambda^\epsilon)^{1/2}\phi_p + 2g(\lambda^\epsilon)^{-1}\phi\big|_T\end{aligned}$$

for $\phi\in X$. Consider the one-parameter family of linear operators

$$(A_t,B_t) : X\to Y \qquad \text{for } t\in[0,1],$$

where

$$A_t = (1-t)A_0 + tA^\epsilon(\lambda^\epsilon, 0),$$
$$B_t = B(\lambda^\epsilon, 0).$$

Below we prove that (A_t, B_t) is semi-Fredholm for all $t \in [0,1]$; i.e., (A_t, B_t) has a closed range and finite-dimensional kernel. On the other hand, $(A_0, B_0) : X \to Y$ is bijective (see [25, Chapter 3], for instance). In particular, it is a Fredholm operator of index zero. Therefore, by the homotopy invariance of Fredholm index [21, Chapter 4, Theorems 2.23 and 5.17], $(A_1, B_1) = (A^\epsilon(\lambda^\epsilon, 0), B(\lambda^\epsilon, 0))$ is also a Fredholm operator of index zero.

Note that (A_t, B_t) satisfies the complementing condition on the interval $[0,1]$. Indeed, A_t is uniformly elliptic for each $t \in [0,1]$, whose coefficients are bounded in $C^{2+\alpha}(\overline{R})$; $B(\lambda^\epsilon, 0)$ is uniformly oblique. Therefore, for each $t \in [0,1]$ the following Schauder estimate [2, Theorem 7.3] holds:

$$\|\phi\|_X \leq C(\|A_t[\phi]\|_{Y_1} + \|B_t[\phi]\|_{Y_2} + \|\phi\|_Z) \tag{4.5}$$

for all $\phi \in X$, where $C > 0$ is independent of ϕ. We remark that the Schauder estimate (4.5) is valid even in an unbounded domain. However, the compact embedding property of Ascoli type is not available. In particular, Z is not compactly embedded in X. (The semi-Fredholm property of an elliptic operator in a bounded domain is a direct consequence of the Schauder estimate (4.5) and the compact embedding properties of Hölder spaces; see [32], for instance.)

In order to obtain the semi-Fredholm property of (A_t, B_t), we need to show that (A_t, B_t) is proper. In general, for linear operators these two concepts are equivalent; see [34, 41]. For $t \in [0,1]$ fixed, consider the equation

$$(A_t, B_t)[w_j] = f_j,$$

where $\{w_j\}$ is a bounded sequence in X and $\{f_j\}$ converges to f, say, in Y. There exists a function $w \in X$ such that $w_j \to w$ in $C^3_{per}(R')$ and $(A_t, B_t)[w] = f$ for any compact subset R' of $\overline{R}$.

We claim that $w_j \to w$ in $C^0_{per}(\overline{R})$. Suppose this convergence does not take place. Accordingly, we may choose a sequence $\{(q_j, p_j)\} \subset \overline{R}$ with $p_j \to -\infty$ such that

$$|w_j(q_j, p_j) - w(q_j, p_j)| \geq c > 0,$$

where $c > 0$ is fixed. Consider the functions $v_j(q,p) = w_j(q, p+p_j) - w(q, p+p_j)$, the operators (A_t, B_t) with the coefficient function $a(p; \lambda^\epsilon)$ shifted by p_j, and the shifted domains $\{(q,p) \in R : p + p_j < 0\}$. Passing to the limit we find a nontrivial limiting function $v_0 \in C^0(\overline{R})$ which is a solution to the limiting problem

$$(A_0, B_0)[v_0] = 0.$$

This contradicts the unique solvability of (A_0, B_0) and proves the claim.

From the convergence of $w_j \to w$ in $C^0_{per}(\overline{R})$, the Schauder estimate (4.5), and the convergence $f_j \to f$ in Y, it follows that $\{w_j\}$ is a Cauchy sequence in X. That is, (A_t, B_t) is proper. Subsequently, (A_t, B_t) is semi-Fredholm for each $t \in [0,1]$. This completes the proof. $\square$

4.2. Preliminary results for degree theory. In order to define a topological degree for F^ϵ we need some properties to establish including the properness and the Fredholm property of the linearized operator.

In the following discussion, we define for $\delta > 0$ the open set

$$\mathcal{O}_\delta = \Big\{ (\lambda, w) \in \mathbb{R} \times X \text{ which satisfies} \\ \lambda > 2\Gamma_\infty + \delta,\ a^{-1}(\lambda) + w_p > \delta \text{ in } R,\ w < \frac{2\lambda - \delta}{4g} \text{ on } T \Big\}. \tag{4.6}$$

Throughout this subsection ϵ and δ are fixed positive constants.

Let us recall that $F^\epsilon : \mathcal{O}_\delta \to Y$ is of class C^2 and that its Fréchet derivative in its second argument is of the form

$$F^\epsilon_w(\lambda, w) = (A^\epsilon(\lambda, w), B(\lambda, w)) \in Y_1 \times Y_2,$$

where

$$\begin{aligned} A^\epsilon(\lambda, w)[\phi] &= (1 + w_q^2)\phi_{pp} - 2(a^{-1}(\lambda) + w_p)w_q\phi_{pq} + (a^{-1}(\lambda) + w_p)^2\phi_{qq} \\ &\quad + \big(-2w_q w_{pq} + 2(a^{-1}(\lambda) + w_p)w_{qq} \\ &\qquad + 3\gamma(-p)(a^{-1}(\lambda) + w_p)^2 + 3\epsilon(a^{-1}(\lambda) + w_p)^2 w\big)\phi_p \\ &\quad + \big(2w_q w_{pp} - 2(a^{-1}(\lambda) + w_p)w_{pq} - 2\gamma(-p)a^{-3}(\lambda)w_q\big)\phi_q \\ &\quad - \epsilon(a^{-1}(\lambda) + w_p)^3\phi, \end{aligned} \tag{4.7}$$

$$B(\lambda, w)[\phi] = 2(2gw - \lambda)(\lambda^{-1/2} + w_p)\phi_p + 2w_q\phi_q + 2g(\lambda^{-1/2} + w_p)^2\phi\big|_T \tag{4.8}$$

for $\phi \in X$. The principal parts of operators $A^\epsilon(\lambda, w)$ and $B(\lambda, w)$ are denoted by

$$\tilde{A}(\lambda, w)[\phi] = (1 + w_q^2)\phi_{pp} - 2(a^{-1}(\lambda) + w_p)w_q\phi_{pq} + (a^{-1}(\lambda) + w_p)^2\phi_{qq}, \tag{4.9}$$

$$\tilde{B}(\lambda, w)[\phi] = 2(2gw - \lambda)(\lambda^{-1/2} + w_p)\phi_p + 2w_q\phi_q\big|_T, \tag{4.10}$$

respectively. Note that for each $(\lambda, w) \in \mathcal{O}_\delta$ the differential operator $A^\epsilon(\lambda, w)$ is uniformly elliptic with coefficients bounded in $C^{2+\alpha}(\overline{R})$; the coefficients of the principal part satisfy

$$(1 + w_q^2)\xi_1^2 - 2(a^{-1}(\lambda) + w_p)w_q\xi_1\xi_2 + (a^{-1}(\lambda) + w_p)^2\xi_2^2 \geq 4\delta^2(\xi_1^2 + \xi_2^2)$$

for all $(\xi_1, \xi_2) \in \mathbb{R}^2$. Also note that for each $(\lambda, w) \in \mathcal{O}_\delta$ the boundary operator $B(\lambda, w)$ is uniformly oblique in the sense that it is bounded away from being tangential; the coefficient of ϕ_p is nonzero and

$$|2(2gw - \lambda)(\lambda^{-1/2} + w_p)| > \delta^2 > 0 \qquad \text{on } T.$$

Therefore, $F^\epsilon_w(\lambda, w)$ satisfies the complementing condition for each $(\lambda, w) \in \mathcal{O}_\delta$ and the following Schauder estimate [2] holds:

$$\|\phi\|_X \leq C(\|A^\epsilon(\lambda, w)[\phi]\|_{Y_1} + \|B(\lambda, w)[\phi]\|_{Y_2} + \|\phi\|_Z) \tag{4.11}$$

for all $\phi \in X$, where $C > 0$ is independent of ϕ.

An important observation is that the limiting problem

$$\big(A_\infty(\lambda) - \epsilon(\lambda - 2\Gamma_\infty)^{-3/2}I, B_\infty(\lambda)\big)[\phi] = 0$$

admits only the zero solution.

Our first preliminary result is the Fredholm property of $F^\epsilon_w(\lambda, w)$.

LEMMA 4.4 (Fredholm property). *For each* $(\lambda, w) \in \mathcal{O}_\delta$ *the linear operator* $F^\epsilon_w(\lambda, w) = (A^\epsilon(\lambda, w), B(\lambda, w)) : X \to Y$ *is a Fredholm operator of index zero.*

Proof. The argument in the proof of Lemma 4.3 shows that $F^\epsilon_w(\lambda, w)$ is semi-Fredholm. This involves the Schauder estimate (4.11) and the fact that the limiting problem

$$\big(A_\infty(\lambda) - \epsilon(\lambda - 2\Gamma_\infty)^{-3/2} I, B_\infty(\lambda)\big)[\phi] = 0$$

has only the zero solution.

On the other hand, the result of Lemma 4.3 is that $F^\epsilon_w(\lambda^\epsilon, 0)$ is a Fredholm operator of index zero. Since $\mathcal{O}_\delta$ is connected, the assertion then follows from the continuity of the Fredholm index [21, Chapter 4, Theorem 5.17]. $\square$

For our next preliminary result, we need several notations to define. The domain of the operator $A^\epsilon(\lambda, w)$ is defined by

$$D(A^\epsilon(\lambda, w)) = \{\, w \in X : B(\lambda, w) = 0 \,\}. \tag{4.12}$$

Note that $A^\epsilon(\lambda, w)$ restricted to $D(A^\epsilon(\lambda, w))$ is closed in Y_1. The spectrum of $A^\epsilon(\lambda, w)$ is defined by

$$\begin{aligned} \sigma(\lambda, w) = \{\, \mu \in \mathbb{C} \text{ such that} \\ (A^\epsilon - \mu I) : D(A^\epsilon(\lambda, w)) \to Y_1 \text{ is not an isomorphism} \,\}. \end{aligned} \tag{4.13}$$

Here A^ϵ, $D(A^\epsilon(\lambda, w))$, and Y_1 are complexified in the natural way.

LEMMA 4.5 (spectral properties). *For* $M > 0$ *and* $(\lambda, w) \in \mathcal{O}_\delta$ *with* $|\lambda| + \|w\|_X \leq M$, *there exist positive constants* s, C_1, C_2 *such that*

$$C_1 \|\phi\|_X \leq |\mu|^{\alpha/2} \|(A^\epsilon(\lambda, w) - \mu I)[\phi]\|_{Y_1} + |\mu|^{(\alpha+1)/2} \|B(\lambda, w)[\phi]\|_{Y_2} \tag{4.14}$$

for all $\phi \in X$ *and for all* $\mu \in \mathbb{C}$ *satisfying* $|\arg(\mu)| \leq \pi/2 + s$ *and* $|\mu| \geq C_2$.

In addition,

(a) $\sigma(\lambda, w)$ *possesses only finitely many eigenvalues in the sector* $|\arg(\mu)| \leq \pi/2 + s$, *each of which has a finite multiplicity; and*

(b) *the boundary operator* $B(\lambda, w) : X \to Y_2$ *is surjective.*

Proof. The proof uses the method due to Agmon [1]. See [18, Proposition 4.4] for one in a Hölder-space setting.

We introduce the differential operator $A^\epsilon(\lambda, w) + e^{i\theta} \frac{\partial^2}{\partial t^2}$ on the cylinder $\Omega := R \times (-\infty, \infty)$, where $\theta = \arg(\mu)$. Note that the operator is elliptic if $|\theta| \leq \pi/2 + s$ for small $s > 0$. The boundary condition on $\partial\Omega = \partial R \times (-\infty, \infty)$ is given by $B(\lambda, w) = 0$.

It is known that $\big(A^\epsilon(\lambda, w) + e^{i\theta} \frac{\partial^2}{\partial t^2}, B(\lambda, w)\big)$ satisfies the complementing condition. Accordingly, one may write down the Schauder estimate: there are $s, C_1 > 0$ such that

$$\begin{aligned} C_1 \|e(t)\phi\|_{C^{3+\alpha}(\overline{\Omega}_1)} \leq\ & \|\zeta(t) e(t) (A^\epsilon - \mu I)[\phi]\|_{C^{1+\alpha}(\overline{\Omega}_2)} \\ & + \|\zeta(t) e(t) B[\phi]\|_{C^{2+\alpha}(\partial\Omega_1)} \\ & + |\mu|^{1/2} \|\zeta_1(t) e(t) \phi\|_{C^0(\overline{\Omega}_2)} \end{aligned} \tag{4.15}$$

for all $\phi \in X$ and for all $\mu \in \mathbb{C}$ with $|arg(\mu)| \leq \pi/2 + s$. Here $\overline{\Omega}_r = \overline{R} \times [-r, r]$, $\partial\Omega_r = \partial\Omega \times [-r, r]$ for some $r > 0$, and $e(t) = \exp(i|\mu|^{1/2} t)$; ζ, ζ_1 are smooth functions with support on the interval $[-2, 2]$ and $\zeta, \zeta_1 \equiv 1$ on $[-1, 1]$.

An explicit calculation shows that

$$\|\zeta_1(t)e(t)\phi\|_{C^0(\Omega_2)} \leq C_3|\mu|^{\alpha/2}\|\phi\|_{Y_1}$$

for $|\mu| > C_2$, where $C_2, C_3 > 0$ are independent of ϕ. Analogous estimates are valid for other terms on the right side of (4.15); see [18, Proposition 4.4] for the details. These estimates together with (4.15) yield the spectral estimate (4.14).

Our task now is to verify (a) and (b). The argument in the proof of Lemma 4.4 indicates that $(A^\epsilon - \mu I, B)$ is a Fredholm operator of index zero for all $\mu \in \mathbb{C}$ with $|arg(\mu)| \leq \pi/2 + s$. On the other hand, the estimate (4.14) dictates that for any $|\mu| > C_2$ the operator $A^\epsilon - \mu I$ restricted to $D(A^\epsilon(\lambda, w))$ has trivial kernel. Assertion (a) then follows from known properties of Fredholm operators [20]. In particular, $(A^\epsilon - \mu I, B) : X \to Y_1 \times Y_2$ is bijective for $\mu > C_2$, which proves (b). ∎

Our last preliminary result is the properness of F^ϵ.

LEMMA 4.6 (properness). *The nonlinear elliptic operator F^ϵ is (locally) proper on $\overline{\mathcal{O}}_\delta$; i.e, $(F^\epsilon)^{-1}(K) \cap \overline{D}$ is compact for each bounded set $D \subset \overline{\mathcal{O}}_\delta$ and compact set $K \subset Y$.*

Proof. Let $\{(\lambda_j, w_j)\} \subset D \subset \overline{\mathcal{O}_\delta}$ be a bounded sequence and let $\{(y_{1j}, y_{2j})\} \subset K \subset Y$ converge to (y_1, y_2) in $Y_1 \times Y_2$ satisfying

$$F^\epsilon(\lambda_j, w_j) = (y_{1j}, y_{2j}) \qquad \text{for } j = 1, 2, \dots.$$

Our goal is to find a subsequence of $\{(\lambda_j, w_j)\}$ which converges in $\mathbb{R} \times X$. For the sake of simplicity, we may assume that $\lambda_j \to \lambda$ in $\mathbb{R}$ for some $\lambda \in \mathbb{R}$ and that $w_j \to w$ in $C^3_{per}(R')$ for some $w \in X$ for any compact subset R' of $\overline{R}$. By continuity, $F^\epsilon(\lambda, w) = (y_1, y_2)$.

It is convenient to decompose F^ϵ as

$$\begin{aligned} F^\epsilon_1(\lambda, w) &= \tilde{A}(\lambda, w)[w] + f_1(\lambda, w) - \epsilon a^{-3}(\lambda) w, \\ F_2(\lambda, w) &= \tilde{B}(\lambda, w)[w] + f_2(\lambda, w). \end{aligned} \tag{4.16}$$

Here $\tilde{A}(\lambda, w)$ and $\tilde{B}(\lambda, w)$ are defined in (4.9) and (4.10); $f_1(\lambda, w)$ is a quartic polynomial expression of w, w_q and w_p, and $f_2(\lambda, w) = \lambda^{-1/2}(2gw - \lambda)(\lambda^{-1/2} + w_p)$.

We claim that $w_j \to w$ in $C^0_{per}(\overline{R})$. Otherwise, there would be a sequence $\{(q_j, p_j)\} \subset \overline{R}$ with $p_j \to -\infty$ such that

$$|w_j(q_j, p_j) - w(q_j, p_j)| \geq c > 0.$$

As is done in Lemma 4.3, consider the shifted domains $\{(q, p) \in R : p < p_j\}$, the operator $(\tilde{A}(\lambda_j, \tilde{w}_j, \tilde{a}) - \epsilon\tilde{a}^{-3}(\lambda_j)I, \tilde{B}(\lambda_j, \tilde{w}_j, \tilde{a}))$ with coefficients shifted by p_j, i.e., $\tilde{w}_j(q, p) = w_j(q, p + p_j)$ and $\tilde{a}(p; \lambda) = a(p + p_j; \lambda)$, and the functions $v_j(q, p) = w_j(q, p + p_j) - w(q, p + p_j)$. Passing to subsequences we conclude that there exist a limiting domain $R_\infty = R$, a limiting operator $(A_\infty - \epsilon(\lambda - 2\Gamma_\infty)^{-3/2}I, B_\infty)$, and a nonzero limiting function $v_\infty \in C^0(\overline{R})$ such that

$$\big(A_\infty - \epsilon(\lambda - 2\Gamma_\infty)^{-3/2}I, B_\infty\big)[v_\infty] = 0.$$

However, the limiting problem has only the zero solution. This proves by contradiction the uniform convergence of $\{w_j\}$ to w.

Since $\{w_j\}$ is uniformly bounded in X, an interpolation inequality (see [15, Lemma 6.32] and [25, Theorem 3.2.1]) asserts that $w_j \to w$ in $C^{k'+\alpha'}(\overline{R})$ for $k' + \alpha' < 3 + \alpha$. Indeed, for any $s > 0$ there exists a constant $C = C(s) > 0$ such that

$$\|w_j - w_k\|_{C^{k'+\alpha'}(\overline{R})} \leq s\|w_j - w_k\|_{C^{3+\alpha}(\overline{R})} + C\|w_j - w_k\|_{C^0(\overline{R})}. \tag{4.17}$$

Next, in order to obtain the convergence in $C^{3+\alpha}(\overline{R})$, Schauder theory is employed. Since $\{(\lambda_j, w_j)\} \subset \overline{\mathcal{O}}_\delta$ is bounded, the uniform Schauder estimates [2] yield

$$\|\phi\|_X \leq C(\|\tilde{A}(\lambda_j, w_j)[\phi]\|_{Y_1} + \|\tilde{B}(\lambda_j, w_j)[\phi]\|_{Y_2} + \|\phi\|_Z)$$

for all $\phi \in X$, where $C > 0$ is independent of ϕ and the index j. In particular,

$$\|w_j - w_k\|_X \leq C(\|\tilde{A}(\lambda_j, w_j)[w_j - w_k]\|_{Y_1} + \|\tilde{B}(\lambda_j, w_j)[w_j - w_k]\|_{Y_2} + \|w_j - w_k\|_Z). \tag{4.18}$$

Already shown is that the last term on the right side of the inequality approaches zero as $j, k \to \infty$. By virtue of the decomposition (4.16), the first two terms on the right side of (4.18) may be written as

$$\begin{aligned}\tilde{A}(\lambda_j, w_j)[w_j - w_k] &= y_{1j} - y_{1k} - \big(\tilde{A}(\lambda_j, w_j) - \tilde{A}(\lambda_k, w_k)\big)w_k \\ &\quad - \big(f_1(\lambda_j, w_j) - f_1(\lambda_k, w_k)\big) \\ &\quad + \epsilon a^{-3}{}_{(\lambda_j)}(w_j - w_k) + \epsilon\big(a^{-3}{}_{(\lambda_j)} - a^{-3}{}_{(\lambda_k)}\big)w_k, \\ \tilde{B}(\lambda_j, w_j)[w_j - w_k] &= y_{2j} - y_{2k} - \big(\tilde{B}(\lambda_j, w_j) - \tilde{B}(\lambda_k, w_k)\big)w_k \\ &\quad - \big(f_2(\lambda_j, w_j) - f_2(\lambda_k, w_k)\big).\end{aligned}$$

Since $\{w_j\} \subset X$ is bounded, by the interpolation inequality (4.17) it follows that the coefficients of the quasi-linear operator $\tilde{A}(\lambda_j, w_j)$ are equicontinuous in j, and thus

$$\|(\tilde{A}(\lambda_j, w_j)[w_k] - \tilde{A}(\lambda_k, w_k))[w_k]\|_{Y_1} \to 0 \qquad \text{as } j, k \to \infty.$$

The convergence

$$\|f_1(\lambda_j, w_j) - f_1(\lambda_k, w_k)\|_{Y_1} \to 0 \qquad \text{as } j, k \to \infty$$

uses the fact that f is a quartic polynomial expression for w and w_p, w_q. These together with the convergence of $\{y_{1j}\}$ in Y_1 yield that

$$\|\tilde{A}(\lambda_j, w_j)[w_j - w_k]\|_{Y_1} \to 0 \qquad \text{as } j, k \to \infty.$$

On the other hand, the standard Schauder estimates and the embedding properties for Hölder spaces in a bounded domain confirm

$$\|\tilde{B}(\lambda_j, w_j)[w_j - w_k]\|_{Y_2} \to 0 \qquad \text{as } j, k \to \infty.$$

Therefore, $w_j \to w$ in $C^{3+\alpha}(\overline{R})$. The assertion then follows since X is a closed subspace of $C^{3+\alpha}(\overline{R})$. ∎

We are now in a position to define a degree for $F^\epsilon(\lambda, w)$. Summarized in Appendix B is the general development in [18, 23], which takes into account the nonlinear boundary operator.

4.3. Global theory of bifurcation for approximate problems. Now the presentation is the existence in the large of nontrivial solutions to (3.11) via global results of abstract bifurcation theory, which uses the generalized degree theory developed in the previous subsection and Appendix B.

For each $\delta > 0$ and $0 < \epsilon < \epsilon_0$, let $\mathcal{S}_\delta^\epsilon$ be the closure in $\mathbb{R} \times X$ of the set of all nontrivial solution pairs $(\lambda, w) \in \mathcal{O}_\delta$ to (3.11), where $\mathcal{O}_\delta$ is defined in (4.6), and let

$\mathcal{C}_\delta^\epsilon \subset \mathbb{R} \times X$ be the connected component of $\mathcal{S}_\delta^\epsilon$ which contains the bifurcation point $(\lambda^\epsilon, 0)$ determined in Lemma 4.1. The local curve of solutions $\mathcal{C}_{loc}^\epsilon$ constructed in Proposition 4.2 is contained in $\mathcal{C}_\delta^\epsilon$ for the parameter $s_0 > 0$ of the curve sufficiently small.

The following global bifurcation result is immediate.

THEOREM 4.7. *For any $\delta > 0$ and $0 < \epsilon < \epsilon_0$, at least one of the following holds:*

(i) *$\mathcal{C}_\delta^\epsilon$ is unbounded in $\mathbb{R} \times X$;*

(ii) *$\mathcal{C}_\delta^\epsilon$ contains another trivial solution pair $(\lambda, 0)$ with $\lambda \neq \lambda^\epsilon$;*

(iii) *$\mathcal{C}_\delta^\epsilon$ meets $\partial\mathcal{O}_\delta$.*

Proof. The proof is almost identical to that of [33, Theorem 1.3] except that we now use the generalized degree defined in [18, 23] (see Appendix B) in place of the Leray–Schauder degree. We assume on the contrary that $\mathcal{C}_\delta^\epsilon$ is not characterized by any of (i), (ii), or (iii), and we argue by contradiction. ∎

The remainder of this subsection is devoted to extending and refining the result of Theorem 4.7.

First, the exploitation of symmetry rules out the second alternative from Theorem 4.7. In order to state our result more precisely, we define the nodal sets

$$R^+ = (0, \pi) \times (-\infty, 0), \qquad T^+ = (0, \pi) \times \{0\},$$
$$\partial R_l^+ = \{(0, p) : p \in (-\infty, 0)\}, \qquad \partial R_r^+ = \{(\pi, p) : p \in (-\infty, 0)\}.$$

It follows as an application of the maximum principle and its sharp form at the corner points [35] that any nontrivial solution $w \in \mathcal{C}_\delta^\epsilon$ to (3.11) possesses the following nodal pattern:

$$w_q < 0 \quad \text{in } R^+ \cup T^+, \tag{4.19}$$
$$w_{qq} < 0 \quad \text{on } \partial R_l^+, \qquad w_{qq} > 0 \quad \text{on } \partial R_r^+, \tag{4.20}$$
$$\text{either} \quad w_{qq}(0,0) < 0 \quad \text{or} \quad w_{qqp}(0,0) > 0, \tag{4.21}$$
$$\text{either} \quad w_{qq}(\pi,0) > 0 \quad \text{or} \quad w_{qqp}(\pi,0) < 0. \tag{4.22}$$

In other words, inequalities (4.19)–(4.22) hold along the entire continuum $\mathcal{C}_\delta^\epsilon \setminus (\lambda^\epsilon, 0)$. The proof is given in Appendix C. Let us define the open set

$$\mathcal{N} = \{\, w \in X : w \text{ satisfies (4.19)–(4.22)}\}. \tag{4.23}$$

By evenness and periodicity of $w \in X$, we infer that $w_q = 0$ on $\partial R_l^+ \cup \partial R_r^+$.

LEMMA 4.8. *If a trivial solution $(\lambda, 0)$ belongs to the continuum $\mathcal{C}_\delta^\epsilon$, then $\lambda = \lambda^\epsilon$.*

Proof. Suppose there is a sequence of nontrivial solution pairs $\{(\lambda_j, w_j)\} \subset \mathcal{C}_\delta^\epsilon$ to (3.11) which converges to $(\lambda, 0)$ in $\mathbb{R} \times X$. We consider a sequence of functions $v_j = \partial_q w_j / \|\partial_q w_j\|_{C^{2+\alpha}(\overline{R})}$, each of which solves

$$(A^\epsilon(\lambda_j, w_j), B(\lambda_j, w_j))[v_j] = 0.$$

The difference $v_j - v_k$ with $j, k \geq 1$ in turn solves the following linear boundary value problem:

$$A^\epsilon(\lambda_j, w_j)[v_j - v_k] = \big(A^\epsilon(\lambda_k, w_k) - A^\epsilon(\lambda_j, w_j)\big)[v_k] \qquad \text{in } R,$$
$$B(\lambda_j, w_j)[v_j - v_k] = \big(B(\lambda_k, w_k) \ - B(\lambda_j, w_j)\big)[v_k] \qquad \text{on } T.$$

The Schauder estimates [2] yield

$$\|v_j - v_k\|_{C^{2+\alpha}(\overline{R})} \le C\big(\|A^\epsilon(\lambda_j, w_j)[v_j - v_k]\|_{C^\alpha(\overline{R})} + \|B(\lambda_j, w_j)[v_j - v_k]\|_{C^{1+\alpha}(T)} + \|v_j - v_k\|_{C^0(\overline{R})}\big), \tag{4.24}$$

where $C > 0$ is independent of index j, k. As is done in Lemma 4.6, we show that

$$\begin{aligned}\|\big(A^\epsilon(\lambda_k, w_k) - A^\epsilon(\lambda_j, w_j)\big)[v_k]\|_{C^\alpha(\overline{R})} &\to 0,\\ \|\big(B(\lambda_k, w_k) - B(\lambda_j, w_j)\big)[v_k]\|_{C^{1+\alpha}(T)} &\to 0,\end{aligned}$$

as $j, k \to \infty$. This utilizes the facts that $\{(\lambda_j, w_j)\}$ converges in $\mathbb{R} \times X$ and that $\{v_j\}$ is bounded in $C^{2+\alpha}(\overline{R})$. The last term on the right side of (4.24) approaches zero as $j, k \to \infty$ since the limiting problem of $(A^\epsilon(\lambda_j, w_j), B(\lambda_j, w_j))$ does not have nontrivial solutions, completely analogous to the argument in the proof of Lemma 4.3 or Lemma 4.6. Therefore, $\{v_j\}$ converges to v, say, in $C^{2+\alpha}(\overline{R})$.

Note that $\partial^\beta v \in o(1)$ as $p \to -\infty$ uniformly for q for all multi-indices $|\beta| \le 2$ and $\|v\|_{C^{2+\alpha}(\overline{R})} = 1$. Since each v_j is 2π periodic in the q-variable and of mean zero over one period, $v = \partial_q \phi$ for some function ϕ which is 2π periodic in the q-variable. By continuity,

$$F^\epsilon_w(\lambda, 0)[\partial_q \phi] = 0, \tag{4.25}$$

$\partial_q \phi \le 0$ on R^+, and $\partial_q \phi = 0$ on $\partial R^+_l \cup \partial R^+_r$. Furthermore, since $\partial_q \phi$ satisfies (4.25) and $\partial_q \phi \not\equiv 0$ in R^+, the maximum principle ensures that $\partial_q \phi < 0$ in R^+.

We now write $\partial_q \phi$ as a sine series

$$\partial_q \phi(q, p) = \sum_{k=0}^\infty \phi_k(p) \sin kq,$$

whence (4.25) is written as

$$\begin{aligned}\sum_{k=0}^\infty \big((a^3(\lambda)\phi_k')' - \epsilon\phi_k - k^2 a(\lambda)\phi_k\big) \sin kq = 0 \qquad &\text{in } R^+,\\ \sum_{k=0}^\infty \big(-2\sqrt{\lambda}\phi_k'(0) + 2g/\lambda \phi_k(0)\big) \sin kq = 0.&\end{aligned}$$

Here $\partial_q \phi$ is subject to the vanishing condition $\partial_q \phi_k \to 0$ as $p \to -\infty$. In particular, ϕ_1 solves the boundary value problem

$$\begin{aligned}(a^3(\lambda)\phi_1')' - \epsilon\phi_1 = a(\lambda)\phi_1 \quad &\text{for } p \in (-\infty, 0),\\ \lambda^{3/2}\phi_1'(0) = g\phi_1(0),&\\ \phi_1(p) \to 0 \qquad\qquad &\text{as } p \to -\infty.\end{aligned}$$

In view of the definitions (4.3) and (4.4) it follows that $\Lambda^\epsilon(\lambda) \le R^\epsilon(\phi_1; \lambda) = -1$. Suppose $\Lambda^\epsilon(\lambda) < -1$; the minimizer Φ would be an eigenfunction corresponding to the simple eigenvalue $\Lambda^\epsilon(\lambda)$ (such that $R^\epsilon(\Phi; \lambda) = \Lambda^\epsilon(\lambda) = \inf R^\epsilon(\lambda)$), and hence Φ does not vanish on $p \in (-\infty, 0)$. On the other hand,

$$\phi_1(p) = \frac{2}{\pi} \int_0^\pi \phi_q(q, p) \sin q \, dq < 0 \qquad \text{for all } p \in (-\infty, 0). \tag{4.26}$$

This contradicts the orthogonality

$$\int_{-\infty}^{0} a(\lambda)\Phi^{\epsilon}(p)\phi_1(p)dp = 0.$$

Therefore, $\Lambda^{\epsilon}(\lambda) = -1$, and $\lambda = \lambda^{\epsilon}$ follows from the monotonicity of Λ (see [10, Lemma 3.4]). $\square$

We summarize our results from Theorem 4.7 and Lemma 4.8.

THEOREM 4.9. *For $\delta > 0$ and $0 < \epsilon < \epsilon_0$, the global continuum $\mathcal{C}_\delta^\epsilon$ either is unbounded in $\mathbb{R} \times X$ or intersects $\partial\mathcal{O}_\delta$. Each nontrivial solution lying on $\mathcal{C}_\delta^\epsilon$ has the nodal configuration* (4.19)–(4.22)*, i.e., $\mathcal{C}_\delta^\epsilon \setminus (\lambda^\epsilon, 0) \subset \mathbb{R} \times \mathcal{N}$.*

The following theorem describes global bifurcation in terms of open sets in $\mathcal{O}_\delta$, which was first developed in [33] and stated explicitly in [5].

THEOREM 4.10 (see [5, Theorem A6]). *Suppose $S \subset \mathcal{O}_\delta$ is a closed set with $(\lambda, 0) \in S$ such that every bounded subset of S is relatively compact in $\mathbb{R} \times X$. Let C be the maximal connected subset of S which contains $(\lambda, 0)$. Then C either is unbounded in $\mathbb{R} \times X$ or meets $\partial\mathcal{O}_\delta$ if and only if $\partial U \cap S \neq \emptyset$ for every bounded open set U with $(\lambda, 0) \in U$ and $\overline{U} \subset \mathcal{O}_\delta$.*

Proof. The proof is almost identical to that of [5, Theorem A6]. We choose an open set U as described above. If C is unbounded, then $(\lambda, 0) \in U \cap C \neq \emptyset$ and $(\overline{\mathcal{O}}_\delta \setminus \overline{U}) \cap C \neq \emptyset$. Since C is connected, $\partial U \cap C \neq \emptyset$, and thus $\partial U \cap S \neq \emptyset$. Similarly, if C intersects $\partial\mathcal{O}_\delta$, then $(\lambda, 0) \in U \cap C \neq \emptyset$ and $(\overline{\mathcal{O}}_\delta \setminus \overline{U}) \cap C \neq \emptyset$. Therefore, $\partial U \cap S \neq \emptyset$.

Conversely, we assume that C is bounded and does not intersect $\partial\mathcal{O}_\delta$, yet $\partial U \cap S \neq \emptyset$ for every bounded open set U with $(\lambda, 0) \in U$ and $\overline{U} \subset \mathcal{O}_\delta$. We choose $R > 0$ large enough that $C \subset B(R)$, where $B(R) = \{(\lambda, w) : |\lambda| + \|w\|_X < R\}$. Let us denote by M a compact metric space $\overline{B(R)} \cap S$; $E = C$ and $F = \{(\lambda, w) \in S : |\lambda| + \|w\|_X = 2R\}$ are nonempty disjoint compact subsets of M. Then Whyburn's lemma [33, Lemma 1.2] applies, and there exist disjoint compact subsets M_1 and M_2 such that $M_1 \cup M_2 = M$, $E \subset M_1$, and $F \subset M_2$. Let

$$U = \{(\lambda, w) : |\lambda - \lambda'| + \|w - w'\|_X < \delta/2 \text{ for } (\lambda', w') \in M_1\},$$

where $\delta = \min(\operatorname{dist}(M_1, M_2), \operatorname{dist}(M_1, \partial B(2R)))$. By assumption, we can find a $(\lambda, w) \in \partial U \cap S$. However, this contradicts that $\operatorname{dist}((\lambda, w), M_1) \leq \delta/2$ and $\operatorname{dist}(M_1, M_2) \geq \delta$. This completes the proof. $\square$

COROLLARY 4.11. *Let U be a bounded open set in $\mathcal{O}_\delta$ such that $(\lambda^0, 0) \in U$ and $\overline{U} \subset \mathcal{O}_\delta$. Then for ϵ sufficiently small,*

$$\partial U \cap \mathcal{C}_\delta^\epsilon \neq \emptyset.$$

Proof. This result is an immediate application of Theorems 4.9 and 4.10. The result of Lemma 4.6 says that any bounded subset of $\mathcal{S}_\delta^\epsilon$ is relatively compact in $\mathbb{R} \times X$. By Lemma 4.1(b), $(\lambda^\epsilon, 0) \in U$ for sufficiently small ϵ. $\square$

5. Existence theory for deep-water waves.

5.1. Global theory of existence for the singular problem. In this subsection, we demonstrate the existence of a global connected set in $\mathbb{R} \times X$ of nontrivial solutions to the original singular problem (3.10).

Let $\mathcal{S}_\delta$ be the closure in $\mathbb{R} \times X$ of the set

$$\{ (\lambda, w) \in \mathcal{O}_\delta : F(\lambda, w) = 0, \ \ w \not\equiv 0, \ \ w \in \mathcal{N} \} \cup \{(\lambda^0, 0)\},$$

the set of nontrivial solutions to (3.10) with the nodal properties (4.19)–(4.22).

LEMMA 5.1. *A bounded subset of $\mathcal{S}_\delta$ is relatively compact in $\mathbb{R} \times X$.*

Proof. Let $\{(\lambda_j, w_j)\} \subset \mathcal{S}_\delta$ be a bounded sequence in $\mathbb{R} \times X$. Then $\{\lambda_j\}$ has a convergent subsequence in $\mathbb{R}$; we say $\lambda_j \to \lambda$ in $\mathbb{R}$. Our goal is to show that $\{w_j\}$ has a convergent subsequence in X.

As is indicated in the proof of Lemma 4.6, an important step is to obtain the convergence in the $C^0_{per}(\overline{R})$-norm. To accomplish this, it is convenient to express w_j as

$$w_j(q,p) = \int_0^q \partial_q w_j(s,p)ds + w_j(0,p). \tag{5.1}$$

Note that $\{\partial_q w_j(q,p)\}$ and $\{w_j(0,p)\}$ are bounded sequences in $C^{2+\alpha}_{per}(\overline{R})$ and $C^{3+\alpha}_{per}((-\infty, 0])$, respectively.

Our task is to prove that $\partial_q w_j$ decays exponentially as $p \to -\infty$ uniformly for q and the index j. The transformations in section 2.2 and section 3.1 assign to each (λ_j, w_j) a solution pair of $\eta_j(x)$ and $\psi_j(x,y)$ of the vorticity-stream formulation (2.9). With this prescription, $w_j \in \mathcal{N}$ means that $\partial_x \psi_j$ is odd and 2π periodic in the x-variable, and that $\partial_x \psi_j(x,y) > 0$ for $x \in (-\pi, 0)$. Lemma 2.1 applies, and $\partial_x \psi_j(x,y)$ acquires the exponential decay property (2.10). A straightforward change of variables back into (q,p)-variables via definitions (2.7) and (3.6) and the change of variables (3.1) yields the corresponding pointwise exponential decay of $\partial_q w_j$:

$$|\partial_q w_j(q,p)| \le \delta A \exp(p/\sqrt{\lambda_j}) \qquad \text{in } \{(q,p) \in \overline{R},\ \ p < p_0\}, \tag{5.2}$$

where $A > 0$ and $p_0 < 0$ are independent of j. Indeed,

$$\begin{aligned} y = H(p) + w(q,p) &= \int_0^p \frac{ds}{\sqrt{\lambda - 2\Gamma(s)}} - \frac{\lambda}{2g} + w(q,p) \\ &\le \frac{p}{\sqrt{\lambda}} + |w(q,p)|. \end{aligned} \tag{5.3}$$

We remark that the choice of A indeed depends only on $\sup_j |w_j(\pm\pi, 0)|$, which is bounded a priori once a bounded sequence $\{w_j\}$ is chosen.

We proceed similarly on the estimate (2.11) to obtain an analogous exponential decay property of $w_j(0,p)$:

$$|\partial_q^2 w_j(0,p)| \le B \exp(p/\sqrt{\lambda_j}) \qquad \text{on } p \in (-\infty, p_0), \tag{5.4}$$

where $B > 0$ is independent of the index j. When restricted on $q = 0$, oddness of w_j reduces (3.7a) to

$$\begin{aligned} &\partial_p^2 w_j + (a^{-1}(\lambda) + \partial_p w_j)\partial_q^2 w_j \\ &\qquad + \gamma(-p)(a^{-1}(\lambda) + \partial_p w_j)^3 - \gamma(-p)a^{-3}(\lambda) = 0. \end{aligned} \tag{5.5}$$

Since $\gamma \in O(s^{-2-2l})$ as $s \to \infty$ for $l > 0$, it follows from the exponential decay (5.4) and from (5.5) that $w_j(0,p)$ decays as $p \to -\infty$ uniformly for j.

Having now established the uniform decay of $\partial_q w_j(q,p)$ and $w_j(0,p)$, we employ an argument of Ascoli type to conclude that both $\{\partial_q w_j(q,p)\}$ and $\{w_j(0,p)\}$ have subsequences which converge uniformly in the C^0-norm. These uniform convergences together with the expression (5.1) ensure that $\{w_j\}$ has a convergent subsequence in the C^0-norm; we say $w_j \to w$ in $C^0_{per}(\overline{R})$ for some $w \in X$.

The remainder of the proof is nearly identical to that of Lemma 4.6, and we only outline the various stages of the proof.

Since $\{w_j\}$ is bounded in X, an interpolation inequality (see (4.17) in Lemma 4.6 and [25, Theorem 3.2.1]) asserts that $w_j \to w$ in $C^3_{per}(\overline{R})$.

As is done in (4.16) in the proof of Lemma 4.6, we decompose F as

$$F_1(\lambda, w) = \tilde{A}(\lambda, w)[w] + f_1(\lambda, w),$$
$$F_2(\lambda, w) = \tilde{B}(\lambda, w)[w] + f_2(\lambda, w).$$

Here $\tilde{A}(\lambda, w)$, $\tilde{B}(\lambda, w)$, and f_2 are same as in the proof of Lemma 4.6; f_1 is a cubic polynomial expression of w_p and w_q. An estimate of Schauder type [2]

$$\|w_j - w_k\|_X \le C(\|\tilde{A}(\lambda_j, w_j)[w_j - w_k]\|_{Y_1} + \|\tilde{B}(\lambda_j, w_j)[w_j - w_k]\|_{Y_2} + \|w_j - w_k\|_Z)$$

holds, where $C > 0$ is independent of w_j. From the interpolation inequality (4.17) and the above decomposition follows

$$\|\tilde{A}(\lambda_j, w_j)[w_j - w_k]\|_{Y_1} \to 0,$$
$$\|\tilde{B}(\lambda_j, w_j)[w_j - w_k]\|_{Y_2} \to 0$$

as $j, k \to \infty$. The Schauder estimate then asserts $w_j \to w$ in X. This completes the proof. □

The global result of the existence of the deep-water waves is now immediate and is described in the next theorem.

THEOREM 5.2. *Let $\mathcal{C}'_\delta$ denote the maximal connected subset of $\mathcal{S}_\delta$ which contains* $(\lambda^0, 0)$.

(i) *The continuum $\mathcal{C}'_\delta$ either is unbounded in $\mathbb{R} \times X$ or intersects $\partial\mathcal{O}_\delta$.*

(ii) *Each nontrivial solution lying on $\mathcal{C}'_\delta$ has precisely the nodal properties* (4.19)–(4.22).

Proof. (i) By virtue of Theorem 4.10, it suffices to show that if U is a bounded open set with $(\lambda^0, 0) \in U$ and $\overline{U} \subset \mathcal{O}_\delta$, then $\partial U \cap \mathcal{S}_\delta \neq \emptyset$. Let U be such an open set. The result of Corollary 4.11 is that there are sequences $\epsilon_j \to 0$ as $j \to \infty$ and $\{(\lambda_j, w_j)\} \subset \mathcal{O}_\delta$ such that

$$(\lambda_j, w_j) \in \partial U \cap \mathcal{C}^{\epsilon_j}_\delta.$$

In other words, $\{(\lambda_j, w_j)\} \subset \mathcal{O}_\delta$ is a bounded sequence in $\mathbb{R} \times X$ and

$$F(\lambda_j, w_j) = \epsilon_j (a^{-1}(\lambda_j) + \partial_p w_j)^3 w_j. \tag{5.6}$$

It will follow by the methods in the proof of Lemma 5.1 that $\{(\lambda_j, w_j)\}$ has a subsequence which converges in $\mathbb{R} \times X$ to an element in $\partial U \cap \mathcal{S}_\delta$. For the rest of the proof it is assumed that $\lambda_j \to \lambda$ in $\mathbb{R}$ as $j \to \infty$.

As in done in the previous lemma, to each solution (λ_j, w_j) of (5.6) we associate a pair of functions $\eta_j(x)$ and $\psi_j(x, y)$ via the definitions (2.9) and (3.6) and the change of variables (3.1). It immediately follows that $\eta_j(x)$ and $\psi_j(x, y)$ are even and 2π periodic in the x-variable. From the nodal property $\partial_q w_j(q, p) < 0$ for $q \in (-\pi, 0)$, we infer that $\partial_x \psi_j(x, y) < 0$ for $(x, y) \in$ and $x \in (-\pi, 0)$.

It is straightforward to see that $\partial_x \psi_j$ satisfies the following Poisson equation in $\overline{\mathcal{D}}_{\eta_j}$:

$$-\Delta \partial_x \psi_j + \epsilon H'(-\psi_j) \partial_x \psi_j = \gamma'(\psi_j) \partial_x \psi_j.$$

Note that $H'(-\psi_j) > 0$ and $\gamma'(\psi_j) \leq 0$. One exercises the argument of the maximum principle type employed in the proof of Lemma 2.1 to conclude that $\partial_x \psi_j(x,y)$ decays exponentially uniformly for x:

$$|\partial_x \psi_j(x,y)| < Ae^y \qquad \text{on } (-\pi,\pi) \times (-\infty, \eta_j(\pm\pi)),$$

where $A > 0$ depends only on $\sup_j |\eta_j(\pm\pi)|$. In particular, A is independent of ϵ. An exponential decay property of $\partial_x^2 \psi_j(0,y)$, analogous to (2.11), is valid on $(-\infty, \eta_j(\pm\pi))$.

Henceforth, the exponential decay estimates for $\partial_q w_j(q,p)$ and $\partial_q^2 w_j(0,p)$ follow with the prescribed change of variables back into (q,p)-variables:

$$\begin{aligned}
|\partial_q w_j(q,p)| \leq& A_1 \exp(p/\sqrt{\lambda_j}) \qquad \text{in } \{(q,p) \in \overline{R},\ \ p < p_0\},\\
|\partial_q^2 w_j(0,p)| \leq& B_1 \exp(p/\sqrt{\lambda_j}) \qquad \text{on } p \in (-\infty, p_0),
\end{aligned}$$

where p_0 depends only on $\sup_j |w_j(\pm\pi, 0)|$. In particular, p_0 is independent of ϵ and the index j. As is done in the previous lemma, one can show that $w_j(0,p)$ decays as $p \to -\infty$ uniformly for the index j. Since $\{\partial_q w_j(q,p)\}$ and $\{w_j(0,p)\}$ are bounded in the Hölder norms, an argument of Ascoli type yields that both $\{\partial_q w_j(q,p)\}$ and $\{w_j(0,p)\}$ have subsequences which converge in the C^0-norm. We will denote the convergent subsequence by $\{w_j\}$. From the expression (5.1) follows $w_j \to w$ in $C^0_{per}(\overline{R})$ for some $w \in X$.

The remainder of the proof is nearly identical to those of Lemmas 4.6 and 5.1. Since $\{w_j\}$ is bounded in X, an interpolation inequality (4.17) (see [25, Theorem 3.2.1], for instance) asserts that $w_j \to w$ in $C^3_{per}(\overline{R})$.

We employ Schauder theory [2] to obtain the inequality

$$\begin{aligned}
\|w_j - w_k\|_X \leq C(&\|\tilde{A}(\lambda_j, w_j)[w_j - w_k]\|_{Y_1}\\
&+ \|\tilde{B}(\lambda_j, w_j)[w_j - w_k]\|_{Y_2} + \|w_j - w_k\|_Z),
\end{aligned}$$

where $C > 0$ is independent of the index j and k. With the decomposition (4.16),

$$\begin{aligned}
\tilde{A}(\lambda_j, w_j)[w_j - w_k] =& \epsilon_j a^{-3}(\lambda_j)(w_j - w_k) + w_k(\epsilon_j a^{-3}(\lambda_j) - \epsilon_k a^{-3}(\lambda_k))\\
&- \big(\tilde{A}(\lambda_j, w_j) - \tilde{A}(\lambda_k, w_k)\big)w_k - \big(f_1(\lambda_j, w_j) - f_1(\lambda_k, w_k)\big),\\
\tilde{B}(\lambda_j, w_j)[w_j - w_k] =& -\big(\tilde{B}(\lambda_j, w_j) - \tilde{B}(\lambda_k, w_k)\big)w_k\\
&- \big(f_2(\lambda_j, w_j) - f_2(\lambda_k, w_k)\big).
\end{aligned}$$

An interpolation inequality (4.17) and the above observation imply

$$\|\tilde{A}(\lambda_j, w_j)[w_j - w_k]\|_{Y_1} \to 0 \qquad \text{as } j,k \to \infty,$$

completely analogous to that in the proof of Lemma 4.6. Evidently,

$$\|\epsilon_j a^{-3}(\lambda_j)(w_j - w_k) + w_k(\epsilon_j a^{-3}(\lambda_j) - \epsilon_k a^{-3}(\lambda_k))\|_{Y_1} \to 0 \qquad \text{as } j,k \to \infty.$$

The above Schauder estimate together with the elliptic estimate on the bounded domain T yields

$$\|\tilde{B}(\lambda_j, w_j)[w_j - w_k]\|_{Y_2} \to 0 \qquad \text{as } j,k \to \infty,$$

whence it follows that $w_j \to w$ in X.

(ii) The proof is nearly identical to that of Lemma C.3 in Appendix C. Suppose the contrary. Since $\mathcal{C}'_\delta \subset \mathcal{S}_\delta$ is connected, there must be a nontrivial solution $(\lambda, w) \in \mathcal{C}'_\delta$ with $w_q \not\equiv 0$ such that at least one of the nodal properties (4.19)–(4.22) would fail with (λ, w). We argue by contradiction using the maximum principle, the Hopf boundary lemma, and its sharp form at corner points due to Serrin [35] (see Lemma C.2). □

Remark 5.3. In case $\mathcal{C}'_\delta$ is unbounded in $\mathbb{R} \times X$, there is a sequence of solution pairs $\{(\lambda_j, w_j)\} \subset \mathcal{C}'_\delta$ such that either

(1) $\lim_{j\to\infty} \lambda_j = \infty$ or

(2) $\lim_{j\to\infty} \|w_j\|_X = \infty$.

If the other alternative that $\mathcal{C}'_\delta$ intersects $\partial\mathcal{O}_\delta$ holds, there is a solution pair $(\lambda, w) \in \mathcal{C}'_\delta$ such that one of the following holds:

(3) $\lambda = 2\Gamma_\infty + \delta$,

(4) $a^{-1}(\lambda) + w_p = \delta$ somewhere in $\overline{R}$,

(5) $w = \frac{2\lambda-\delta}{4g}$ somewhere on T.

5.2. Uniform regularity. The purpose of this subsection is to obtain bounds for the higher derivatives of w in terms of w and w_p uniformly along the continuum $\mathcal{C}'_\delta$.

THEOREM 5.4. *For each* $\delta > 0$, $\sup_{w\in\mathcal{C}'_\delta} \sup_{\overline{R}}(|w| + |w_p|) < \infty$ *implies*

$$\sup_{w\in\mathcal{C}'_\delta} \sup_{\overline{R}} \|w\|_X < \infty.$$

Proof. First, to demonstrate a uniform bound for w_q we consider the following linear elliptic boundary value problem:

$$\begin{aligned} A(\lambda, w)[w_q] = \tilde{A}(\lambda, w)[w_q] \\ &+ \big(-2w_q w_{pq} + 3\gamma(-p)(a^{-1}(\lambda) + w_p)^2\big) w_{qp} \\ &+ \big(2w_q w_{pp} - 2\gamma(-p)a^{-3}(\lambda) w_q\big) w_{qq} = 0, \end{aligned} \tag{5.7}$$

$$B(\lambda, w)[w_q] = \tilde{B}(\lambda, w)[w_q] + 2g(\lambda^{-1/2} + w_p)^2 w_q\big|_T = 0. \tag{5.8}$$

Here the principal parts $\tilde{A}(\lambda, w)$ and $\tilde{B}(\lambda, w)$ are defined in (4.9). We assume the solutions take the form $w(q,p) = k(q,p) - sqe^p$, where $s > 0$ is a constant. Substituting this into (5.7) results in the equation

$$\begin{aligned} &\tilde{A}(\lambda, k - sqe^p)[k_q] \\ &\quad - 2k_q k_{pq}^2 + 3\gamma(-p)(a^{-1}(\lambda) + k_p)^2 k_{qp} + 2k_q k_{pp} k_{qq} - 2\gamma(-p)a^{-3}(\lambda) k_q k_{qq} \\ &\quad - se^p\big((1 + k_q)^2 - 4k_q k_{pq} - 2k_{pq}^2 + 2k_{pp}k_{qq} - 2qk_q k_{qq} - 2\gamma(-p)k_{qq} \\ &\qquad + 3\gamma(-p)(a^{-1}(\lambda) + k_p)^2 + 6q\gamma(-p)(a^{-1}(\lambda) + k_p)k_{qp}\big) + O(s^2) = 0. \end{aligned} \tag{5.9}$$

Suppose k_q had an interior maximum in $\overline{R}$. At such a point the following would hold:

$$\begin{cases} k_{qp} = k_{qq} = 0, \\ k_{qpp}, k_{qqq} \le 0 \quad \text{and} \quad k_{qpp}k_{qqq} \ge k_{qpq}^2, \end{cases}$$

whence (5.9) would become

$$\tilde{A}(\lambda, k - sqe^p)[k_p] - se^p\big((1 + k_q^2) + 3\gamma(-p)(a^{-1}(\lambda) + k_p)^2\big) + O(s^2) = 0.$$

Since $(1+k_q^2)+3\gamma(-p)(a^{-1}(\lambda)+k_p)^2 > 1$, (5.9) is further reduced to the elliptic inequality $\tilde{A}(\lambda, k-sqe^p)[k_q] > 0$ for $s>0$ sufficiently small. Then, by the maximum principle, $k_q = w_q + se^p$ cannot have an interior maximum along $\mathcal{C}'_\delta$.

Note that $k_q \to 0$ as $p \to -\infty$. On the other hand, the nonlinear boundary condition (3.2b),

$$1 + (2gw-\lambda)(1/\sqrt{\lambda} + w_p^2) + (k_q - s)^2 = 0,$$

dictates that k_q on T is bounded by w and w_p.

Since w and w_p are uniformly bounded along $\mathcal{C}'_\delta$, the maximum of k_q and also the maximum of w_q are uniformly bounded along $\mathcal{C}'_\delta$. One repeats the same consideration with $w(q,p) = k(q,p) + sqe^p$ to conclude that the minimum of w_q along $\mathcal{C}'_\delta$ is bounded by w and w_p. Therefore, $\sup_{w\in\mathcal{C}'_\delta} \sup_{\overline{R}} |w_q|$ is finite.

Next, in order to obtain bounds for the second derivatives, we proceed with the a priori estimates of Schauder type due to Lieberman and Trudinger [27] for nonlinear elliptic partial differential equations with a nonlinear oblique boundary condition.

We consider a quasi-linear boundary value problem of the form

$$\begin{cases} F_1(\nabla w)[w] = \displaystyle\sum_{i,j=1}^{2} a^{ij}(\nabla w)\partial_{ij} w + f_1(\nabla w) = 0 & \text{in } R, \\ F_2(w, \nabla w) = 0 & \text{on } T. \end{cases} \tag{5.10}$$

Here $a^{ij}, f_1 \in C^2(\mathbb{R}^2)$, $F_2 \in C^2(\mathbb{R}\times\mathbb{R}^2)$; $\nabla w = (w_q, w_p)$ denotes the gradient of w.

Theorem 5.5. *Let $w \in C^2_{per}(\overline{R})$ be a solution to (5.10) such that $\partial^\beta w \in o(1)$ as $p \to -\infty$ for all multi-indices $|\beta| \le 2$. Suppose that $|w| + |\nabla w| \le K$ in $\overline{R}$ for some constant $K>0$ and that there exist $\delta, M > 0$ such that the following conditions hold for all $(z,r) \in \mathbb{R}\times\mathbb{R}^2$ with $|z|+|r| \le K$:*

$$\sum_{i,j=1}^{2} a^{ij}\xi_i\xi_j \ge \delta(\xi_1^2 + \xi_2^2) \qquad \textit{for all } (\xi_1,\xi_2) \in \mathbb{R}^2, \tag{5.11}$$

$$|a^{ij}|, |a^{ij}_r|, |f_r| \le \delta M, \tag{5.12}$$

$$\chi := F_{2r}\cdot(0,-1) > 0 \qquad \textit{on } T\times\mathbb{R}, \tag{5.13}$$

$$|F_2|, |F_{2z}|, |F_{2r}|, |F_{2zz}|, |F_{2zr}|, |F_{2rr}| \le \chi M. \tag{5.14}$$

Then for $\alpha' \in (0,1)$ and $C = C(K,M) > 0$ constants, $w \in C^{2+\alpha'}(\overline{R})$ and

$$\|w\|_{C^{2+\alpha'}(\overline{R})} \le C.$$

The assertion for any compact subset of $\overline{R}$ follows as a direct consequence of [27, Theorem 1.1]. Then we employ the treatment in [2, Theorem 6.3] and an interpolation theorem to obtain estimates valid on the entire domain $\overline{R}$.

In order to adapt Theorem 5.5 to the present setting, we take

$$\begin{aligned} \sum_{i,j=1}^{2} a^{ij}(r)\partial_{ij} &= (1+r_1^2)\partial_p^2 - 2(a^{-1}(\lambda)+r_2)r_1\partial_p\partial_q + (a^{-1}(\lambda)+r_2^2)\partial_q^2, \\ f_1(r) &= \gamma(-p)(a^{-1}+r_2)^3 - \gamma(-p)a^{-3}(\lambda)(1+r_1^2), \\ F_2(z,r) &= 1 + (2gz-\lambda)(\lambda^{-1/2}+r_2)^2 + r_1^2, \end{aligned}$$

where $z \in \mathbb{R}$ and $r = (r_1, r_2) \in \mathbb{R}^2$. Conditions (5.11) and (5.13) are fulfilled evidently; F_1 is uniformly elliptic and F_2 is uniformly oblique. In particular, there exists $s > 0$ such that

$$\chi = F_{2r} \cdot (0, -1) = 2(\lambda - 2gz)(\lambda^{-1/2} + r_2)\big|_T \geq s > 0,$$

since by (3.7b) we have

$$(\lambda - 2gw)(\lambda^{-1/2} + w_p) = \frac{1 + w_q^2}{\lambda^{-1/2} + |w_p|} \geq \frac{1}{(2\Gamma_\infty + \delta)^{-1/2} + \sup_{\overline{R}} |w_p|}$$

and the right side of the inequality is bounded along $\mathcal{C}'_\delta$. It is straightforward to verify the structure conditions (5.12) and (5.14). By virtue of Theorem 5.5 we then conclude that $\|w\|_{C^{2+\alpha'}(\overline{R})}$ for some $\alpha' \in (0,1)$ is uniformly bounded by $\sup_{\mathcal{C}'_\delta} \sup_{\overline{R}}(|w| + |w_p|)$ along the entire continuum $\mathcal{C}'_\delta$.

Our final task is to gain bounds for the higher derivatives. To do so, it is convenient to write (5.7) and (5.8) in the form

$$\begin{cases} \tilde{A}(\lambda, w)[w_q] = f_1(w_p, w_q, w_{pp}, w_{pq}, w_{qq}) & \text{in } R, \\ \tilde{B}(\lambda, w)[w_q] = -2g(\lambda^{-1/2} + w_p)^2 w_q & \text{on } T. \end{cases}$$

Note that $\tilde{A}(\lambda, w)$ is a uniformly elliptic differential operator whose coefficients are bounded in $C^{1+\alpha}(\overline{R})$ and that $\tilde{B}(\lambda, w)$ is a uniformly oblique boundary operator with bounded coefficients in $C^{1+\alpha}(T)$. Note as well that $f_1 \in C^{\alpha'}(\overline{R})$, where α' is as in the previous step. Accordingly, the (standard) Schauder estimate [2]

$$\begin{aligned} \|w_q\|_{C^{2+\alpha'}(\overline{R})} &\leq C\left(\|f_1\|_{C^{\alpha'}(\overline{R})} + \|2g(\lambda^{-1/2} + w_p)^2 w_q\|_{C^{1+\alpha'}(T)} + \|w_q\|_{C^0(\overline{R})}\right) \\ &\leq \tilde{C}(\|w\|_{C^{2+\alpha'}(\overline{R})}) \end{aligned}$$

holds, where $\tilde{C}$ depends on $\|w\|_{C^{2+\alpha'}(\overline{R})}$. That is to say, w_q earns a uniform $C^{2+\alpha'}(\overline{R})$ bound along $\mathcal{C}'_\delta$.

It remains to prove a uniform $C^{2+\alpha'}(\overline{R})$ estimate for w_p along $\mathcal{C}'_\delta$. We use the differential equation (3.7a) to express w_{pp} in terms of other derivatives of w of order less than or equal to 2, each of which is bounded along $\mathcal{C}'_\delta$ in the $C^{2+\alpha'}(\overline{R})$-norm. Thus, $\|w\|_{C^{3+\alpha'}(\overline{R})}$ remains uniformly bounded along $\mathcal{C}'_\delta$. In particular, $\|w\|_{C^{2+\alpha}(\overline{R})}$ is bounded along $\mathcal{C}'_\delta$. The assertion of Theorem 5.4 then follows by repeating the same argument as in the previous step with $\alpha' = \alpha$. □

5.3. Proof of the main result (Theorem 2.3). Throughout this subsection, $(u - c, v, \eta)$ represents the solution triple of the deep-water problem (2.1)–(2.6) corresponding to the solution pair (λ, w) of (3.10) via the transform in section 3.1.

The following observations are useful.

LEMMA 5.6 (speed at crest). *The relative flow speed at the crest of any nontrivial flow is bounded:*

$$(c - u(0,0))^2 = Q < \lambda. \tag{5.15}$$

Proof. Any nontrivial solution w belongs to $\mathcal{N}$, and thus it enjoys the nodal properties, $w_q = 0$ and $w_{qq} < 0$ on $q = 0$. Subsequently, $h_q = 0$ and $h_{qq} < 0$ on $q = 0$. Under the circumstances, (3.2a) reduces to the inequality $h_{pp} > -\gamma(-p)h_p^3$. Further,

$$-\left(\frac{1}{h_p^2(0,p)}\right)_p > -2\gamma(-p) \qquad \text{for } p \in (-\infty, 0).$$

Integration over $p \in (-\infty, 0)$ then yields $c^2 - (c - u(0,0))^2 > -2\Gamma_\infty$, from which (5.15) follows. $\square$

The same calculation carried out on $q = \pm\pi$ leads to an analogous bound for the relative flow speed at the trough:

$$(c - u(\pm\pi, \eta(\pm\pi)))^2 > \lambda. \tag{5.16}$$

The following properties of nontrivial solutions are direct consequences of (5.15) and (5.16):

$$w(0,0) = \ h(0,0) - \ H(0) = \frac{\lambda - Q}{2g} > 0, \tag{5.17}$$

$$w(\pm\pi, 0) = h(\pm\pi, 0) - \ H(0) = \frac{\lambda - (c - u(\pi, \eta(\pi)))^2}{2g} < 0. \tag{5.18}$$

Lemma 5.7 (pressure estimate). *The pressure satisfies the inequality*

$$P + gy - \Gamma(-\psi) \geq P_0 + g\eta(\pm\pi). \tag{5.19}$$

Proof. Let $s > 0$ and define a function on $\mathcal{D}_\eta$ by

$$M^s(x,y) = \tfrac{1}{2}(\psi_x^2 + \psi_y^2) + sy - E + P_0 = (s-g)y - P - \Gamma(-\psi) + P_0. \tag{5.20}$$

Note that

$$\Delta M^s + 2\gamma'(\psi) M^s = \psi_{xx}^2 + 2\psi_{xy}^2 + \psi_{yy}^2 + 2(sy - E + P_0)\gamma'(\psi) \geq 0.$$

The inequality utilizes the fact that $\gamma'(\psi) \leq 0$ and that $E - P_0 = \frac{1}{2}(c - u(0,0))^2 > 0$. The weak maximum principle asserts that M^s in $\overline{\mathcal{D}}_\eta$ attains its maximum on the free surface or at the infinite bottom. On the other hand,

$$\begin{aligned} M_y^s &= \psi_x\psi_{xy} + \psi_y\psi_{yy} + s \\ &= \psi_x\psi_{xy} - \psi_y\psi_{xx} - \psi_y\gamma(\psi) + s > 0 \end{aligned}$$

as $y \to -\infty$, since $\psi_x, \psi_{xx} \to 0$ and $\gamma(\psi) \to 0$ as $y \to -\infty$ uniformly for x while ψ_y is bounded. Therefore, M^s attains its maximum on the surface, and

$$(s-g)y - P - \Gamma(-\psi) + P_0 \leq (s-g)\eta(\pm\pi)$$

for each $s > 0$. Inequality (5.19) then follows as the limit $s \to 0$. $\square$

By the same token, the function

$$M(x,y) := \tfrac{1}{2}(\psi_x^2 + \psi_y^2) + gy - E + P_0 = -P - \Gamma(-\psi) + P_0$$

attains its maximum on the free surface. By the Hopf boundary point lemma,

$$e(x) := (-\eta_x, 1) \cdot (M_x, M_y)\big|_{\mathcal{S}_\eta} > 0 \qquad \text{for } x \in [0, \pi]. \tag{5.21}$$

On the other hand, since M is constantly zero on the surface,

$$(1, \eta_x) \cdot (M_x, M_y)\big|_{\mathcal{S}_\eta} = 0 \qquad \text{for } x \in [0, \pi], \tag{5.22}$$

and hence

$$M_x\big|_{\mathcal{S}_\eta} = -\frac{e(x)}{1+\eta_x^2}\eta_x \ge 0 \qquad \text{for } x \in [0,\pi].$$

Therefore, for $x \in [0,\pi]$

$$\begin{aligned}\frac{d}{dx}\left(\psi_y^2(x,\eta(x))\right) &= 2\psi_y\psi_{xy} + 2\psi_y\psi_{yy}\eta_x = 2\psi_y\psi_{xy} - 2\psi_x\psi_{yy} \\ &= 2\psi_y\psi_{xy} + 2\psi_x(\psi_{xx}+\gamma(0)) \\ &= -2\eta_x\left(\frac{e(x)}{1+\eta_x^2} - \gamma(0)(c-u)\big|_{\mathcal{S}_\eta}\right) \ge 0,\end{aligned}$$

provided that

$$\gamma(0) \le \min_{x\in[0,\pi]}\left(\frac{e(x)}{1+\eta_x^2(x)}\cdot\frac{1}{c-u(x,\eta(x))}\right). \tag{5.23}$$

Since $\psi_y(x,\eta(x))<0$, this says that $\psi_y(x,y)$ is nondecreasing for $x\in[0,\pi]$.

This is summarized in the following lemma.

LEMMA 5.8. *If $\gamma(0)$ satisfies* (5.23), *then the relative flow speed $c-u(x,\eta(x))$ is nondecreasing from crest to trough:*

$$c-u(x,\eta(x)) \ge c-u(0,0) > 0 \qquad \textit{for } x\in[0,\pi]. \tag{5.24}$$

Proof of Theorem 2.3. For any $\delta>0$, due to Theorem 5.2, Remark 5.3, and Theorem 5.4, at least one of the following holds:

(1) there exists a sequence $\{(\lambda_j,w_j)\}\subset\mathcal{C}'_\delta$ such that $\lim_{j\to\infty}\lambda_j=\infty$;
(2) there exists a sequence $\{(\lambda_j,w_j)\}\subset\mathcal{C}'_\delta$ such that $\sup_{\overline{R}}|w_j|\to\infty$;
(3) there exists a sequence $\{(\lambda_j,w_j)\}\subset\mathcal{C}'_\delta$ such that $\sup_{\overline{R}}|\partial_p w_j|\to\infty$;
(4) there exists a solution pair $(\lambda,w)\in\mathcal{C}'_\delta$ such that $\lambda=2\Gamma_\infty+\delta$;
(5) there exists a solution pair $(\lambda,w)\in\mathcal{C}'_\delta$ such that $a^{-1}(\lambda)+w_p=\delta$ at some point in $\overline{R}$;
(6) there exists a solution pair $(\lambda,w)\in\mathcal{C}'_\delta$ such that $w=\frac{2\lambda-\delta}{4g}$ at some point on T.

Our task is to give an interpretation of each alternative (1) through (6) in terms of the relative flow speed $c-u$ to prove assertion (ii) in Theorem 2.3. We note that

$$a^{-1}(\lambda)+w_p = h_p = \frac{1}{c-u}. \tag{5.25}$$

Alternative (1). From (5.16) it follows that $\lim_{j\to\infty}(c_j-u_j(\pm\pi,\eta_j(\pm\pi)))=\infty$.

Alternative (2). The nodal configuration says that $\partial_q w_j<0$ on R^+ whereas oddness of $\partial_q w_j$ requires that $w_q>0$ on R with $q\in(-\pi,0)$. That is, w is increasing where $q\in(-\pi,0)$ and decreasing where $q\in(0,\pi)$, and thus w attains its maximum on $q=0$ and its minimum on $q=\pm\pi$. Therefore, in the case of Alternative (2) either $\sup_p w_j(0,p)\uparrow\infty$ or $\inf_p w_j(\pm\pi,p)\downarrow-\infty$ must hold.

Suppose $\sup_p w_j(0,p)\uparrow\infty$. The nodal properties, $\partial_q w_j=0$ and $\partial_q^2 w_j<0$ on $q=0$, reduce (3.7a) to the inequality

$$\partial_p^2 w_j + \gamma(-p)\left((\partial_p w_j + \tfrac{3}{2}a^{-1}(\lambda_j))^2 + \tfrac{3}{4}a^{-2}(\lambda_j)\right)\partial_p w_j > 0 \qquad \text{on } q=0.$$

Since $w_j(0,0) > 0$ and $w_j(0,p)$ tends to zero as $p \to -\infty$, the maximum principle implies that $\sup_p w_j(0,p) = w_j(0,0)$. From (5.17) it follows $\lambda_j \uparrow \infty$, and in turn it follows $\lim_{j\to\infty}(c_j - u_j(\pm\pi, \eta_j(\pm\pi))) = \infty$, as is done previously for Alternative (1).

A similar consideration on $q = \pm\pi$ implies that $w(\pm\pi, p)$ on $p \in (-\infty, 0)$ attains a negative minimum at $p = 0$. Therefore, in case $\inf_p w_j(\pm\pi, p) \downarrow -\infty$, (5.18) implies that $\lim_{j\to\infty}(c_j - u_j(\pm\pi, \eta_j(\pm\pi))) = \infty$.

Alternative (3). Since $\delta < a^{-1}(\lambda) + w_p < \delta^{-1/2} + w_p$ for $(\lambda, w) \in \mathcal{O}_\delta$, i.e., w_p is bounded from below by $\delta - \delta^{-1/2}$, it must hold $\sup_{\overline{R}} \partial_p w_j \uparrow \infty$ under the circumstance. Then, by (5.25), $\sup_{\overline{R}} \partial_p h_j \uparrow \infty$ holds, and consequently $\inf_{\overline{\mathcal{D}}_{\eta_j}}(c_j - u_j) \downarrow 0$.

Alternative (4). To each $\delta_j \downarrow 0$ we associate $(\lambda_j, w_j) \in \mathcal{C}'_{\delta_j}$ such that $\lambda_j = 2\Gamma_\infty + \delta_j$. We may assume that $\sup_R \partial_p w_j$ is bounded; otherwise, $\inf_{\overline{\mathcal{D}}_{\eta_j}}(c_j - u_j) \downarrow 0$ must hold by the previous treatment for Alternative (3).

For a sequence $p_j \to -\infty$ with $2\Gamma_\infty - 2\Gamma(p_j) = \delta_j$, we have

$$\partial_p h_j(0,p) = a^{-1}(\lambda_j) + \partial_p w_j(0,p) > (2\delta_j)^{-1/2} - \sup_R \partial_p w_j(0,p) \uparrow \infty.$$

Therefore, by (5.25) $\inf_{\overline{\mathcal{D}}_{\eta_j}}(c_j - u_j) \downarrow 0$.

Alternative (5). Choose sequences $\delta_j \downarrow 0$ and $(\lambda_j, w_j) \in \mathcal{C}'_{\delta_j}$ such that $a^{-1}(\lambda_j) + \partial_p w_j = \delta_j$. By (5.25) it follows that $\sup_{\overline{\mathcal{D}}_{\eta_j}}(c_j - u_j) \uparrow \infty$.

Alternative (6). For a sequence $\delta_j \downarrow 0$ there exists a sequence $(\lambda_j, w_j) \in \mathcal{C}'_{\delta_j}$ such that $\lambda_j - 2gw_j = \frac{1}{2}\delta_j$ somewhere on T. The nonlinear boundary condition on top (3.2b) then yields

$$\frac{1}{\partial_p h_j^2} \le \frac{1 + \partial_q h_j^2}{\partial_p h_j^2} = \frac{\delta_j}{2} \downarrow 0,$$

whence $\inf_{\overline{\mathcal{D}}_{\eta_j}}(c_j - u_j) \downarrow 0$.

Let $\mathcal{C}' = \cup_{\delta>0}\mathcal{C}'_\delta$. By construction, $\mathcal{C}'_\delta \subset \mathcal{C}'_{\delta'}$ if $\delta > \delta'$. At this stage, the conclusion is that there is a sequence of solution triples $\{(u_j - c_j, v_j, \eta_j)\} \subset \mathcal{C}$ in the space $C^{2+\alpha}(\overline{\mathcal{D}}_\eta) \times C^{2+\alpha}(\overline{\mathcal{D}}_\eta) \times C^{3+\alpha}(\mathbb{R})$ such that

$$\text{either} \quad \sup_{\overline{\mathcal{D}}_{\eta_j}}(c_j - u_j) \uparrow \infty \quad \text{or} \quad \inf_{\overline{\mathcal{D}}_{\eta_j}}(c_j - u_j) \downarrow 0.$$

The remainder of the proof consists of studying the above two alternatives further.

First, let the first alternative $\sup_{\overline{\mathcal{D}}_{\eta_j}}(c_j - u_j) \uparrow \infty$ hold. We recall that

$$E = \tfrac{1}{2}\left((c-u)^2 + v^2\right) + gy + P + \Gamma(-\psi).$$

Evaluated at the trough $(\pm\pi, \eta(\pm\pi))$, it reduces to

$$E = \tfrac{1}{2}\left(c - u(\pm\pi, \eta(\pm\pi))\right)^2 + g\eta(\pm\pi) + P_0.$$

Subtracting the above two expressions for E, we arrive at

$$\begin{aligned}\tfrac{1}{2}(c-u)^2 &\le \tfrac{1}{2}\left(c - u(\pm\pi, \eta(\pm\pi))\right)^2 + g\eta(\pi) + P_0 - P - gy + \Gamma(-\psi) \\ &\le \tfrac{1}{2}\left(c - u(\pm\pi, \eta(\pm\pi))\right)^2.\end{aligned}$$

The last inequality uses Lemma 5.7. Therefore, if the first alternative $\sup_{\overline{\mathcal{D}}_{\eta_j}}(c_j - u_j) \uparrow \infty$ holds, $\lim_{j\to\infty}(c_j - u_j(\pm\pi, \eta_j(\pm\pi))) = \infty$ must hold.

Next, we assume the second alternative $\inf_{\overline{\mathcal{D}}_{\eta_j}}(c_j - u_j) \downarrow 0$ holds. We may choose a sequence $\{s_j\}$ and a sequence of solution triples $\{(u_j - c_j, v_j, \eta_j)\} \subset \mathcal{C}$ such that $s_j \downarrow 0$ and $c_j - u_j(x_j, y_j) = s_j$ at some point (x_j, y_j) in $\overline{\mathcal{D}}_{\eta_j}$; $\{y_j\}$ is assumed to be bounded from below as $u_j(x,y) \to 0$ as $y \to -\infty$ uniformly for x.

Differentiating the Poisson equation $-\Delta\psi = \gamma(\psi)$ leads to $\Delta(u-c)+\gamma'(\psi)(u-c) = 0$ in $\mathcal{D}_\eta$. Recall that $\psi_y = u - c$. Let us introduce a sequence of functions,

$$W_j(x,y) = u_j(x,y) - c_j + s_j e^{\beta(y-y_j)} \qquad \text{for } (x,y) \in \overline{\mathcal{D}}_{\eta_j}, \tag{5.26}$$

where $\beta > 0$ is a constant such that $\beta^2 + \gamma'(s) \geq 0$ for all $s \in [0,\infty)$. Note that

$$\Delta W_j + \gamma'(\psi) W_j = s_j(\beta^2 + \gamma'(\psi)) e^{\beta(y-y_j)} \geq 0 \qquad \text{in } \overline{\mathcal{D}}_{\eta_j}$$

and that $W_j(x_j, y_j) = 0$.

Since $\gamma'(\psi) \leq 0$, the weak maximum principle ensures that W_j in $\overline{\mathcal{D}}_{\eta_j}$ attains its maximum either on the surface or at the infinite bottom. On the other hand,

$$W_j \to -c_j < 0 \qquad \text{as } y \to -\infty.$$

Therefore, W_j in $\overline{\mathcal{D}}_{\eta_j}$ attains its maximum on the free surface $y = \eta_j(x)$. In particular, there is a sequence of points $\{(\xi_j, \eta_j(\xi_j)) : \xi_j \in [0,\pi]\}$ such that

$$u_j(\xi_j, \eta_j(\xi_j)) - c_j + s_j e^{\beta(\eta_j(\xi_j) - y_j)} \geq 0,$$

whence

$$0 \leq c_j - u_j(\xi_j, \eta_j(\xi_j)) \leq s_j e^{\beta(\eta_j(\xi_j) - y_j)}.$$

Since $\eta_j(\xi_j) \leq \eta_j(0) = 0$ and $\{y_j\}$ is bounded from below, we conclude that

$$\lim_{j\to\infty} (c_j - u_j(\xi_j, \eta_j(\xi_j))) = 0.$$

This proves (ii) of Theorem 2.3.

Furthermore, if $\gamma(0)$ satisfies (5.23), then, by Lemma 5.8, W_j in $\overline{\mathcal{D}}_{\eta_j}$ attains its maximum at the crest $(0,0)$, and therefore

$$\lim_{j\to\infty} (c_j - u_j(0,0)) = 0.$$

This completes the proof. $\square$

Remark 5.9. When the first alternative $\lim_{j\to\infty}(c_j - u_j(\pm\pi, \eta_j(\pm\pi))) = \infty$ holds, either the speed of wave propagation or the wave amplitude becomes unboundedly large. Indeed, as is done previously in the proof,

$$E = \tfrac{1}{2}(c - u(0,0))^2 = \tfrac{1}{2}(c - u(\pm\pi, \eta(\pm\pi)))^2 + g\eta(\pm\pi) + P_0,$$

from which it follows that

$$\begin{aligned}(c - u(\pm\pi, \eta_j(\pm\pi)))^2 &= (c - u(0,0))^2 + 2g|\eta(\pm\pi)| \\ &< \lambda + 2g|\eta(\pm\pi)|.\end{aligned}$$

Therefore, when the left side approaches infinity as $j \to \infty$,

$$\text{either} \quad \lim_{j\to\infty} c_j = \infty \quad \text{or} \quad \lim_{j\to\infty} |\eta_j(\pm\pi)| = \infty$$

holds.

Appendix A. Proof of Proposition 4.2 (local bifurcation). This section is devoted to the detailed analysis of local bifurcation for the regular approximate problem (3.11). We begin by invoking the theorem of local bifurcation from a simple eigenvalue due to Crandall and Rabinowitz.

THEOREM A.1 (see [11, Theorem 1] and [10, Theorem 3.6]). *Suppose that X and Y are Banach spaces and $I \subset \mathbb{R}$ is an open interval of λ^*. Also suppose that $F : I \times X \to Y$ is of class C^2 and satisfies*

(i) $F(\lambda, 0) = 0$ *for all* $\lambda \in I$*;*

(ii) $F_w(\lambda^*, 0)$ *is a Fredholm operator of index zero;*

(iii) $\ker F_w(\lambda^*, 0)$ *is one-dimensional and generated by* ϕ^**; and*

(iv) $F_{w\lambda}(\lambda^*, 0)[\phi^*] \notin \mathrm{range} F_w(\lambda^*, 0)$.

Then $(\lambda^*, 0)$ is a local bifurcation point. More precisely, there exists $s_0 > 0$ and a local curve of solutions

$$\{ (\lambda, w) = (\lambda(s), w(s)),\ |s| < s_0 : F(\lambda(s), w(s)) = 0 \}$$

in $\mathbb{R} \times X$ such that $(\lambda(0), w(0)) = (\lambda^*, 0)$. Furthermore, any solution near $(\lambda^*, 0)$ is of the form

$$w(s) = s\phi^* + o(s) \qquad \text{for} \quad |s| < s_0.$$

The curve of solutions is unique in the sense that there is an open neighborhood $U \subset I \times X$ of $(\lambda^*, 0)$ such that

$$\{ (\lambda, w) \in U : F(\lambda, w) = 0,\ w \not\equiv 0 \} = \{ (\lambda(s), w(s)) : 0 < |s| < s_0 \}.$$

Our task is to verify conditions (i) through (iv) for the setting in section 4.1. From the definition of F^ϵ in (3.11) follows that $F^\epsilon(\lambda, 0) = 0$ for $\lambda \in (2\Gamma_\infty, \infty)$. The Fredholm property of $F^\epsilon_w(\lambda^\epsilon, 0)$ is proved in Lemma 4.3.

LEMMA A.2 (null space). *The null space of $F^\epsilon_w(\lambda^\epsilon, 0)$ is one-dimensional and generated by ϕ^ϵ. Here ϕ^ϵ is an eigenfunction of (4.1) with $\lambda = \lambda^\epsilon$.*

Proof. As is established in Lemma 4.1, $\phi^\epsilon(q, p) = \Phi^\epsilon(p) \cos q$ belongs to $\ker F^\epsilon_w(\lambda^\epsilon, 0)$. Conversely, let $\phi \in X$ be in $\ker F^\epsilon_w(\lambda^\epsilon, 0)$. Such a function ϕ is even and 2π periodic in the q-variable, and hence it can be written as

$$\phi(q, p) = \sum_{k=0}^{\infty} \phi_k(p) \cos kq.$$

Each ϕ_k solves the ordinary differential equation

$$(a^3(\lambda^\epsilon)\phi_k')' - \epsilon\phi_k = k^2 a(\lambda^\epsilon)\phi_k \qquad \text{for } p \in (-\infty, 0),$$

with the boundary conditions

$$\begin{aligned} (\lambda^\epsilon)^{3/2}\phi_k'(0) &= g\phi_k(0), \\ \phi_k, \phi_k' \to 0 \qquad &\text{as } \ p \to -\infty. \end{aligned}$$

In case $k = 1$, it follows from Lemma 4.1 that ϕ_1 is a constant multiple of Φ^ϵ. For $k \geq 2$, any nontrivial solution ϕ_k would satisfy

$$\begin{aligned} R(\phi_k; \lambda^\epsilon) &= \frac{-g\phi_k^2(0) + \int_{-\infty}^0 a^3(\lambda^\epsilon)(\phi_k')^2 + \epsilon \int_{-\infty}^0 \phi_k^2\, dp}{\int_{-\infty}^0 a(\lambda^\epsilon)\phi_k^2\, dp} \\ &= -k^2 < -1. \end{aligned}$$

This contradicts (4.4), the characterization of λ^ϵ. Therefore, $\phi_k \equiv 0$ for $k \geq 2$. Consider a solution ϕ_0 of $k = 0$. From the comparison theorem for ordinary differential equations follows that any nontrivial solution ϕ_0 vanishes nowhere. This, however, contradicts the orthogonality,

$$\int_{-\infty}^{0} a(\lambda^\epsilon)\Phi^\epsilon(p)\phi_0(p)dp = 0,$$

which proves the assertion. ∎

Proof of Proposition 4.2. It remains to verify (iv) of Theorem A.1. We first claim that if $(f_1, f_2) \in Y$ belongs to the range of $F_w^\epsilon(\lambda^\epsilon, 0)$, then the orthogonality condition

$$\int_R f_1 \cdot a^3(\lambda^\epsilon)\phi^\epsilon dqdp + \frac{1}{2}\int_T f_2 \cdot \lambda^\epsilon \phi^\epsilon dq = 0 \tag{A.1}$$

holds. We remark that for any $f_1 \in Y_1$ the first integral converges since ϕ^ϵ decays exponentially as $p \to -\infty$ uniformly for q and $a(\lambda^\epsilon)$ is bounded.

To prove the claim, let $f_1 = A(\lambda^\epsilon, 0)v$ and $f_2 = B(\lambda^\epsilon, 0)v$ for some $v \in X$, i.e.,

$$\begin{aligned} f_1 &= a^{-3}(\lambda^\epsilon)(a^3(\lambda^\epsilon)v_p)_p + (a(\lambda^\epsilon)v_q)_q - \epsilon a^{-3}(\lambda^\epsilon)v, \\ f_2 &= (2g/\lambda^\epsilon)v - 2\sqrt{\lambda^\epsilon}v_p\big|_T. \end{aligned}$$

An integration by parts shows that

$$\begin{aligned} \iint_R f_1 \cdot a^3(\lambda^\epsilon)\phi^\epsilon\, dqdp &= \iint_R [(a^3(\lambda^\epsilon)v_p)_p + (a(\lambda^\epsilon)v_q)_q - \epsilon v]\phi^\epsilon\, dqdp \\ &= \iint_R [(a^3(\lambda^\epsilon)\phi^\epsilon_p)_p + (a(\lambda^\epsilon)\phi^\epsilon_q)_q - \epsilon\phi^\epsilon]v\, dqdp \\ &\quad + \int_T (a^3(\lambda^\epsilon)v_p\phi^\epsilon - a^3(\lambda^\epsilon)v\phi^\epsilon_p)\, dq \\ &= -\frac{1}{2}\int_T f_2 \cdot \lambda^\epsilon\phi^\epsilon\, dp. \end{aligned}$$

The last equality uses that ϕ^ϵ solves (4.1) with $\lambda = \lambda^\epsilon$. This proves the claim.

An explicit calculation yields

$$\begin{aligned} F^\epsilon_{w\lambda}(\lambda^\epsilon, 0)[\phi^\epsilon] = \big(&-a^{-4}(\lambda^\epsilon)\phi^\epsilon_{qq} - 3\gamma(-p)\, a^{-4}(\lambda^\epsilon)\phi^\epsilon_p + \frac{3}{2}\epsilon a^{-5}(\lambda^\epsilon)\phi^\epsilon, \\ &-2g(\lambda^\epsilon)^{-2}\phi^\epsilon - (\lambda^\epsilon)^{-1/2}\phi^\epsilon_p\big|_T\big). \end{aligned}$$

We shall show that (A.1) fails with $(f_1, f_2) = F^\epsilon_{w\lambda}(\lambda^\epsilon, 0)[\phi^\epsilon]$:

$$\begin{aligned} &\iint_R \left(-a^{-4}(\lambda^\epsilon)\phi^\epsilon_{qq} - 3\gamma(-p)a^{-4}(\lambda^\epsilon)\phi^\epsilon_p + \frac{3}{2}\epsilon a^{-5}(\lambda^\epsilon)\phi^\epsilon\right) \cdot a^3(\lambda^\epsilon)\phi^\epsilon\, dqdp \\ &\quad + \frac{1}{2}\int_T \left(2g(\lambda^\epsilon)^{-2}\phi^\epsilon - (\lambda^\epsilon)^{-1/2}\phi^\epsilon_p\right) \cdot \lambda^\epsilon\phi^\epsilon\, dq \\ &= -\iint_R a^{-1}(\lambda^\epsilon)\phi^\epsilon\phi^\epsilon_{qq}\, dqdp - 3\iint_R \gamma(-p)a^{-1}(\lambda^\epsilon)\phi^\epsilon\phi^\epsilon_p\, dqdp \\ &\quad - \frac{3}{2}\epsilon\iint_R a^{-2}(\lambda^\epsilon)(\phi^\epsilon)^2\, dqdp \\ &\quad + \frac{1}{2}\int_T \left(-2g(\lambda^\epsilon)^{-1}(\phi^\epsilon)^2 - \sqrt{\lambda^\epsilon}\phi^\epsilon\phi^\epsilon_p\right)\, dq \end{aligned}$$

$$
\begin{aligned}
&= \iint_R a^{-1}(\lambda^\epsilon)(\phi^\epsilon)^2\,dqdp - 3\iint_R \gamma(-p)a^{-1}(\lambda^\epsilon)\phi^\epsilon\phi^\epsilon_p\,dqdp\\
&\quad - \frac{3}{2}\epsilon\iint_R a^{-2}(\lambda^\epsilon)(\phi^\epsilon)^2\,dqdp - \frac{3}{2}\int_T g(\lambda^\epsilon)^{-1}(\phi^\epsilon)^2\,dq\\
&= -\frac{1}{2}\iint_R a^{-1}(\lambda^\epsilon)(\phi^\epsilon)^2\,dqdp - \frac{3}{2}\iint_R a(\lambda^\epsilon)(\phi^\epsilon_p)^2\,dqdp < 0.
\end{aligned}
$$

The last equality uses (4.1) with $\lambda = \lambda^\epsilon$ and $a'(p;\lambda^\epsilon) = \gamma(-p)a^{-1}(p;\lambda^\epsilon)$:

$$
\begin{aligned}
&\iint_R \gamma(-p)a^{-1}(\lambda^\epsilon)\phi^\epsilon\phi^\epsilon_p\,dqdp\\
&\qquad = \frac{1}{2}\iint_R a^{-1}(\lambda^\epsilon)(\phi^\epsilon)^2\,dqdp + \frac{1}{2}\iint_R a(\lambda^\epsilon)(\phi^\epsilon_p)^2\,dqdp\\
&\qquad\qquad + \frac{1}{2}\epsilon\iint_R a^{-2}(\lambda^\epsilon)(ep)^2\,dqdp - \frac{1}{2}\int_T g(\lambda^\epsilon)^{-1}(\phi^\epsilon)^2\,dq.
\end{aligned}
$$

Therefore, $F^\epsilon_{w\lambda}(\lambda^\epsilon,0)[\phi^\epsilon] \notin \mathrm{range} F^\epsilon_w(\lambda^\epsilon,0)$.

We deduce from Theorem A.1 the existence of a local bifurcation curve $(\lambda(s),w(s)) \in \mathcal{C}^\epsilon_{loc}$ in $\mathbb{R}\times X$ such that $(\lambda(0),w(0)) = (\lambda^\epsilon,0)$ and $F^\epsilon(\lambda(s),w(s)) = 0$. ◻

Appendix B. Generalized degree theory. This section is the summary of the general development in [18, 23] to define a generalized degree for a large class of mappings which includes $F^\epsilon(\lambda,\cdot)$ defined in (3.11). Let X, Y_1, and Y_2 be real Banach spaces, with X continuously embedded in Y_1. The Banach space $Y = Y_1\times Y_2$ is equipped with the product topology. Let $\mathcal{W}$ be a bounded open set of X and $F : \overline{\mathcal{W}} \to Y$ be of $C^2(\mathcal{W},Y)\cap C^0(\overline{\mathcal{W}},Y)$. The mapping F is supposed to be proper, i.e., $F^{-1}(K)\cap\overline{\mathcal{W}}$ is compact in X for any $K\subset Y$ compact. The linearized operator $F_w(w) = (A(w),B(w))$ is *admissible* in the following sense:

(L1) F_w is a Fredholm operator of index zero;
(L2) $B : X\to Y_2$ is surjective;
(L3) for each $w\in\mathcal{W}$ there exist $C_1, C_2 > 0$ such that

$$C_1\|\phi\|_X \le \mu^{\alpha/2}\|(A(w)-\mu I)\phi\|_{Y_1} + \mu^{(\alpha+1)/2}\|B(w)\phi\|_{Y_2}$$

for all $\phi\in X$ and for all $\mu\ge C_2$; and
(L4) there exists an open neighborhood N of the ray $\{\mu\in\mathbb{R},\,\mu\ge 0\}$ in the complex plane such that $\sigma(A)\cap N$ consists of finitely many eigenvalues, each of finite algebraic multiplicity; the spectrum $\sigma(A)$ is defined in (4.13).

Consider the equation

$$F(w) = y \quad \text{for } y\notin F(\partial\mathcal{W}).$$

We assume for now that y is a regular value, i.e., $F_w(w) = (A(w),B(w))$ is surjective for all $w\in F^{-1}(y)\cap\mathcal{W}$. It is in turn bijective by the Fredholm property. Then (L3) and (L4) ensure that $A(w)$ has only finitely many positive real eigenvalues, $\nu(w)$ say, counted by algebraic multiplicity. By properness, the inverse image $F^{-1}(y)\cap\mathcal{W}$ is compact and thereby a finite set. Accordingly, we define the *degree* of F with respect to y by

$$\deg(F,\mathcal{W},y) = \sum_{w\in F^{-1}(y)\cap\mathcal{W}} (-1)^{\nu(w)}. \tag{B.1}$$

Set $\deg(F,\mathcal{W},y) = 0$ if $F^{-1}(y)\cap\mathcal{W} = \emptyset$.

Suppose that $y \notin F(\partial\mathcal{W})$ is not a regular value. For this we employ the Sard–Smale–Quinn theorem which asserts that the set of regular values of F is dense in Y. We may choose a nearby regular value $\tilde{y}$ of F with $\|\tilde{y} - y\|_Y \le \inf_{\hat{y}\in F(\partial\mathcal{W})} \|\hat{y} - y\|_Y$ and define

$$\deg(F, \mathcal{W}, y) = \deg(F, \mathcal{W}, \tilde{y}), \tag{B.2}$$

where the right side is well defined.

An important property of degree is the following generalized homotopy invariance valid on noncylindrical domains. Let $\mathcal{U} \subset [0,1] \times \mathcal{W}$ be open and

$$\mathcal{U}_t = \{w \in \mathcal{W} : (t, w) \in \mathcal{U}\}, \qquad \partial\mathcal{U}_t = \{w \in \mathcal{W} : (t, w) \in \partial\mathcal{U}\}.$$

By an *admissible homotopy* we mean a proper map $H : \overline{\mathcal{U}} \to Y$ of class C^2 such that $H(t, \cdot) : \mathcal{U}_t \to Y$ is admissible in the sense of (L1)–(L4) above for each $t \in [0,1]$.

Proposition B.1 (generalized homotopy invariance). *The degree defined by* (B.1) *and* (B.2) *is invariant under admissible homotopies, i.e.,*

$$\deg(H(0,\cdot), \mathcal{U}_0, y) = \deg(H(1,\cdot), \mathcal{U}_1, y)$$

provided $y \notin H(t, \partial\mathcal{U}_t)$ *for all* $t \in [0,1]$.

The proof is standard. See [18, Appendix] for one in the present setting.

Appendix C. Preservation of nodal structure. Here we demonstrate that the continuum obtained in Theorem 4.7 globally preserves the nodal configuration which it inherits from the eigenfunction of the linearized problem at the local bifurcation point. Accordingly, the second alternative in Theorem 4.7 is eliminated.

As is done in section 4.2, let

$$R^+ = (0, \pi) \times (-\infty, 0), \qquad T^+ = (0, \pi) \times \{0\},$$
$$\partial R^+_l = \{(0, p) : p \in (-\infty, 0)\}, \qquad \partial R^+_r = \{(\pi, p) : p \in (-\infty, 0)\}.$$

Our goal is to show that any nontrivial solution to (3.11) in $\mathcal{C}^\epsilon_\delta$ possesses the following nodal pattern:

$$w_q < 0 \quad \text{in } R^+ \cup T^+, \tag{C.1}$$
$$w_{qq} < 0 \quad \text{on } \partial R^+_l, \qquad w_{qq} > 0 \quad \text{on } \partial R^+_r, \tag{C.2}$$
$$\text{either} \quad w_{qq}(0,0) < 0 \quad \text{or} \quad w_{qqp}(0,0) > 0, \tag{C.3}$$
$$\text{either} \quad w_{qq}(\pi,0) > 0 \quad \text{or} \quad w_{qqp}(\pi,0) < 0. \tag{C.4}$$

Inequalities (C.1)–(C.4) define the open set $\mathcal{N}$; see (4.23). By evenness and periodicity of $w \in X$, we infer that $w_q = 0$ on $\partial R^+_l \cup \partial R^+_r$.

Lemma C.1. $\mathcal{C}^\epsilon_{loc} \cap (\mathbb{R} \times \mathcal{N}) \neq \emptyset$*; i.e.,* (C.1)–(C.4) *hold along the local curve* $\mathcal{C}^\epsilon_{loc} \setminus (\lambda^\epsilon, 0)$ *of bifurcation in a small neighborhood of* $(\lambda^\epsilon, 0)$ *in* $\mathbb{R} \times X$.

Proof. We begin with the observation that an eigenfunction $\phi^\epsilon(q,p) = \Phi^\epsilon(p)\cos q$ of (4.1) at $(\lambda^\epsilon, 0)$ belongs to $\mathcal{N}$. This uses the fact that

$$\Phi^\epsilon(p) > 0 \quad \text{for } p \in (-\infty, 0] \qquad \text{and} \qquad (\Phi^\epsilon)'(0) > 0. \tag{C.5}$$

The alternatives (C.3) and (C.4) hold since $\phi^\epsilon_{qqp}(0,0) < 0$ and $\phi^\epsilon_{qq}(\pi,0) > 0$.

By virtue of Proposition 4.2, any solution of (3.11) contained in $\mathbb{R}\times X$ near $(\lambda^\epsilon, 0)$ is of the form

$$w(q,p) = s\Phi^\epsilon(p)\cos q + o(s) \quad \text{in } C^{3+\alpha}(\overline{R})$$

for $s > 0$ sufficiently small. By restricting it to $\overline{R}^+$ and differentiating, we arrive at

$$\begin{aligned} w_q(q,p) &= -s\Phi^\epsilon(p)\sin q &&+ o(s) \quad \text{in } C^{2+\alpha}(R^+ \cup T^+), \\ w_{qq}(0,p) &= -s\Phi^\epsilon(p) &&+ o(s) \quad \text{in } C^{1+\alpha}(\partial R_l^+), \\ w_{qq}(\pi,p) &= \ \ s\Phi^\epsilon(p) &&+ o(s) \quad \text{in } C^{1+\alpha}(\partial R_r^+), \\ w_{qqp}(q,p) &= -s(\Phi^\epsilon)'(p)\cos q &&+ o(s) \quad \text{in } C^{\alpha}(\overline{R}^+). \end{aligned} \tag{C.6}$$

Choose $s_1 > 0$ small enough, and (C.1), (C.2) follow for $w(s)$ with $0 < |s| < s_1$.

Next, at the left corner point $(0,0)$, oddness of w_q leads to

$$w_q(0,0) = w_{qp}(0,0) = w_{qqq}(0,0) = w_{qpp}(0,0) = 0.$$

Accordingly, we may write down Taylor expansions in the neighborhood $\overline{R}^+ \cap B_{1/k}((0,0))$ of w_q and w_{qq} as

$$\begin{aligned} w_q(q,p) &= w_{qq}(0,0)q + w_{qqp}(0,0)qp + O(k^{-3}), \\ w_{qq}(q,p) &= w_{qq}(0,0) \ \ + w_{qqp}(0,0)p \ \ + O(k^{-2}), \end{aligned} \tag{C.7}$$

respectively. Here $B_r(Q)$ denotes the open ball in the (q,p)-plane of radius r centered at Q. Analogous expansions are valid at the right corner point $(\pi, 0)$,

$$\begin{aligned} w_q(q,p) &= w_{qq}(\pi,0)(q-\pi) + w_{qqp}(\pi,0)(q-\pi)p + O(k^{-3}), \\ w_{qq}(q,p) &= w_{qq}(\pi,0) \qquad\quad + w_{qqp}(\pi,0)p \qquad\quad + O(k^{-2}), \end{aligned}$$

in $\overline{R}^+ \cap B_{1/k}((\pi,0))$. From (C.5) and (C.6) it follows that

$$\begin{aligned} w_{qq}(0,0) &\le -\tfrac{s}{2}\Phi^\epsilon(0) < 0, \quad & w_{qqp}(0,0) &\le -\tfrac{s}{2}(\Phi^\epsilon)'(0) < 0, \\ w_{qq}(\pi,0) &\ge \ \ \tfrac{s}{2}\Phi^\epsilon(0) > 0, \quad & w_{qqp}(\pi,0) &\ge \ \ \tfrac{s}{2}(\Phi^\epsilon)'(0) > 0 \end{aligned} \tag{C.8}$$

for $s > 0$ sufficiently small.

In view of (C.7) and (C.8), we deduce that for large $k_2 > 0$ an integer and $0 < s_2 < s_1$ sufficiently small, (C.1)–(C.3) hold for $w(s)$ with $0 < s < s_2$, w restricted to $B_{1/k_2}((0,0))$. We proceed similarly at the right corner point $(\pi, 0)$ to conclude that (C.1), (C.2), and (C.4) hold for $w(s)$ restricted to $B_{1/k_3}((\pi,0))$, with $0 < s_3 < s_1$, for a large integer $k_3 \ge 0$ and $0 < s_3 < s_1$ small enough. Finally, the assertion follows if we take $s = \min(s_2, s_3)$ and consider $\overline{R}^+ = R^+ \cup T^+ \cup \partial R_l^+ \cup \partial R_r^+ \cup \overline{B}_{1/k_2}((0,0)) \cup \overline{B}_{1/k_3}((\pi,0))$. ∎

For our next result, we appeal to the following sharp form of the maximum principle.

Lemma C.2. *Let R^+ be the open semi-infinite strip $(0,\pi)\times(-\infty,0)$ and let w be a C^2-subsolution of the uniformly elliptic differential operator*

$$L = a_{11}(q,p)\partial_q^2 + 2a_{12}(q,p)\partial_p\partial_q + a_{22}(q,p)\partial_p^2 + b_1(q,p)\partial_q + b_2(q,p)\partial_p$$

in R^+. We assume that the coefficients a_{ij}, b_i $(i, j = 1, 2)$ are continuous and uniformly bounded on $\overline{R}^+$. Suppose $w \geq 0$ in $\overline{R}^+$ and $w \to 0$ as $p \to -\infty$ uniformly for q.

(a) *$w > 0$ in R^+ unless $u \equiv 0$.*

(b) *We further assume that there is a constant $k > 0$ such that*

$$|a_{12}(q,p)| \leq k \min(q, \pi - q) \tag{C.9}$$

for all $(q,p) \in \overline{R}^+$. If $w = 0$ at some corner point Q of $\overline{R}^+$, then

$$\text{either} \quad \frac{\partial w}{\partial \vec{s}}(Q) > 0 \quad \text{or} \quad \frac{\partial^2 w}{\partial \vec{s}^2}(Q) > 0 \tag{C.10}$$

unless $w \equiv 0$. Here $\vec{s}$ is any direction vector at Q which enters R^+ nontangentially.

Assertion (a) follows from the Phragmén-Lindelöf theorem [14] and the weak maximum principle. Assertion (b) is the edge point lemma due to Serrin [35, Lemma 2].

LEMMA C.3. *The nodal properties* (C.1)–(C.4) *hold along the entire continuum $\mathcal{C}^\epsilon_\delta \setminus (\lambda^\epsilon, 0)$ unless $(\lambda, 0) \in \mathcal{C}^\epsilon_\delta$ with $\lambda \neq \lambda^\epsilon$.*

Proof. Assume the contrary. Then by the previous lemma and the connectedness of $\mathcal{C}^\epsilon_\delta$, there must be a nontrivial solution $(\hat{\lambda}, \hat{w}) \in \mathcal{C}^\epsilon_\delta$ which belongs to the boundary of $\mathbb{R} \times \mathcal{N}$. That is, $\hat{w}_q \not\equiv 0$, $\hat{w}_q \leq 0$ on $\overline{R}^+$, and at least one of the following conditions holds:

(1) $\hat{w}_q = 0$ at some point $(\hat{q}, 0) \in T^+$;

(2) $\hat{w}_q = 0$ and $\hat{w}_{qq} = 0$ at some point $(\hat{q}, \hat{p}) \in \overline{R}^+ \setminus ((0,0) \cup (\pi, 0))$;

(3) $\hat{w}_{qq} = 0$ and $\hat{w}_{qqp} = 0$ at some corner point.

Observe that $\hat{w}_q$ solves the linear partial differential equation in R^+

$$\begin{aligned} A(\hat{\lambda}, \hat{w})[\hat{w}_q] &= \tilde{A}(\hat{\lambda}, \hat{w})[\hat{w}_q] \\ &\quad + [-2\hat{w}_q \hat{w}_{pq} + 3\gamma(-p)(a^{-1}(\hat{\lambda}) + \hat{w}_p)^2]\hat{w}_{qp} \\ &\quad + [2\hat{w}_q \hat{w}_{pp} - 2\gamma(-p) a^{-3}(\hat{\lambda}) \hat{w}_q]\hat{w}_{qq} = 0 \end{aligned} \tag{C.11}$$

with the boundary condition

$$\begin{aligned} B(\hat{\lambda}, \hat{w})[\hat{w}_q] &= 2(2g\hat{w} - \hat{\lambda})(\hat{\lambda}^{-1/2} + \hat{w}_p)\hat{w}_{qp} \\ &\quad + 2\hat{w}_q \hat{w}_{qq} + 2g(\hat{\lambda}^{-1/2} + \hat{w}_p)^2 \hat{w}_q\big|_T = 0. \end{aligned} \tag{C.12}$$

Since $(\hat{\lambda}, \hat{w}) \in \mathcal{C}^\epsilon_\delta \subset \mathcal{O}_\delta$, $A(\hat{\lambda}, \hat{w})$ is uniformly elliptic and its coefficients are bounded and continuous in $\overline{R}^+$. By continuity, $\hat{w}_q \leq 0$ in $\overline{R}^+$. The bluntness condition (C.9) is fulfilled since $w_q = 0$ on $\partial R^+_l \cup \partial R^+_r$.

We first suppose $\hat{w}_q(\hat{q}, 0) = 0$ for some $(\hat{q}, 0) \in T^+$. At such a point the boundary condition becomes

$$2(2g\hat{w} - \hat{\lambda})(\hat{\lambda}^{-1/2} + \hat{w}_p)\hat{w}_{qp} = 0.$$

The Hopf boundary lemma requires $\hat{w}_{qp}(\hat{q}, 0) > 0$. Since $\hat{\lambda}^{-1/2} + \hat{w}_p \geq \delta > 0$ for $(\hat{\lambda}, \hat{w}) \in \mathcal{O}_\delta$, it must hold $2g\hat{w} - \hat{\lambda} = 0$. This contradicts the boundary condition on top (3.2b) since

$$1 \leq 1 + \hat{w}_q^2 = 1 + \hat{w}_q^2 + (2g\hat{w} - \hat{\lambda})(\hat{\lambda}^{-1/2} + \hat{w}_p)^2 = 0 \qquad \text{at } (\hat{q}, 0).$$

Therefore, (C.1) holds for $\hat{w}_q$.

Next, provided that $\hat{w}_q \leq 0$ in $\overline{R}^+$, we employ Lemma C.2(a) and the Hopf boundary lemma to conclude that condition (2) implies that $\hat{w}_q \equiv 0$.

Finally, suppose that condition (3) holds at $(0,0)$. By oddness of w_q, we infer that

$$w_q(0,0) = w_{pq}(0,0) = w_{qqq}(0,0) = w_{qpp}(0,0) = 0.$$

Hence,

$$\frac{\partial w_q}{\partial \vec{s}}(0,0) = 0 \quad \text{and} \quad \frac{\partial^2 w_q}{\partial \vec{s}^2}(0,0) = 0$$

for any direction $\vec{s}$ which enters R^+ nontangentially. This contradicts Lemma C.2(b). The same consideration at $(\pi, 0)$ excludes possibility (3). This completes the proof. ∎

At this stage, we have shown that (C.1)–(C.4) hold along the entire continuum $\mathcal{C}_\delta^\epsilon$ except in the case $(\lambda, 0) \in \mathcal{C}_\delta^\epsilon$ with $\lambda \neq \lambda^\epsilon$. Lemma 4.8 rules out this possibility and proves the preservation of the nodal configuration globally along the continuum.

Acknowledgment. This work was done as part of the author's doctoral dissertation. The author would like to thank her advisor, Walter Strauss, for his excellent guidance and for reading many drafts of this work.

REFERENCES

[1] S. Agmon, *On the eigenfunctions and on the eigenvalues of general elliptic boundary value problems*, Comm. Pure Appl. Math., 15 (1962), pp. 119–147.

[2] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[3] C. Amick, *Bounds for water waves*, Arch. Rational Mech. Anal., 99 (1987), pp. 91–114.

[4] C. Amick, L. Fraenkel, and J. Toland, *On the Stokes conjecture for the wave of extreme form*, Acta Math., 148 (1982), pp. 193–214.

[5] C. Amick and J. Toland, *On solitary water-waves of finite amplitude*, Arch. Rational Mech. Anal., 76 (1981), pp. 9–95.

[6] A. Constantin, *On the deep water wave motion*, J. Phys. A, 34 (2001), pp. 1405–1417.

[7] A. Constantin and J. Escher, *Symmetry of steady deep-water waves with vorticity*, European J. Appl. Math., 15 (2004), pp. 755–768.

[8] E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, 1955.

[9] A. Constantin and W. Strauss, *Exact periodic traveling water waves with vorticity*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 797–800.

[10] A. Constantin and W. Strauss, *Exact steady periodic water waves with vorticity*, Comm. Pure Appl. Math., 57 (2004), pp. 481–527.

[11] M. Crandall and P. Rabinowitz, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.

[12] M.-L. Dubreil-Jacotin, *Sur la détermination rigoreuse des ondes permanentes périodiques d'ampleur finie*, J. Math. Pures Appl., 13 (1934), pp. 217–291.

[13] F. Gerstner, *Theorie der Wellen samt einer daraus abgeleiteten Theorie der Deichprofile*, Ann. Phys., 2 (1809), pp. 412–445.

[14] D. Gilbarg, *The Phragmén-Lindelöf theorem for elliptic partial differential equations*, J. Rational Mech. Anal., 1 (1952), pp. 411–417.

[15] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 2001.

[16] R. Goyon, *Contribution à la théorie des houles*, Ann. Fac. Sci. Univ. Toulouse, 22 (1958), pp. 1–55.

[17] T. Healey and H. Kielhöfer, *Symmetry and nodal properties in the global bifurcation analysis of quasi-linear elliptic equations*, Arch. Ration. Mech. Anal., 113 (1991), pp. 299–311.

[18] T. Healey and H. Simpson, *Global continuation in nonlinear elasticity*, Arch. Ration. Mech. Anal., 143 (1998), pp. 1–28.

[19] R. S. Johnson, *A Modern Introduction to the Mathematical Theory of Water Waves*, Cambridge University Press, Cambridge, UK, 1997.

[20] S. Kaniel and M. Schechter, *Spectral theory for Fredholm operators*, Comm. Pure Appl. Math., 16 (1963), pp. 423–448.

[21] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1967.

[22] G. Keady and J. Norbury, *On the existence theory for irrotational water waves*, Math. Proc. Cambridge Philos. Soc., 83 (1978), pp. 137–157.

[23] H. Kielhöfer, *Multiple eigenvalue bifurcation for Fredholm operator*, J. Reine Angew. Math., 358 (1985), pp. 104–124.

[24] J. Krasovskiĭ, *On the theory of steady-state waves of finite amplitude*, Ž. Vyčisl. Mat. i Mat. Fiz., 1 (1961), pp. 836–855 (in Russian); U.S.S.R. Comput. Math. and Math. Phys., 1 (1961), pp. 996–1018 (in English).

[25] N. V. Krylov, *Lectures on Elliptic and Parabolic Equations in Hölder Spaces*, AMS, Providence, RI, 1996.

[26] T. Levi-Civita, *Determinazione rigorosa delle onde irrotazionali periodiche in acqua profonda*, Rend. Accad. Lincei, 33 (1924), pp. 141–150.

[27] G. Lieberman and N. Trudinger, *Nonlinear oblique boundary value problems for nonlinear elliptic equations*, Trans. Amer. Math. Soc., 295 (1986), pp. 509–546.

[28] J. Lighthill, *Waves in Fluids*, Cambridge University Press, Cambridge, UK, 1978.

[29] J. McLeod, *The Stokes and Krasovskii conjectures for the wave of greatest height*, Stud. Appl. Math., 98 (1997), pp. 311–334.

[30] A. Nekrasov, *On steady waves*, Izv. Ivanovo-Voznesenk. Politekhn., 3 (1921), p. 40.

[31] H. Okamoto and M. Shoji, *The Mathematical Theory of Permanent Progressive Water Waves*, Adv. Ser. Nonlinear Dynam. 20, World Scientific, River Edge, NJ, 2001.

[32] J. Peetre, *Another approach to elliptic boundary value problems*, Comm. Pure Appl. Math., 14 (1961), pp. 711–731.

[33] P. Rabinowitz, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.

[34] M. Schechter, *Principles of Functional Analysis*, 2nd ed., Grad. Stud. Math. 36, AMS, Providence, RI, 2001.

[35] J. Serrin, *A symmetry problem in potential theory*, Arch. Ration. Mech. Anal., 43 (1971), pp. 304–318.

[36] J. Simmen and P. Saffman, *Steady deep-water waves on a linear shear current*, Stud. Appl. Math., 73 (1985), pp. 35–57.

[37] G. G. Stokes, *Considerations relative to the greatest height of oscillatory irrotational waves which can be propagated without change of form*, in Mathematical and Physical Papers, Vol. 1, Cambridge University Press, Cambridge, UK, 1880, pp. 225–228.

[38] J. Toland, *On the existence of a wave of greatest height and Stokes's conjecture*, Proc. Roy. Soc. London Ser. A, 363 (1978), pp. 469–485.

[39] J. Toland, *Stokes waves*, Topol. Methods Nonlinear Anal., 7 (1996), pp. 1–48.

[40] J. M. Vanden-Broeck, *Periodic waves with constant vorticity in water of infinite depth*, IMA J. Appl. Math., 56 (1996), pp. 207–217.

[41] V. Volpert and A. Volpert, *Properness and topological degree for general elliptic operators*, Abst. Appl. Anal., 8 (2003), pp. 129–181.

[42] E. Zeidler, *Existenzbeweis für permanente Kapillar-Schwerewellen mit allgemeinen Wirbelverteilungen*, Arch. Ration. Mech. Anal., 50 (1973), pp. 34–72.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1522–1534

ASYMPTOTIC STABILITY IN A NEUTRAL DELAY DIFFERENTIAL SYSTEM WITH VARIABLE DELAYS*

WU-HUA CHEN†, XIAOMEI LU†, ZHI-HONG GUAN‡, AND WEI XING ZHENG§

Abstract. By applying M-matrix theory and some new analysis techniques, we have succeeded in establishing the asymptotic stability of the zero solution of the "pure-delay type" neutral system with variable delays. The criteria obtained in this paper extend and improve the existing ones.

Key words. neutral delay system, pure-delay type, asymptotic stability, M-matrix

AMS subject classifications. 34K20, 34K60

DOI. 10.1137/S0036141004441167

1. Introduction. In this paper, we will be concerned with establishing new results for asymptotic stability of the trivial solution of the perturbed nonlinear neutral systems of the form

$$\begin{aligned} \frac{d}{dt}\left(x_i(t) - p_i(t)x_i(t-q_i)\right) = &-b_i(t)x_i(t-\tau_i(t)) \\ &+ f_i\left(t, x_1(t-\tau_{i1}(t)), \dots, x_n(t-\tau_{in}(t))\right), \\ &\qquad i = 1, 2, \dots, n, \end{aligned} \tag{1.1}$$

where $q_i > 0$ is a constant, $p_i(t),\ b_i(t) \geq 0,\ \tau_i(t) \geq 0$, and $\tau_{ij}(t) \geq 0$ are continuous on $R_+ = [0, +\infty)$, $i, j = 1, 2, \dots, n$, and $|p_i(t)| \leq p_i$, $\tau_i(t) \leq \tau_i$, $\tau_{ij}(t) \leq \tau_{ij}$. For each $i \in \{1, 2, \dots, n\}$, we assume that $f_i : R_+ \times R^n \to R$ is a continuous function and there exists a nonnegative matrix $D = (d_{ij})_{n\times n}$ such that

$$|f_i(t, x_1, x_2, \dots, x_n)| \leq b_i(t)\sum_{j=1}^{n} d_{ij}|x_j|,\ i = 1, 2, \dots, n. \tag{1.2}$$

When $p_i(t) = 0,\ \tau_i(t) = 0$, and $b_i(t) \equiv b_i$, $i = 1, 2, \dots, n$, (1.1) reduces to

$$\dot{x}_i(t) = -b_i x_i(t) + f_i\left(t, x_1(t-\tau_{i1}(t)), \dots, x_n(t-\tau_{in}(t))\right),\ i = 1, 2, \dots, n. \tag{1.3}$$

Equation (1.1) is usually used to model the delayed neural networks and has many important applications. The interested reader can refer to [1] and [2] and the references therein. A typical result for asymptotic stability of system (1.1) is that if

$$|f_i(t, x_1, x_2, \dots, x_n)| \leq \sum_{j=1}^{n} |a_{ij}||x_j|,\ i = 1, 2, \dots, n, \tag{1.4}$$

---

*Received by the editors February 18, 2004; accepted for publication (in the revised form) July 6, 2005; published electronically January 12, 2006.
http://www.siam.org/journals/sima/37-5/44116.html

†College of Mathematics and Information Science, Guangxi University, Nanning, Guangxi, 530004, People's Republic of China (wuhua_chen@163.com, wuhua_chen2002@yahoo.com.cn, lu_xiaomei@126.com). The research of the first author was partially supported by the Guangxi Natural Science Foundation (0542032) and the National Natural Science Foundation of China under grant 10461001.

‡Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, People's Republic of China (zhguan@mail.hust.edu.cn).

§School of QMMS, University of Western Sydney, Sydney, NSW 1797, Australia (w.zheng@uws.edu.au). The research of this author was partially supported by a research grant from the Australian Research Council.

and $\mathrm{diag}\{b_1, b_2, \ldots,\ b_n\} - |A|$ is a nonsingular M-matrix, where $|A| = (|a_{ij}|)_{n\times n}$, then the trivial solution of system (1.4) is globally asymptotically stable; see, for example, [17] .

When $\tau_i(t) \neq 0,\ i = 1, \ldots, n$, system (1.1) is referred to as a "pure-delay type" system [3], [4], [5] which has many important applications, such as in congestion control for the internet [15] and in delay feedback control for chaotic systems [16]. For pure-delay type systems, the stability problem becomes much harder since there is no linear nondelayed term that dominates the others. In [14], Györi considered the stability problem for a special case of system (1.1):

$$\dot{x}_i(t) = -b_i x_i(t-\tau_i) + f_i\left(t, x_1(t-\tau_{i1}), \ldots, x_n(t-\tau_{in})\right),\ i = 1, 2, \ldots, n. \tag{1.5}$$

He established the following result.

THEOREM 1.1. *If condition* (1.4) *is satisfied,* $b_i\tau_i < 1/\mathrm{e},\ i = 1, 2, \ldots, n,$ *and*

$$\mathrm{diag}\{b_1, b_2, \ldots,\ b_n\} - |A| \tag{1.6}$$

is a nonsingular M-matrix, then the trivial solution of system (1.5) *is globally asymptotically stable for all constant delays* τ_{ij}.

The main techniques used in [14] are based on the asymptotic representation of the solutions of linear differential equations with constant coefficients and constant delays, so the method in [14] seems to be useless for system (1.1) with variable delays and variable coefficient.

When $n = 1$, (1.1) reduces to the scalar neutral delay differential equation

$$\frac{d}{dt}\left(x(t) - p(t)x(t-q)\right) = -b(t)x(t-\tau(t)) + a(t)x(t-r(t)), \tag{1.7}$$

where $p(t),\ b(t),\ a(t),\ \tau(t),\ r(t)$ are continuous on $[0, +\infty)$, and $b(t) \geq 0$, $0 \leq \tau(t) \leq \tau$, $0 \leq r(t) \leq r$, q is a positive constant.

It is shown in Yu [10] that if

$$\int_t^{t+\tau} b(s)ds < \frac{3}{2} - 2p(2-p), \quad \int_0^{+\infty} b(s)ds = +\infty,$$

then the zero solution of (1.7) with $a(t) = 0$ is asymptotically stable. This is the first 3/2 stability result in the literature for neutral differential equations. In [19], Bartha investigated the stability properties of neutral differential equations with state-dependent delay. Recently, Wang and Liao [13] studied the global attractivity of the zero solution of (1.7) under the assumptions of $\tau(t) = \tau$ and $r(t) = r$. By making the change of variable

$$z(t) = x(t) - p(t)x(t-q) - \int_{t+r-\tau}^{t} a(s)ds, \tag{1.8}$$

Wang and Liao transformed (1.7) with $\tau(t) = \tau$ and $r(t) = r$ into the following neutral differential equation with distributed delay:

$$\dot{z}(t) = -\bar{b}(t)x(t-\tau),$$

where $\bar{b}(t) = b(t) - a(t+r-\tau)$. Then applying the modified techniques used in [10], they derived the following theorem.

THEOREM 1.2. *Consider* (1.7) *with* $\tau(t) = \tau$ *and* $r(t) = r$. *Set* $\eta = \sup_{t\geq 0}\int_t^{t+\tau} b(s)ds$. *Assume that there exists a constant* $p \geq 0$ *such that*

$$|p(t)| + \int_{t-(\tau-r)}^{t} \operatorname{sign}(\tau - r)|a(s)|ds \leq p, \tag{1.9}$$

and

$$p < \frac{1}{4},\ \eta < \frac{3}{2} - 2p\,, \tag{1.10}$$

or

$$\frac{1}{4} \leq p < \frac{1}{2},\ \eta < \sqrt{2(1-2p)}\,. \tag{1.11}$$

If

$$\bar{b}(t) \geq 0,\ \int_0^{+\infty} \bar{b}(t)dt = +\infty,$$

then every solution of (1.7) *tends to zero as* $t \to +\infty$.

The techniques in [13] are based on the variable change (1.8), so the method in [13] is not applicable to (1.7) with variable delays. Moreover, the result in [13] depends on the delay of the perturbation (see condition (1.9)) and cannot guarantee the zero solution of (1.7) to be asymptotically stable.

Recently, So, Tang, and Zou [11] studied the global attractivity for the following linear differential system with constant delays:

$$\dot{x}_i(t) = -b_i x(t - \tau_i) + \sum_{j=1}^{n} a_{ij} x_j(t - \tau_{ij}),\ i = 1, 2, \ldots, n. \tag{1.12}$$

By using M-matrix theory and some inequality techniques, So, Tang, and Zou [11] proved the following theorem.

THEOREM 1.3. *Let* $\hat{A} = (\hat{a}_{ij})$ *be defined by*

$$\hat{a}_{ij} = \frac{1 + 1/9 b_i\tau_i(3 + 2b_i\tau_i)}{1 - 1/9 b_i\tau_i(3 + 2b_i\tau_i)}|a_{ij}|,\ i, j = 1, \ldots, n.$$

If $b_i\tau_i < 3/2,\ i = 1, \ldots, n,$ *and*

$$\operatorname{diag}\{b_1, \ldots, b_n\} - \hat{A} \tag{1.13}$$

is a nonsingular M-matrix, then every solution of (1.12) *tends to zero as* $t \to \infty$.

The techniques in [11] may be applicable to studying the global attractivity for system (1.1); however, they cannot be used to investigate the asymptotic stability for nonlinear system (1.1) with variable delays. Therefore, the stability problem for system (1.1) has not been fully investigated. In this paper, we will consider the following question: When all subsystems of (1.1),

$$\frac{d}{dt}\left(x_i(t) - p_i(t)x_i(t - q_i)\right) = -b_i(t)x_i(t - \tau_i(t)),\ i = 1, 2, \ldots, n,$$

are asymptotically stable, what conditions should be given on the perturbation terms $f_i(t, x_1(t - \tau_{i1}(t)), \ldots, x_n(t - \tau_{in}(t))),\ i = 1, 2, \ldots, n,$ such that the trivial solution of perturbed system (1.1) is globally asymptotically stable for all bounded variable delays $\tau_{ij}(t)$?

Before moving on, we need to introduce some notation and definitions.

For vector $x = (x_1, x_2, \dots, x_n)^T \in R^n$, $|x|$ denotes a vector norm defined by $|x| = \max_{1\le i\le n}\{|x_i|\}$. Let $h = \max_{1\le i,j\le n}\{q_i, \tau_i, \tau_{ij}\}$. $C = C([-h,\ 0],\ R^n)$ denotes the space of bounded, continuous functions $\phi:\ [-h,\ 0] \to R^n$ with norm $\|\phi\| = \sup_{s\in[-h,\ 0]}|\phi(s)|$. If $y \in C([-h,\ \alpha),\ R^n)$ with $\alpha > 0$ and $t \in [0,\ \alpha)$, then $y_t \in C$ is defined by $y_t(s) = y(t+s)$, $s \in [-h,\ 0]$. For $t_0 \ge 0$, $\phi \in C$, vector function $x \in C([-h,\ \alpha),\ R^n)$ with $\alpha > 0$ is called a solution of system (1.1) on $[0,\ \alpha)$ through (t_0, ϕ), denoted by $x(t, t_0, \phi)$ if $x_{t_0} = \phi$ and x satisfies (1.1) on $[0,\ \alpha)$.

DEFINITION 1.1. *The zero solution of system* (1.1) *is said to be stable if for any* $\epsilon > 0$ *there exists* $\delta(t_0, \epsilon) > 0$ *such that* $t \ge t_0 \ge 0$ *and* $\phi \in C$ *with* $\|\phi\| < \delta$ *implies* $|x(t, t_0, \phi)| < \epsilon$. *It is uniformly stable if the above* δ *is independent of* t_0.

DEFINITION 1.2. *The zero solution of system* (1.1) *is said to be globally asymptotically stable if it is stable and for any* $t_0 \ge 0$ *and for any* $\phi \in C$, $|x(t, t_0, \phi)| \to 0$ *as* $t \to \infty$.

Set $\eta_i = \sup_{t\ge 0}\int_t^{t+\tau_i} b_i(s)ds$. *Associated with matrix* $D = (d_{ij})_{n\times n}$, *we define the new matrix* $\tilde{D} = (\tilde{d}_{ij})_{n\times n}$ *by* $\tilde{d}_{ij} = d_{ij}/c_i$ *for* $i,\ j = 1, 2, \dots, n$, *where*

$$c_i = \begin{cases} \dfrac{2 - \eta_i^2 - 4p_i}{2 + \eta_i^2} & \text{if } \ \eta_i < 1\,, \\ \dfrac{3/2 - 2p_i - \eta_i}{1/2 + \eta_i} & \text{if } \ \eta_i \ge 1\,. \end{cases}$$

LEMMA 1.1. *If*

$$p_i < \frac{1}{4},\ \eta_i < \frac{3}{2} - 2p_i,\ i = 1, 2, \dots, n, \tag{1.14}$$

or

$$\frac{1}{4} \le p_i < \frac{1}{2},\ \eta_i < \sqrt{2(1-2p_i)},\ i = 1, 2, \dots, n, \tag{1.15}$$

then $0 < c_i < 1 - 2p_i$ *for* $i = 1, 2, \dots, n$.

Proof. For each $i \in \{1, 2, \dots, n\}$, we have $3/2 - 2p_i \ge \sqrt{2(1-2p_i)}$. So if condition (1.15) holds, then we have $c_i > 0$. Now we consider the case that condition (1.14) holds. If $\eta_i \ge 1$, it is obvious that $c_i > 0$. If $\eta_i < 1$, since $p_i < \frac{1}{4}$, so $2 - \eta_i^2 - 4p_i > 1 - \eta_i^2 > 0$. Thus, $c_i > 0$. The proof for $c_i < 1 - 2p_i$ is direct; we omit it. □

Now we can state our main result.

THEOREM 1.4. *Assume that* (1.2) *is satisfied, and* (1.14) *or* (1.15) *holds. If* $I - \tilde{D}$ *is a nonsingular* M*-matrix,* I *is an* $n \times n$ *identity matrix, and*

$$\int_0^{+\infty} b_i(t)dt = +\infty,\ i = 1, 2, \dots, n, \tag{1.16}$$

then the zero solution of (1.1) *is globally asymptotically stable.*

Applying Theorem 1.3 to system (1.12) yields the following result.

COROLLARY 1.1. *Consider system* (1.12). *Let* $\bar{A} = (\bar{a}_{ij})$ *be defined by* $\bar{a}_{ij} = \mu_i|a_{ij}|$, $i, j = 1, 2, \dots, n$, *where*

$$\mu_i = \begin{cases} \dfrac{2 + (b_i\tau_i)^2}{2 - (b_i\tau_i)^2} & \text{if } \ b_i\tau_i < 1, \\ \dfrac{1 + 2b_i\tau_i}{3 - 2b_i\tau_i} & \text{if } \ b_i\tau_i \ge 1. \end{cases}$$

If $b_i\tau_i < 3/2,\ i = 1, \dots, n,$ *and*

$$\operatorname{diag}\{b_1,\ b_2, \dots,\ b_n\} - \bar{A} \tag{1.17}$$

is a nonsingular M-matrix, then the trivial solution of system (1.12) *is globally asymptotically stable.*

Remark 1.1. It is easy to check that $\hat{a}_{ij} \geq \bar{a}_{ij}$ for $i, j = 1, 2, \dots, n$. So the condition (1.13) in Theorem 1.3 is improved by the condition (1.17) in Corollary 1.1. Moreover, the conditions in Corollary 1.1 can guarantee the trivial solution of (1.5) to be globally asymptotically stable, which cannot be achieved by using the method proposed by So, Tang, and Zou in [11]. Therefore, the results obtained in this paper extend and improve the ones in [11].

Remark 1.2. Corollary 1.1 is not an extension of Theorem 1.1, since the condition (1.17) in Corollary 1.1 is stronger than the condition (1.6) in Theorem 1.1. This suggests that there is room for improving condition (1.17).

Applying Theorem 1.4 to (1.7), we get the following corollary.

Corollary 1.2. *Assume that there exists a constant* $p \geq 0$ *such that*

$$|p(t)| \leq p. \tag{1.18}$$

Set

$$\eta = \sup_{t\geq 0} \int_t^{t+\tau} b(s)ds, \quad c = \begin{cases} \dfrac{2-\eta^2-4p}{2+\eta^2} & \text{if } \ \eta < 1\,, \\ \dfrac{3/2-2p-\eta}{1/2+\eta} & \text{if } \ \eta \geq 1\,. \end{cases}$$

If condition (1.10) *or* (1.11) *is satisfied,* $\int_0^{+\infty} b(t)dt = +\infty$*, and there exists a constant* $\delta,\ 0 \leq \delta < c$*, such that* $|a(t)| \leq \delta b(t)$*, then the zero solution of* (1.7) *is globally asymptotically stable.*

Remark 1.3. Comparing Corollary 1.2 with Theorem 1.2 obtained in [13], condition (1.9) in Theorem 1.2 is improved by (1.18). Moreover, our result is independent of the delay r. When $\delta = 0$, we will get the 3/2 asymptotic stability result for neutral differential delay equations. Therefore, Corollary 1.2 extends the 3/2 asymptotic stability results obtained in [6], [7], [8], [9], [10] to the perturbed equations. However, as pointed out in Corollary 1.1, the upper bound c of δ in Corollary 1.2 may not be the best possible. A special example is given in section 3 which shows that the best possible upper bound of δ may be 1.

2. Proof of Theorem 1.4. First we introduce some notation. Since $I - \tilde{D}$ is a nonsingular M-matrix, there exist positive constants $r_1,\ r_2, \dots,\ r_n$, such that

$$-r_i + \sum_{j=1}^n r_j \tilde{d}_{ij} < 0.$$

By the definition of $\tilde{d}_{ij}$, we get

$$-r_i + 1/c_i \sum_{j=1}^n r_j d_{ij} < 0. \tag{2.1}$$

In what follows, we set $\bar{d} = \max_{1\leq i\leq n}\{1 + \sum_{j=1}^n d_{ij} r_j / r_i\}$, $q = \min_{1\leq i\leq n}\{q_i\}$, $\bar{r} = \max_{1\leq i\leq n}\{r_i\}$, $\underline{r} = \min_{1\leq i\leq n}\{r_i\}$.

LEMMA 2.1. *Consider* (1.1) *and assume that condition* (1.2) *is satisfied for* $\eta_i < +\infty$ *for* $i \in \{1, 2, \dots, n\}$. *Set* $\eta = \sum_{j=1}^{n} \eta_j$. *For* $\phi \in C$, *let* $x(t) = x(t, t_0, \phi)$, *and* $v(t) = \max_{1\le i\le n} \{\|x_{it}\|/r_i\}$ *for* $t \ge t_0$. *Then*

$$v(t) \le \left[\max_{1\le i\le n} \{1 + 2p_i\}\right]^m \exp\left(m\bar{d}\eta\right) v(t_0), \quad t \in [t_0+(m-1)q,\ t_0+mq], \quad m = 1, 2, \dots.$$

Proof. For $t \in [t_0,\ t_0 + q]$, integrating (1.1) on $[t_0,\ t]$, we get

$$\begin{aligned} x_i(t) = x_i(t, t_0, \phi) &= x_i(t_0) + p_i(t)x_i(t - q_i) - p_i(t_0)x_i(t_0 - q_i) \\ &\quad + \int_{t_0}^{t} (-b_i(s)x_i(s - \tau_i(s)) + f_i(s, x_{i1}(s - \tau_{i1}(s)), \dots, x_{in}(s - \tau_{in}(s))))\, ds \end{aligned}$$

for $i = 1, \dots, n$. So, by (1.2), for $i = 1, \dots, n$, $t \in [t_0,\ t_0 + q]$,

$$\begin{aligned} |x_i(t)| &\le (1 + 2p_i)\|x_{it_0}\| + \int_{t_0}^{t} b_i(s) \left[\|x_{is}\| + \sum_{j=1}^{n} d_{ij}\|x_{js}\|\right] ds \\ &= (1 + 2p_i)r_i\|x_{it_0}\|/r_i + \int_{t_0}^{t} b_i(s) \left[r_i\left(\|x_{is}\|/r_i\right) + \sum_{j=1}^{n} d_{ij}r_j\left(\|x_{js}\|/r_j\right)\right] ds \\ &\le (1 + 2p_i)r_i v(t_0) + \int_{t_0}^{t} b_i(s) \left[r_i + \sum_{j=1}^{n} d_{ij}r_j\right] v(s)ds. \end{aligned}$$

It follows that for $i = 1, \dots, n$, $t \in [t_0,\ t_0 + q]$,

$$\|x_{it}\|/r_i \le (1 + 2p_i)v(t_0) + \int_{t_0}^{t} b_i(s) \left[1 + \sum_{j=1}^{n} d_{ij}r_j/r_i\right] v(s)ds.$$

Hence, for $t \in [t_0,\ t_0 + q]$,

$$v(t) \le \max_{1\le i\le n} \{1 + 2p_i\}\, v(t_0) + \int_{t_0}^{t} \sum_{j=1}^{n} b_j(s)\bar{d}v(s)ds.$$

By Gronwall's inequality, for $t \in [t_0,\ t_0 + q]$,

$$v(t) \le \max_{1\le i\le n} \{1 + 2p_i\} \exp\left(\bar{d}\sum_{j=1}^{n}\int_{t_0}^{t} b_j(s)ds\right) v(t_0) \le \max_{1\le i\le n} \{1 + 2p_i\} \exp\left(\bar{d}\eta\right) v(t_0).$$

Then, using the induction method, for any $m \ge 1$, we get

$$v(t) \le \left[\max_{1\le i\le n} \{1 + 2p_i\}\right]^m \exp\left(m\bar{d}\eta\right) v(t_0), \quad t \in [t_0 + (m - 1)q,\ t_0 + mq]. \qquad \square$$

LEMMA 2.2. *Under the conditions of Theorem* 1.4, *the zero solution of* (1.1) *is uniformly stable.*

Proof. Let $y_i(t) = x_i(t) - p_i(t)x_i(t - q_i)$, $z(t) = \max_{1\le i\le n}\{|x_i(t)|/r_i\}$. Choose positive integer l such that $lq \ge 2h$. For given $\epsilon > 0$, let

$$\delta = \frac{\underline{r}}{\bar{r}}\frac{1-\bar{p}}{1+\bar{p}}\left[\max_{1\le i\le n}\{1+2p_i\}\right]^{-l}\exp\left(-\bar{d}l\eta\right)\epsilon.$$

From Lemma 2.1, for $\|\phi\| < \delta$,

$$\begin{aligned} z(t) &\le v(t) \le \left[\max_{1\le i\le n}\{1+2p_i\}\right]^{l}\exp\left(\bar{d}l\eta\right)v(t_0) \\ &\le \left[\max_{1\le i\le n}\{1+2p_i\}\right]^{l}\exp\left(\bar{d}l\eta\right)\|\phi\|/\underline{r} \\ &< \frac{1-\bar{p}}{1+\bar{p}}\bar{\epsilon},\quad t\in[t_0,\ t_0+lq], \end{aligned} \tag{2.2}$$

where $\bar{\epsilon} = \epsilon/\bar{r}$.

Now we use the techniques in [12] and [13] to prove that

$$|z(t)| < \bar{\epsilon}\ \text{ for }\ t > t_0 + lq. \tag{2.3}$$

If this is true, then by (2.2), we get

$$|x(t)| \le \bar{r}z(t) < \epsilon\ \text{ for } t > t_0,$$

which implies that the zero solution of (1.1) is uniformly stable. Suppose that (2.3) is not true; then there exists $t^* > t_0 + lq$ such that

$$z(t) < \bar{\epsilon} \text{ for } t\in[t_0,\ t^*] \text{ and }\ z(t^*) = \bar{\epsilon}.$$

Then there exists some $k \in \{1, 2, \ldots, n\}$ such that $z(t^*) = |x_k(t^*)|/r_k$. Without loss of generality, we may assume that $x_k(t^*) > 0$; then $x_k(t^*) = r_k\bar{\epsilon}$, and

$$|x_i(t)| \le r_i z(t) < r_i\bar{\epsilon} \text{ for } t\in[t_0,\ t^*],\ i = 1, 2, \ldots, n. \tag{2.4}$$

It follows that

$$y_k(t^*) = x_k(t^*) - p_k(t^*)x_k(t^* - q_k) > (1 - p_k)r_k\bar{\epsilon}. \tag{2.5}$$

On the other hand, by (2.2) and (2.4), for $t \in [t_0,\ t_0 + lq]$,

$$\begin{aligned} |y_k(t)| &= |x_k(t) - p_k(t)x_k(t - q_k)| \\ &< (1 + p_k)r_k\bar{\epsilon}(1 - \bar{p})/(1 + \bar{p}) \le (1 - p_k)r_k\bar{\epsilon}. \end{aligned} \tag{2.6}$$

Combining (2.5) and (2.6) yields that there exists $\xi \in (t_0 + lq,\ t^*]$ such that

$$\dot{y}_k(\xi) \ge 0,\quad \text{and}\quad y_k(\xi) = \max_{s\in[t_0,\ t^*]} y_k(s) > (1 - p_k)r_k\bar{\epsilon}. \tag{2.7}$$

Set $u(t) = y_k(t) - p_k r_k\bar{\epsilon}$; then for $t \in [t_0, t^*]$,

$$r_k\bar{\epsilon} \ge x_k(t) = y_k(t) + p_k(t)x_k(t - q_k) \ge y_k(t) - p_k r_k\bar{\epsilon} = u(t). \tag{2.8}$$

So,

$$\begin{aligned} \dot{u}(t) &= \dot{y}_k(t) < b_k(t)(-x_k(t - \tau_k(t)) + c_k r_k\bar{\epsilon}) \\ &\le b_k(t)(-u(t - \tau_k(t)) + c_k r_k\bar{\epsilon}). \end{aligned} \tag{2.9}$$

In particular,

$$\dot{u}(\xi) < b_k(\xi)(-u(\xi - \tau_k(\xi)) + c_k r_k \bar{\epsilon}). \tag{2.10}$$

If we can prove that

$$u(t) > c_k r_k \bar{\epsilon} \text{ for } t \in [\xi - \tau_k,\ \xi], \tag{2.11}$$

then by (2.10), we will get $\dot{u}(\xi) < 0$, which contradicts $\dot{u}(\xi) = \dot{y}_k(\xi) \geq 0$, so the proof will be complete. In what follows, we will prove (2.11). Suppose the contrary. Since $u(\xi) = y_k(\xi) - p_k r_k \bar{\epsilon} > (1 - 2p_k) r_k \bar{\epsilon} > c_k r_k \bar{\epsilon}$ by Lemma 1.1 and (2.7), there exists $t_1 \in [\xi - \tau_k,\ \xi)$ such that

$$u(t_1) = c_k r_k \bar{\epsilon} \text{ and } \ u(t) > c_k r_k \bar{\epsilon} \text{ for } t \in (t_1,\ \xi]. \tag{2.12}$$

We assert that

$$\dot{u}(t) < (1 + c_k) b_k(t) \int_{t-\tau_k}^{t_1} b_k(s) ds r_k \bar{\epsilon},\ \ t \in [t_1 - \tau_k, \xi]. \tag{2.13}$$

In fact, if $t - \tau_k(t) > t_1$ for some $t \in [t_1 - \tau_k, \xi]$, then $\dot{u}(t) < 0$ by (2.9) and (2.12), which yields (2.13). On the other hand, by (2.8) and (2.9), we have

$$\dot{u}(t) < (1 + c_k) b_k(t) r_k \bar{\epsilon},\ \ t \in [t_1 - \tau_k, \xi]. \tag{2.14}$$

If $t - \tau_k(t) < t_1$ for some $t \in [t_1 - \tau_k, \xi]$, integrating the above inequality from $t - \tau_k(t)$ to t_1, we have

$$c_k r_k \bar{\epsilon} - u(t - \tau_k(t)) = u(t_1) - u(t - \tau_k(t)) = \int_{t-\tau_k(t)}^{t_1} \dot{u}(s) ds < (1 + c_k) \int_{t-\tau_k}^{t_1} b_k(s) ds r_k \bar{\epsilon}.$$

Substituting the above inequality into (2.9), we get (2.13).

Combining (2.13) and (2.14), we have

$$\begin{aligned}
\dot{u}(t) &< (1 + c_k) b_k(t) \min \left\{ 1,\ \int_{t-\tau_k}^{t_1} b_k(s) ds \right\} r_k \bar{\epsilon} \\
&= (1 + c_k) b_k(t) \min \left\{ 1,\ \int_{t-\tau_k}^{t} b_k(s) ds - \int_{t_1}^{t} b_k(s) ds \right\} r_k \bar{\epsilon} \\
&\leq (1 + c_k) b_k(t) \min \left\{ 1,\ \eta_k - \int_{t_1}^{t} b_k(s) ds \right\} r_k \bar{\epsilon}.
\end{aligned}$$

For $t \in [t_1,\ \xi]$, integrating the above inequality on the two side from t_1 to ξ, we get

$$(1 - 2p_k - c_k) r_k \bar{\epsilon} < u(\xi) - u(t_1) < (1 + c_k) I_k r_k \bar{\epsilon}. \tag{2.15}$$

Let $I_k = \int_{t_1}^{\xi} b_k(t) \min\{1,\ \eta_k - \int_{t_1}^{t} b_k(s) ds\} dt$. Now we give an estimation of I_k. There are two possible cases.

Case 1. $\int_{t_1}^{\xi} b_k(t) dt \geq 1$. In this case, choose $t_2 \in (t_1,\ \xi)$ such that $\int_{t_2}^{\xi} b_k(t) dt = 1$.

$$\begin{aligned}
I_k &\leq \int_{t_1}^{t_2} b_k(t) dt + \int_{t_2}^{\xi} b_k(t) \left[\eta_k - \int_{t_1}^{t} b_k(s) ds \right] dt \\
&\leq \int_{t_1}^{t_2} b_k(t) dt + \eta_k \int_{t_2}^{\xi} b_k(t) dt - \frac{1}{2} \int_{t_2}^{\xi} b_k(t) dt \left[\int_{t_2}^{\xi} b_k(t) dt \right. \\
&\qquad \left. + 2 \int_{t_1}^{t_2} b_k(t) dt \right] \\
&= \eta_k - \frac{1}{2}.
\end{aligned} \tag{2.16}$$

Case 2. $\int_{t_1}^{\xi} b_k(t)dt < 1$.

$$\begin{aligned} I_k &\le \eta_k \int_{t_1}^{\xi} b_k(t)dt - \int_{t_1}^{\xi} b_k(t) \int_{t_1}^{t} b_k(s)dsdt \\ &= \eta_k \int_{t_1}^{\xi} b_k(t)dt - \frac{1}{2}\left[\int_{t_1}^{\xi} b_k(t)dt\right]^2 . \end{aligned}$$

Note that $\int_{t_1}^{\xi} b_k(t)dt \le \eta_k$ and the function $-1/2\,x^2 + \eta_k x$ is increasing on $x \in [0, \eta_k]$. Therefore, if $\eta_k \ge 1$, then $I_k \le \eta_k - \frac{1}{2}$. If $\eta_k < 1$, then $I_k \le \frac{1}{2}{\eta_k}^2$.

Since $\frac{1}{2}{\eta_k}^2 \ge \eta_k - \frac{1}{2}$, combining Cases 1 and 2 yields

$$I_k \le \begin{cases} \frac{1}{2}\eta_k^2 & \text{if } \eta_k < 1\,, \\ \eta_k - \frac{1}{2} & \text{if } \eta_k \ge 1\,. \end{cases}$$

Substituting the above inequality into (2.15), we get

$$c_k > \begin{cases} \dfrac{2 - 4p_k - \eta_k^2}{2 + \eta_k^2} & \text{if } \eta_k < 1, \\ \dfrac{3/2 - 2p_k - \eta_k}{\eta_k + 1/2} & \text{if } \eta_k \ge 1, \end{cases}$$

which contradicts the definition of $c_i,\ i = 1, 2, \dots, n$. The proof is complete. $\square$

Proof of Theorem 1.4. In view of Lemma 2.2, the zero solution is uniformly stable. Moreover, from the proof of Lemma 2.2, every solution $x(t)$ of (1.1) is bounded. Now we prove that

$$\lim_{t\to\infty} x(t) = 0. \tag{2.17}$$

Set $\gamma = \max_{1\le i\le n} t\{\sum_{j=1}^{n} d_{ij} r_j/(c_i r_i)\}$. Then by (2.1), $0 < \gamma < 1$. Set $\beta_i = \limsup_{t\to\infty}|x_i(t)|$, then $0 \le \beta_i < \infty$. We assume that $\beta_k/r_k = \max_{1\le i\le n}\{\beta_i/r_i\}$ for some $k \in \{1, 2, \dots, n\}$. If $\beta_k = 0$, then (2.17) holds. If $\beta_k > 0$, choose

$$\bar\sigma_k = \begin{cases} (1-\gamma)\frac{1-2p_k-\frac{1}{2}\eta_k^2}{1+\gamma\left(1-2p_k-\frac{1}{2}\eta_k^2\right)\bar r/\underline{r}+\frac{1}{2}\eta_k^2\bar r/\underline{r}}\beta_k & \text{if } \ \eta_k < 1, \\ (1-\gamma)\frac{3/2-2p_k-\eta_k}{1+\gamma(3/2-2p_k-\eta_k)\bar r/\underline{r}+(\eta_k-1/2)\bar r/\underline{r}}\beta_k & \text{if } \ \eta_k \ge 1. \end{cases}$$

It is easy to see that $0 < \bar\sigma_k < \beta_k$. Moreover, for any $0 < \sigma_k < \min\{\bar\sigma_k,\ (1 - 2p_k - \gamma c_k)\beta_k/\left(1 + \gamma c_k \bar r/\underline{r}\right)\}$, there exists $T > t_0 + 2h$, such that

$$|x_i(t)| < \beta_i + \sigma_k \text{ for } t \ge T, \quad i = 1, \dots, n.$$

Let $y_k(t) = x_k(t) - p_k(t)x_k(t - q_k)$. We distinguish the two cases.

Case A. *$\dot y_k(t)$ is eventually sign-definite.* By the boundedness of $y_k(t)$, the limit $\mu_k = \lim_{t\to+\infty} y_k(t)$ exists. We may assume that $\limsup_{t\to+\infty} x_k(t) = \beta_k$. Then,

$$\beta_k = \limsup_{t\to+\infty} x_k(t) \le \limsup_{t\to+\infty} y_k(t) + p_k \limsup_{t\to+\infty} |x_k(t - q_k)| = \mu_k + p_k\beta_k.$$

So $\mu_k \ge (1 - p_k)\,\beta_k$. Set $\alpha_k = \liminf_{t\to+\infty} x_k(t)$. Then

$$\alpha_k = \liminf_{t\to+\infty} x_k(t) \ge \liminf_{t\to+\infty} y_k(t) - p_k \limsup_{t\to+\infty} |x_k(t - q_k)| = \mu_k - p_k\beta_k \ge (1 - 2p_k)\,\beta_k.$$

Thus, there exists $T_1 \geq T$ such that

$$(1-2p_k)\,\beta_k - \sigma_k < x_k(t) < \beta_k + \sigma_k,\ \ t \geq T_1 + h.$$

By (1.1) and (2.1), for $t \geq T_1 + h$,

$$\begin{aligned}
\dot{y}_k(t) &\leq -b_k(t)x_k(t-\tau_k(t)) + b_k(t)\sum_{j=1}^{n} d_{kj}|x_j(t-\tau_{kj}(t))| \\
&< -b_k(t)\left((1-2p_k)\,\beta_k - \sigma_k\right) + b_k(t)\sum_{j=1}^{n} d_{kj}\left(\beta_j + \sigma_k\right) \\
&= -b_k(t)\left((1-2p_k)\,\beta_k - \sigma_k\right) + b_k(t)\sum_{j=1}^{n} d_{kj}r_j\left(\beta_j/r_j + \sigma_k/r_j\right) \\
&\leq -b_k(t)\left((1-2p_k)\,\beta_k - \sigma_k\right) + b_k(t)\sum_{j=1}^{n} d_{kj}r_j\left(\beta_k/r_k + \sigma_k/\underline{r}\right) \\
&\leq -b_k(t)\left((1-2p_k)\,\beta_k - \sigma_k\right) + \gamma b_k(t)c_k r_k\left(\beta_k/r_k + \sigma_k/\underline{r}\right) \\
&= -b_k(t)\left\{(1-2p_k-\gamma c_k)\,\beta_k - \sigma_k\left[1+\gamma c_k\bar{r}/\underline{r}\right]\right\}.
\end{aligned}$$

By using (1.16) and the definition of σ_k, the above inequality implies that $y_k(t) \to -\infty$. This is a contradiction. So, $\beta_k = 0$.

Case B. *$\dot{y}_k(t)$ is oscillatory.* Noticing that

$$\limsup_{t\to\infty}|y_k(t)| \geq \limsup_{t\to\infty}|x_k(t)| - p_k\limsup_{t\to\infty}|x_k(t-q_k)| \geq (1-p_k)\beta_k,$$

there exists an increasing sequence $\{t_{km}\}$ such that $t_{km} \to \infty$ as $m \to \infty$, $\dot{y}_k(t_{km}) \geq 0$, and

$$|y_k(t_{km})| > (1-p_k)(\beta_k - \sigma_k).$$

Without loss of generality, we assume $t_{km} > T + h$ and $y_k(t_{km}) > 0$. Set $u(t) = y_k(t) - p_k(\beta_k + \sigma_k)$; then $\dot{u}(t_{km}) = \dot{y}_k(t_{km}) \geq 0$ and

$$\begin{aligned}
\beta_k + \sigma_k \ &\geq\ x_k(t) = y_k(t) + p_k(t)x_k(t-q_k) \\
&>\ y_k(t) - p_k(\beta_k+\sigma_k) = u(t),\ \ t \geq T+h.
\end{aligned} \tag{2.18}$$

Moreover, By (1.1), (2.1), and the above inequality,

$$\begin{aligned}
\dot{u}(t) \ &=\ \dot{y}_k(t) < -b_k(t)x_k(t-\tau_k(t)) + \gamma b_k(t)c_k\left(\beta_k + \sigma_k\bar{r}/\underline{r}\right) \\
&<\ -b_k(t)u(t-\tau_k(t)) + \gamma b_k(t)c_k\left(\beta_k + \sigma_k\bar{r}/\underline{r}\right).
\end{aligned} \tag{2.19}$$

In particular,

$$\begin{aligned}
0 \leq \dot{u}(t_{km}) = \dot{y}_k(t_{km}) \ &<\ -b_k(t_{km})u(t_{km} - \tau_k(t_{km})) \\
&\quad +\gamma b_k(t_{km})c_k\left(\beta_k + \ \sigma_k\bar{r}/\underline{r}\right).
\end{aligned} \tag{2.20}$$

If we can prove that

$$u(t) > \gamma c_k(\beta_k + \sigma_k\bar{r}/\underline{r}) \text{ for } t \in [t_{km} - \tau_k,\ t_{km}], \tag{2.21}$$

then by (2.20), we will obtain a contradiction, so we can deduce $\beta_k = 0$ and the proof is complete. Now under the assumption of $\beta_k > 0$, we prove that (2.21) holds. Note that by the definition of σ_k,

$$u(t_{km}) = y_k(t_{km}) + p_k(t_{km})x_k(t_{km} - q_k) > (1 - 2p_k)\beta_k - \sigma_k > \gamma c_k(\beta_k + \sigma_k\bar{r}/\underline{r}).$$

So, if (2.21) is not true, then there exists $t^1_{km} \in [t_{km} - \tau_k,\ t_{km})$, such that

$$u\left(t^1_{km}\right) = \gamma c_k(\beta_k + \sigma_k\bar{r}/\underline{r}) \text{ and } u(t) > \gamma c_k\left(\beta_k + \sigma_k\bar{r}/\underline{r}\right) \tag{2.22}$$

for $t \in \left(t^1_{km},\ t_{km}\right]$. We assert that

$$\dot{u}(t) < (1 + \gamma c_k)b_k(t)\int_{t-\tau_k}^{t^1_{km}} b_k(s)ds(\beta_k + \sigma_k\bar{r}/\underline{r}),\ t \in [t^1_{km} - \tau_k, t_{km}]. \tag{2.23}$$

In fact, by (2.18) and (2.19), we have

$$\dot{u}(t) < b_k(t)(1 + \gamma c_k)(\beta_k + \sigma_k\bar{r}/\underline{r}),\ t \in [t^1_{km} - \tau_k, t_{km}]. \tag{2.24}$$

If $t-\tau_k(t) \le t^1_{km}$ for some $t \in [t^1_{km}, t_{km}]$, integrating the above inequality from $t-\tau_k(t)$ to t^1_{km}, we get

$$\begin{aligned}\gamma c_k(\beta_k + \sigma_k\bar{r}/\underline{r}) - u(t - \tau_k(t)) &= u(t^1_{km}) - u(t - \tau_k(t))\\ &< (1 + \gamma c_k)\int_{t-\tau_k}^{t^1_{km}} b_k(s)ds(\beta_k + \sigma_k\bar{r}/\underline{r}).\end{aligned}$$

Substituting the above inequality into the second inequality in (2.19), we get (2.23). If $t - \tau_k(t) > t^1_{km}$ for some $t \in [t^1_{km}, t_{km}]$, then by (2.23) and the second inequality of (2.19), we get (2.23).

Combining (2.23) and (2.24) yields

$$\dot{u}(t) < (1 + \gamma c_k)b_k(t)\min\left\{1,\ \int_{t-\tau_k}^{t^1_{km}} b_k(s)ds\right\}\left(\beta_k + \sigma_k\bar{r}/\underline{r}\right).$$

Integrating this inequality from t^1_{km} to t_{km}, we get

$$\begin{aligned}&(1 - \gamma c_k - 2p_k)\beta_k - \sigma_k\left(1 + \gamma c_k\right)\bar{r}/\underline{r} < u(t_{km}) - u\left(t^1_{km}\right)\\ <\ &(1 + \gamma c_k)I_{km}\left(\beta_k + \sigma_k\bar{r}/\underline{r}\right),\end{aligned} \tag{2.25}$$

where $I_{km} = \int_{t^1_{km}}^{t_{km}} b_k(t)\min\{1,\ \int_{t-\tau_k}^{t^1_{km}} b_k(s)ds\}dt$.

Using the same technique as in the proof of Lemma 2.2, we have

$$I_{km} \le \begin{cases} 1/2\eta_k^2 & \text{if } \eta_k < 1\,,\\ \eta_k - 1/2 & \text{if } \eta_k \ge 1\,.\end{cases}$$

Therefore, by (2.25), we have

$$\sigma_k > \bar{\sigma}_k\,,$$

which is impossible since $\sigma_k < \bar{\sigma}_k$. Therefore, $\beta_k = 0$; that is, (2.17) holds, and so the proof is complete. $\square$

3. A special example.

Example 3.1. Consider the delay differential equation

$$\dot{x}(t) = -b(t)(x(t-1) - \delta x(t-1/2)), \tag{3.1}$$

where $\delta > 0$ and $b(t)$ is defined by

$$b(t) = \begin{cases} 48(t-3i)(3i+1/2-t) & \text{for } 3i \le t \le 3i+1/2, \\ \beta(t-3i-1/2)(3i+1-t) & \text{for } 3i+1/2 \le t \le 3i+1, \\ 48(t-3i-1)(3i+3/2-t) & \text{for } 3i+1 \le t \le 3i+3/2, \\ 0 & \text{for } 3i+3/2 \le t \le 3(i+1), \end{cases}$$

where $\beta > 0$ is a constant.

The example is a perturbed version of Example 3.1 in [18]. It is easy to see that $\eta = \sup_{t\ge 0} \int_t^{t+1} b(s)ds = 1 + \beta/48$. By Corollary 1.2, if $\beta < 24$ and $\delta < c$, then the zero solution of (3.1) is asymptotically stable.

But by elementary calculations, we have, for $i = 1, 2, \ldots,$

$$x(t) = x\left(3i - 3/2\right)\left\{1 - 48(1-\delta)\left[\frac{1}{4}(t-3i)^2 - \frac{1}{3}(t-3i)^3\right]\right\} \text{ for } t \in [3i, 3i+1/2],$$

$$\begin{aligned} x(t) = x\left(3i - 3/2\right)\Bigg\{&\delta - \frac{\beta}{4}(1-\delta)\left(t - 3i - \frac{1}{2}\right)^2 + \frac{\beta}{3}(1-\delta)\left(t-3i-\frac{1}{2}\right)^3 \\ &-\frac{3}{2}\beta\delta(1-\delta)\left(t-3i-\frac{1}{2}\right)^4 + 4\beta\delta(1-\delta)\left(t-3i-\frac{1}{2}\right)^5 \\ &-\frac{8}{3}\beta\delta(1-\delta)\left(t-3i-\frac{1}{2}\right)^6\Bigg\} \text{ for } t \in [3i+1/2, 3i+1], \end{aligned}$$

$$\begin{aligned} x(t) = x\left(3i - 3/2\right)\Bigg\{&\delta - \frac{\beta}{96}(\delta+2)(1-\delta) - (1-\delta)(t-3i-1)^2[12 - 16(t-3i-1)] \\ &+(1-\delta)(48-\beta\delta)(t-3i-1)^4\left[\frac{3}{2} - 4(t-3i-1) + \frac{8}{3}(t-3i-1)^2\right] \\ &-48\beta\delta^2(1-\delta)(t-3i-1)^6\left[\frac{1}{8} - \frac{1}{2}(t-3i-1) + \frac{8}{3}(t-3i-1)^2\right]\Bigg\} \\ &\text{for } t \in [3i+1, 3i+3/2], \end{aligned}$$

$$x(t) = x\left(3i-3/2\right)g(\delta) \text{ for } t \in [3i+3/2, 3i+3],$$

where $g(\delta) = -1 + 2\delta - \frac{\beta}{96}(\delta+2)(1-\delta) + \frac{1}{96}(1-\delta)(48-\beta\delta) - \frac{\beta}{288}\delta^2(1-\delta)$. By induction, we have $x(3i-3/2) = x(3/2)g^{i-1}(\delta)$ for $i \ge 1$. So there exists a constant $M_0 > 0$ such that $|x(t)| \le M_0|x(3/2)||g(\delta)|^{i-1}$ for $t \in [3i, 3i+3]$ and $i \ge 1$. Note that $\dot{g}(\lambda) = \frac{\beta}{96}\delta^2 + \frac{5\beta}{144}\delta + \frac{3}{2} > 0$ for $\delta \in [0,1)$, so $-\frac{\beta}{48} - \frac{1}{2} \le g(\delta) < 1$ for $\delta \in [0,1)$. Therefore, the necessary and sufficient conditions for asymptotic stability of the zero solution of (3.1) are $\beta < 24$ and $\delta < 1$. This means that condition $\delta < c$ in Corollary 1.2 is a sufficient condition (not a necessary condition) for the zero solution of (1.7) to be asymptotically stable.

Remark 3.1. When $\delta = 0$, this example (see [18]) also shows that the upper bound 3/2 for the stability result in [7] is the best possible when the coefficient is a continuous function (the case when the coefficient is a piecewise continuous function was proved in [7]). However, how to find the sharp condition for the stability of the perturbed system remains an open question.

Acknowledgments. The authors acknowledge helpful comments and suggestions of the anonymous reviewers.

REFERENCES

[1] J. K. Hale, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[2] G. Stépán, *Retarded Dynamical Systems: Stability and Characteristic Functions*, Pitman Res. Notes Math. Ser. 210, Longman Scientific & Technical, Harlow, UK, 1989.

[3] K. Gopalsamy and X. He, *Global stability in n-species competition modelled by "pure-delay type" systems* II: *Nonautonomous case*, Canad. Appl. Math. Quart., 6 (1998), pp. 17–43.

[4] X. He, *Global stability in nonautonomous Lotka-Volterra systems of "pure-delay type"*, Differential Integral Equations, 11 (1998), pp. 293–310.

[5] X. He, *The Lyapunov functionals for delay Lotka–Volterra-type models*, SIAM J. Appl. Math., 58 (1998), pp. 1222–1236.

[6] J. A. York, *Asymptotic stability for one dimensional differential delay equations*, J. Differential Equations, 7 (1970), pp. 189–202.

[7] T. Yoneyama, *On the* 3/2 *stability theorem for one dimensional delay differential equations*, J. Math. Anal. Appl., 125 (1987), pp. 161–173.

[8] T. Krisztin, *On stability properties for one dimensional functional differential equations*, Funckial. Ekvac., 34 (1991), pp. 241–256.

[9] Y. Muroya, *Note on Yoneyama's* 3/2 *stability theorems for one dimensional delay differential equations,* J. Math. Anal. Appl., 247 (2000), pp. 314–322.

[10] J. S. Yu, *Asymptotic stability for nonautonomous scalar neutral differential equations*, J. Math. Anal. Appl., 203 (1996), pp. 850–860.

[11] J. W.-H. So, X. Tang, and X. Zou, *Stability in a linear delay system without instantaneous negative feedback,* SIAM J. Math. Anal., 33 (2002), pp. 1297–1304.

[12] W.-H. Chen and Z.-H. Guan, *Uniform asymptotic stability for perturbed neutral delay differential equations*, J. Math. Anal. Appl., 291 (2004), pp. 578–595.

[13] X. Wang and L. Liao, *Asymptotic behavior of solutions of neutral differential equations with positive and negative coefficients*, J. Math. Anal. Appl., 279 (2003), pp. 326–338.

[14] I. Györi, *Interaction between oscillations and global asymptotic stability in delay differential equations*, Differential Integral Equations, 3 (1990), pp. 181–200.

[15] L. Massoulié, *Stability of distributed congestion control with heterogeneous feedback delays*, IEEE Trans. Automat. Control, 47 (2002), pp. 895–902.

[16] K. Pyragas, *Continuous control of chaos by self-controlling feedback*, Phys. Lett. A, 170 (1992), pp. 421–428.

[17] W.-H. Chen, Z.-H. Guan, and X. Lu, *Delay-dependent exponential stability of neural networks with variable delays*, Phys. Lett. A, 326 (2004), pp. 355–363.

[18] J. S. Yu, *Sharp conditions for the uniform stability of nonautonomous Hutchinson's equation*, Ann. Differential Equations, 20 (2004), pp. 194–207.

[19] M. Bartha, *On stability properties for neutral differential equations with state-dependent delay*, Differential Equations Dynam. Systems, 7 (1999), pp. 197–220.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1535–1559

GLOBAL BRANCHES OF TRAVELLING-WAVES TO A GROSS–PITAEVSKII–SCHRÖDINGER SYSTEM IN ONE DIMENSION*

MIHAI MARIŞ†

Abstract. We are interested in the existence of travelling-wave solutions to a system which modelizes the motion of an uncharged impurity in a Bose condensate. We prove that in space dimension one, there exist travelling-waves moving with velocity c if and only if c is less than the sound velocity at infinity. In this case we investigate the structure of the set of travelling-waves and we show that it contains global subcontinua in appropriate Sobolev spaces.

Key words. Gross–Pitaevskii system, travelling-waves, global bifurcation

AMS subject classifications. 35Q55, 35Q51, 37K50, 35P05, 35J10

DOI. 10.1137/040611914

1. Introduction.

This paper is devoted to the study of a special kind of solutions of a system describing the motion of an uncharged impurity in a Bose condensate. In dimensionless variables, the system reads

$$\left\{\begin{aligned} 2i\frac{\partial\psi}{\partial t} &= -\Delta\psi + \tfrac{1}{\varepsilon^2}(|\psi|^2 + \tfrac{1}{\varepsilon^2}|\varphi|^2 - 1)\psi, \\ 2i\delta\frac{\partial\varphi}{\partial t} &= -\Delta\varphi + \tfrac{1}{\varepsilon^2}(q^2|\psi|^2 - \varepsilon^2 k^2)\varphi. \end{aligned}\right. \tag{1.1}$$

Here ψ and φ are the wavefunctions for bosons, respectively for the impurity, $\delta = \frac{\mu}{M}$, where μ is the mass of impurity, M is the boson mass (δ is supposed to be small), $q^2 = \frac{l}{2d}$, l is the boson-impurity scattering length and d the boson diameter, k is a dimensionless measure for the single-particle impurity energy, and ε is a dimensionless constant ($\varepsilon = (\frac{a\mu}{lM})^{\frac{1}{5}}$, where a is the "healing length"; in applications, $\varepsilon \cong 0.2$). Assuming that we are in a frame in which the condensate is at rest at infinity, the solutions must satisfy the "boundary conditions"

$$|\psi| \longrightarrow 1, \quad \varphi \longrightarrow 0 \quad \text{as } |x| \longrightarrow \infty. \tag{1.2}$$

This system (originally introduced by Clark and Gross) was studied by Grant and Roberts [5]. Using formal asymptotic expansions and numerical experiments, they computed the effective radius and the induced mass of the uncharged impurity.

We consider here the system (1.1) in a one dimensional space and we look for solitary waves, that is, for solutions of the form

$$\psi(x,t) = \tilde{\psi}(x - ct), \quad \varphi(x,t) = \tilde{\varphi}(x - ct). \tag{1.3}$$

This kind of solution corresponds to the case where the only disturbance present in the condensate is that caused by the uniform motion of the impurity with velocity c.

*Received by the editors July 19, 2004; accepted for publication (in revised form) May 9, 2005; published electronically January 27, 2006.
http://www.siam.org/journals/sima/37-5/61191.html

†Université de Franche-Comté, Département de Mathématiques UMR 6623, 16, Route de Gray, 25030 Besançon Cedex, France (mihai.maris@math.univ-fcomte.fr).

In view of the boundary conditions, we seek solutions of the form

$$\tilde{\psi}(x) = (1+\tilde{r}(x))e^{i\psi_0(x)}, \quad \tilde{\varphi}(x) = \tilde{u}(x)e^{i\varphi_0(x)}, \tag{1.4}$$

with $\tilde{r}(x) \longrightarrow 0$, $\tilde{u}(x) \longrightarrow 0$ as $|x| \longrightarrow \infty$. By an easy computation we find that the real functions ψ_0, φ_0, $\tilde{r}$, $\tilde{u}$ must satisfy

$$\psi_0' = c\left(1 - \frac{1}{(1+\tilde{r})^2}\right), \tag{1.5}$$

$$\varphi_0' = c\delta, \tag{1.6}$$

$$\tilde{r}'' = c^2\left(\frac{1}{(1+\tilde{r})^3} - (1+\tilde{r})\right) + \frac{1}{\varepsilon^2}\left((1+\tilde{r})^3 - (1+\tilde{r}) + \frac{1}{\varepsilon^2}(1+\tilde{r})\tilde{u}^2\right), \tag{1.7}$$

$$\tilde{u}'' = \left(\frac{q^2}{\varepsilon^2}(1+\tilde{r})^2 - c^2\delta^2 - k^2\right)\tilde{u}. \tag{1.8}$$

From (1.6) we see that necessarily $\varphi_0(x) = c\delta x + C$. Note that the system is invariant under the transform $(\psi,\varphi) \longmapsto (e^{i\alpha}\psi, e^{i\beta}\varphi)$, so the integration constants in (1.5) and (1.6) are not important. Thus all we have to do is to solve the system (1.7)–(1.8). Thereafter it will be easy to find the corresponding phases from (1.5)–(1.6), and (1.4) will give a solitary-wave solution of (1.1).

After the scale change $\tilde{u}(x) = \frac{1}{\varepsilon}u(\frac{x}{\varepsilon})$, $\tilde{r}(x) = r(\frac{x}{\varepsilon})$, we find that the functions r and u satisfy

$$r'' = (1+r)^3 - (1+r) - c^2\varepsilon^2\left(1 + r - \frac{1}{(1+r)^3}\right) + (1+r)u^2, \tag{1.9}$$

$$u'' = (q^2(1+r)^2 - \lambda)u, \tag{1.10}$$

where

$$\lambda = \varepsilon^2(c^2\delta^2 + k^2). \tag{1.11}$$

The equation $r'' = (1+r)^3 - (1+r) - \frac{v^2}{4}(1 + r - \frac{1}{(1+r)^3}) + (1+r)U$, where U is a positive Borel measure, was studied in [7]. In the case $U \equiv 0$, it has been shown that this equation has only the trivial solution $r \equiv 0$ if $|v| \geq \sqrt{2}$; for $0 < |v| < \sqrt{2}$, it also admits the solution

$$r_v(x) = -1 + \sqrt{\frac{v^2}{2} + \left(1 - \frac{v^2}{2}\right)\tanh^2\left(\frac{\sqrt{2-v^2}}{2}x\right)}. \tag{1.12}$$

Moreover, any other nontrivial solution is of the form $r_v(\cdot - x_0)$ for some $x_0 \in \mathbf{R}$. Equation (1.10) is linear in u; more precisely, u must be an eigenvector of the linear operator $-\frac{d^2}{dx^2} + q^2(1+r)^2$ corresponding to the eigenvalue $\lambda = \varepsilon^2(c^2\delta^2 + k^2)$.

It is now clear that except for translations, the only solutions of (1.9)–(1.10) of the form $(r,0)$ are $(0,0)$ and $(r_{2c\varepsilon}, 0)$ (the latter exists only for $|c\varepsilon| < \frac{1}{\sqrt{2}}$). We call these solutions the *trivial solutions* of (1.9)–(1.10). We will prove that there exist nontrivial solutions of (1.9)–(1.10) in a neighborhood of $(r_{2c\varepsilon}, 0)$ (for suitable values of the parameter λ) and we will study the global structure of the set of nontrivial solutions.

It has been shown (see, e.g., [7] and references therein) that by using the Madelung transform $\psi = \sqrt{\rho}e^{i\psi_0}$, the first equation in (1.1) can be put into a hydrodynamical form (i.e., it is equivalent to a system of Euler equations for a compressible inviscid fluid of density ρ and velocity $\nabla\psi_0$). In this context, $\frac{1}{\varepsilon\sqrt{2}}$ represents the sound velocity at infinity. It will be proved at the beginning of section 3 that (1.1) does not possess nonconstant travelling-waves moving with velocity $|c| \geq \frac{1}{\varepsilon\sqrt{2}}$. Hence we will assume throughout that $|c| < \frac{1}{\varepsilon\sqrt{2}}$.

Observe that the system (1.9)–(1.10) has a good variational formulation: its solutions are critical points of the "energy" functional. Indeed, since $1 + \tilde{r} = |\tilde{\psi}| \geq 0$, it is clear that we must have $\tilde{r} \geq -1$. Therefore we will look for solutions r of (1.9) with $r > -1$. Let $V = \{r \in H^1(\mathbf{R}) \mid \inf_{x\in\mathbf{R}} r(x) > -1\}$. It is obvious that V is open in $H^1(\mathbf{R})$ because $H^1(\mathbf{R}) \subset C_b^0(\mathbf{R})$ by the Sobolev embedding. A pair $(r,u) \in V \times H^1(\mathbf{R})$ satisfy (1.9)–(1.10) if and only if (r,u) is a critical point of the C^∞ functional $E : V \times H^1(\mathbf{R}) \longrightarrow \mathbf{R}$,

$$
\begin{aligned}
E(r,u) \quad = \quad & \int_{\mathbf{R}} |r'|^2 dx + \frac{1}{2}\int_{\mathbf{R}} \Big((1+r)^2 - 1\Big)^2 \left(1 - \frac{2c^2\varepsilon^2}{(1+r)^2}\right) dx \\
& + \int_{\mathbf{R}} u^2(1+r)^2 dx + \frac{1}{q^2}\int_{\mathbf{R}} |u'|^2 dx - \frac{\lambda}{q^2}\int_{\mathbf{R}} u^2 dx.
\end{aligned}
\tag{1.13}
$$

However, $E(r,\cdot)$ is quadratic in u for any fixed r and it would be very difficult to find critical points of E by using a classical topological argument.

In this paper we use bifurcation theory to show the existence of nontrivial solitary waves for the system (1.1). Note that this system (or, equivalently, (1.9)–(1.10)) is invariant by translations. To avoid the degeneracy of the linearized system due to this invariance, we work on symmetric function spaces. Consequently, the travelling-waves that we obtain will also present a symmetry. To be more precise, we will use the spaces

$$\mathbf{H} = H^2_{rad}(\mathbf{R}) = \{u \in H^2(\mathbf{R}) \mid u(x) = u(-x) \ \ \forall x \in \mathbf{R}\} \quad \text{and}$$

$$\mathbf{L} = L^2_{rad}(\mathbf{R}) = \{u \in L^2(\mathbf{R}) \mid u(x) = u(-x) \ \text{ a.e. } \ x \in \mathbf{R}\}.$$

Clearly $\mathbf{H}\cap V$ is an open set of $\mathbf{H}$. We define $S : (\mathbf{H}\cap V)\times\mathbf{H} \longrightarrow \mathbf{L}$, $T : \mathbf{R}\times\mathbf{H}\times\mathbf{H} \longrightarrow \mathbf{L}$,

$$S(r,u) = -r'' + (1+r)^3 - (1+r) - c^2\varepsilon^2\left(1 + r - \frac{1}{(1+r)^3}\right) + (1+r)u^2, \tag{1.14}$$

$$T(\lambda,r,u) = -u'' + (q^2(1+r)^2 - \lambda)u. \tag{1.15}$$

It is obvious that S and T are well defined and of class C^∞ (recall that $\mathbf{H} \subset C_b^1(\mathbf{R})$ and $\mathbf{H}$ is an algebra). Clearly r and u satisfy the system (1.9)–(1.10) if and only if $S(r,u) = 0$ and $T(\lambda,r,u) = 0$.

In the next section, we will study the structure of the set of nontrivial solutions in a neighborhood of the trivial ones. It follows easily from the implicit function theorem that there are no nontrivial solutions of (1.9)–(1.10) in a neighborhood of $(\lambda,0,0)$ for $\lambda < q^2$ (see the proof of Theorem 3.8). It is well known that we may have nontrivial solutions arbitrarily close to $(\lambda, r_{2c\varepsilon}, 0)$ if and only if the differential $d_{(r,u)}(S,T)(\lambda, r_{2c\varepsilon}, 0)$ is not invertible. For $\lambda < q^2$, we will see that $d_{(r,u)}(S,T)(\lambda, r_{2c\varepsilon}, 0)$ is not invertible if and only if λ is an eigenvalue of the particular Schrödinger operator given by (1.10). In this case we show that all the nontrivial solutions in a neighborhood of $(\lambda, r_{2c\varepsilon}, 0)$ form a smooth curve in $\mathbf{R} \times \mathbf{H} \times \mathbf{H}$.

It is natural to ask how long such a branch of solutions exists. Recently, there were obtained general global bifurcation results for C^1 Fredholm mappings of index 0 which apply to a broad class of elliptic equations in $\mathbf{R}^N$ (see, e.g., [9], [10]). Using the ideas and techniques developed in [11] it can be proved that for any fixed $\lambda < q^2$, the mapping $(S, T(\lambda, \cdot, \cdot)) : (\mathbf{H} \cap V) \times \mathbf{H} \longrightarrow \mathbf{L} \times \mathbf{L}$ is Fredholm of index 0. By a general global bifurcation theorem (a variant of Theorem 6.1 in [9]), one can prove that either the branch of nontrivial solutions of (1.9)–(1.10) starting from a bifurcation point $(\lambda, r_{2c\varepsilon}, 0)$ is noncompact in $\mathbf{R} \times \mathbf{H} \times \mathbf{H}$ or it meets $[q^2, \infty) \times \mathbf{H} \times \mathbf{H}$ (note that $[q^2, \infty)$ is the essential spectrum of the linear Schrödinger operator appearing in (1.10)).

To obtain further information (such as unboundedness) about the branches of nontrivial solutions, a key ingredient would be the properness of the operator (S,T), at least on closed bounded sets. Unfortunately, it is easy to see that the operator (S,T) is not proper on closed bounded sets. Indeed, it suffices to take $r_n = r_{2c\varepsilon}(\cdot - n) + r_{2c\varepsilon}(\cdot + n)$ and to observe that $(S,T)(\lambda, r_n, 0) \longrightarrow (0,0)$ as $n \longrightarrow \infty$; the sequence (r_n) is bounded in $\mathbf{H}$ but has no convergent subsequence.

In order to obtain a more precise description of the branches of nontrivial solutions and to avoid troubles due to the lack of properness, we choose a different approach: we reformulate the problem and we work on a weighted Sobolev space (which is a subspace of $\mathbf{H}$). In section 3, we use a variant of the global bifurcation theorem of Rabinowitz [12] to obtain global branches of solutions of (1.9)–(1.10) in that space. Note that the use of a slowly increasing weight (for example, $(1+x^2)^s$ for $s > 0$) is sufficient to eliminate the lack of properness and to obtain global branches of travelling-waves. It is worth noting that for $\lambda < q^2$, any nontrivial travelling-wave which is in $\mathbf{H}$ also belongs to the weighted space which is used (i.e., there is no loss of solutions). We show that there exists exactly one branch of nontrivial solutions bifurcating from the curve $(\lambda, r_{2c\varepsilon}, 0)$ if $q \leq \frac{1}{\sqrt{2 \ln 2}}$. The number of these branches is increasing with q and tends to infinity as $q \longrightarrow \infty$. We will prove that any of these branches is either unbounded (in the weighted space) or λ tends to q^2 along it. On the other hand, we prove that there are no nontrivial solutions of (1.9)–(1.10) for $\lambda > q^2$.

2. Local curves of solutions. In order to prove a local existence result of nontrivial solitary waves for the system (1.1), we have to study the properties of the linear operator $A = -\frac{d^2}{dx^2} + q^2(1 + r_{2c\varepsilon})^2$, which can be written as $A = -\frac{d^2}{dx^2} + q^2 r_{2c\varepsilon}(2 + r_{2c\varepsilon}) + q^2$. Since $-1 < r_{2c\varepsilon}(x) < 0$ for any $x \in \mathbf{R}$, the function $r_{2c\varepsilon}(2 + r_{2c\varepsilon})$ is everywhere negative (and even). Actually, in a slightly more general framework, we will study the operator $L = -\frac{d^2}{dx^2} + V(x)$ for a negative potential V, the properties of A being then deduced from those of L by a shift. For any $\lambda \leq 0$, we also consider the Cauchy problem

$$(2.1) \qquad \begin{cases} -u''(x) + V(x)u(x) = \lambda u(x), \\ u(0) = 1, \quad u'(0) = 0. \end{cases}$$

If V is continuous and even (i.e., $V(x) = V(-x)$), it is clear that problem (2.1) has a unique *global* solution which is also even. We denote by u_λ this solution and by $n(\lambda)$ the number of zeroes of u_λ in $(0, \infty)$.

Proposition 2.1. *Let $V \in L^2 \cap L^\infty(\mathbf{R}^N)$, $V \not\equiv 0$ be continuous, less than or equal to zero, even, and satisfy $\lim_{x \longrightarrow \pm\infty} V(x) = 0$. The operator $L = -\frac{d^2}{dx^2} + V(x) : \mathbf{H} \longrightarrow \mathbf{L}$ is self-adjoint and has the following properties:*

(i) $\sigma_{ess}(L) = [0, \infty)$.

(ii) *L has at least one negative eigenvalue.*

(iii) *Any eigenvalue of L is simple.*

(iv) *For any $\lambda < 0$ and $\varepsilon > 0$, there exists $C > 0$ such that*

$$|u_\lambda^{(m)}(x)| \le Ce^{\sqrt{-\lambda+\varepsilon}|x|}, \qquad m = 0, 1, 2. \tag{2.2}$$

If $\lambda < 0$ is an eigenvalue and $0 < \varepsilon < -\lambda$, there exist $C_1, C_2, M > 0$ such that

$$C_1 e^{-\sqrt{-\lambda+\varepsilon}|x|} \le |u_\lambda^{(m)}(x)| \le C_2 e^{-\sqrt{-\lambda-\varepsilon}|x|} \qquad \text{on} \qquad [M, \infty), \qquad m = 0, 1, 2. \tag{2.3}$$

(v) *For any $\lambda \le 0$, the number of eigenvalues of L in $(-\infty, \lambda)$ is exactly $n(\lambda)$, the number of zeroes of u_λ in $(0, \infty)$.*

(vi) *If $\int_0^\infty x|V(x)|dx < \infty$, then L has at most $1 + \int_0^\infty x|V(x)|dx$ negative eigenvalues.*

Proof. (i) The operator $-\frac{d^2}{dx^2} + V(x)$ on $L^2(\mathbf{R})$ (with domain $H^2(\mathbf{R})$) is self-adjoint, so it is easy to see that L is self-adjoint. Multiplication by V is a relatively compact perturbation of $-\Delta$ and it follows from a classical theorem of Weyl that $\sigma_{ess}(L) = \sigma_{ess}(-\Delta) = [0, \infty)$.

(ii) It suffices to show that there exists $u \in \mathbf{H}$ such that $\langle Lu, u\rangle_{\mathbf{L}} < 0$ and it will follow from the min-max principle (see [13, Theorem XIII.1, p. 76]) that L has negative eigenvalues. Consider an even function $u \in C_0^\infty$ such that $u \equiv 1$ on $[-1, 1]$ and u is nonincreasing on $[0, \infty)$. Let $u_n(x) = u(\frac{x}{n})$. Then

$$\langle Lu_n, u_n\rangle_{\mathbf{L}} = \frac{1}{n}\int_{\mathbf{R}} |u'(x)|^2 dx + \int_{\mathbf{R}} \left|u\left(\frac{x}{n}\right)\right|^2 V(x)dx \longrightarrow \int_{\mathbf{R}} V(x)dx < 0$$

as $n \longrightarrow \infty$, and thus $\langle Lu_n, u_n\rangle_{\mathbf{L}} < 0$ for n sufficiently large.

(iii) Clearly, λ is an eigenvalue of L if and only if $u_\lambda \in \mathbf{H}$. If this is the case, it is obvious that $\mathrm{Ker}(L - \lambda) = Span\{u_\lambda\}$. Since L is self-adjoint, we have $\mathrm{Ker}(L - \lambda) \cap \mathrm{Im}(L - \lambda) = \{0\}$, so for any $n \in \mathbf{N}^*$ we have $\mathrm{Ker}(L - \lambda)^n = \mathrm{Ker}(L - \lambda) = Span\{u_\lambda\}$; that is, λ is a simple eigenvalue.

(iv) By (2.1), u_λ and u_λ' cannot vanish simultaneously, so u_λ must change sign any time it vanishes and u_λ has only isolated zeroes. There exists $d > 0$ such that $V(x) - \lambda > -\frac{\lambda}{2} > 0$ on $[d, \infty)$ because $V(x) \longrightarrow 0$ as $x \longrightarrow \infty$. Two situations may occur:

1. There exists $x_0 > d$ such that $u_\lambda(x_0)$ and $u_\lambda'(x_0)$ have the same sign, say, are positive. Then $u_\lambda'' = (V(x) - \lambda)u_\lambda$, and thus u_λ'' will remain positive after x_0 as long as $u_\lambda > 0$, which implies that u_λ' is increasing, hence it remains positive as long as $u_\lambda > 0$. Consequently, u_λ is increasing after x_0 as long as it remains positive, which implies that u_λ is positive and increasing on $[x_0, \infty)$. Since $u_\lambda'(x) \ge u_\lambda'(x_0) > 0$ for any $x > x_0$, we have necessarily $\lim_{x\to\infty} u_\lambda(x) = \infty$. By (2.1) we find that $\lim_{x\to\infty} u_\lambda''(x) = \infty$, so we have also $\lim_{x\to\infty} u_\lambda'(x) = \infty$. Let $f(x) = (u_\lambda'(x))^2$ and $g(x) = u_\lambda^2(x)$. Clearly, $f(x) \longrightarrow \infty$, $g(x) \longrightarrow \infty$ as $x \longrightarrow \infty$ and

$$\frac{f'(x)}{g'(x)} = \frac{u_\lambda''(x)}{u_\lambda(x)} = V(x) - \lambda \longrightarrow -\lambda \qquad \text{as } x \longrightarrow \infty.$$

L'Hôpital's rule implies that $\lim_{x\to\infty} \frac{f(x)}{g(x)} = -\lambda$, which gives $\lim_{x\to\infty} \frac{u_\lambda'(x)}{u_\lambda(x)} = \sqrt{-\lambda}$. Thus, for any $\epsilon > 0$ there exists $x_\epsilon > 0$ such that

$$\sqrt{-\lambda-\epsilon} < \frac{u_\lambda'(x)}{u_\lambda(x)} < \sqrt{-\lambda+\epsilon} \qquad \text{on } [x_\epsilon, \infty). \tag{2.4}$$

Integrating (2.4) from x_ϵ to x we get, for any $x > x_\epsilon$,

$$\sqrt{-\lambda-\epsilon}(x-x_\epsilon) < \ln u_\lambda(x) - \ln u_\lambda(x_\epsilon) < \sqrt{-\lambda+\epsilon}(x-x_\epsilon),$$

that is,

$$u_\lambda(x_\epsilon)e^{\sqrt{-\lambda-\epsilon}(x-x_\epsilon)} < u_\lambda(x) < u_\lambda(x_\epsilon)e^{\sqrt{-\lambda+\epsilon}(x-x_\epsilon)} \qquad \text{for any } x > x_\epsilon. \tag{2.5}$$

Note that the above situation always occurs if u_λ has a zero in (d,∞). Indeed, if $u_\lambda(x_0) = 0$, then necessarily $u_\lambda(x)$ and $u_\lambda'(x)$ have opposite signs for $x < x_0$ and x close to x_0 (because if u_λ and u_λ' have the same sign at some $x_1 \in (d, x_0)$, we have just seen that u_λ cannot vanish after x_1). But u_λ changes sign at x_0 and $u_\lambda'(x_0) \neq 0$, hence u_λ and u_λ' have the same sign just after x_0.

2. The functions u_λ and u_λ' have opposite signs in (d,∞). Replacing u_λ by $-u_\lambda$ if necessary, we may suppose that $u_\lambda > 0$ and $u_\lambda' < 0$ in (d,∞) (observe that u_λ' cannot vanish because it also changes sign at any zero and we would be in case 1). So u_λ is decreasing and positive on (d,∞). Let $l = \lim_{x\to\infty} u_\lambda(x)$. Clearly, $l \geq 0$. If $l > 0$, then $u_\lambda''(x) \longrightarrow -\lambda l > 0$ as $x \longrightarrow \infty$ by (2.1), which implies $u_\lambda'(x) \longrightarrow \infty$ as $x \longrightarrow \infty$, a contradiction. Thus necessarily $l = 0$. Also, u_λ' is increasing on (d,∞) (because $u_\lambda''(x) = (V(x)-\lambda)u_\lambda(x) > 0$) and negative, so it also has a limit at infinity. Since u_λ converges (to zero) at infinity, we must have $\lim_{x\to\infty} u_\lambda'(x) = 0$. Now we may apply l'Hôpital's rule to get

$$\lim_{x\to\infty} \frac{(u_\lambda'(x))^2}{u_\lambda^2(x)} = \lim_{x\to\infty} \frac{u_\lambda''(x)}{u_\lambda(x)} = \lim_{x\to\infty} (V(x)-\lambda) = -\lambda.$$

Thus $\frac{u_\lambda'(x)}{u_\lambda(x)} \longrightarrow -\sqrt{-\lambda}$ as $x \longrightarrow \infty$ because u_λ and u_λ' have opposite signs at infinity. Given $\epsilon > 0$, there exists $M > d$ such that

$$-\sqrt{-\lambda+\epsilon} < \frac{u_\lambda'(x)}{u_\lambda(x)} < -\sqrt{-\lambda-\epsilon} \qquad \text{on } [M,\infty). \tag{2.6}$$

Integrating (2.6) on $[M,x]$ we obtain, as in case 1,

$$u_\lambda(M)e^{-\sqrt{-\lambda+\epsilon}(x-M)} < u_\lambda(x) < u_\lambda(M)e^{-\sqrt{-\lambda-\epsilon}(x-M)} \qquad \text{for any } x > M. \tag{2.7}$$

Finally, (2.2) and (2.3) follow from (2.5), respectively (2.7), and the fact that $\lim_{x\to\infty} \frac{u_\lambda''(x)}{u_\lambda(x)} = -\lambda$, $\lim_{x\to\infty} \frac{u_\lambda'(x)}{u_\lambda(x)} = \pm\sqrt{-\lambda}$. It is obvious that λ is an eigenvalue of L if and only if $u_\lambda \in \mathbf{H}$, i.e., if and only if we are in case 2. Therefore assertion (iv) is proved.

Note also that u_λ has only a finite number of zeroes. Indeed, it follows from the above arguments that u_λ has at most one zero in (d,∞) and we know that any zero is isolated, and thus there are only finitely many zeroes in $[0,d]$.

The proofs of (v) and (vi) are rather classical and are similar to the proofs of Theorems XIII.8 and XIII.9, pp. 90–94 in [13]. The bound on the number of eigenvalues given by (vi) is due to Bargmann (see [13] and references therein). $\square$

COROLLARY 2.2. *The linear operator* $A = -\frac{d^2}{dx^2} + q^2(1 + r_{2c\varepsilon})^2$ *(considered on* $\mathbf{L}$ *with domain* $D(A) = \mathbf{H}$*) is self-adjoint and has the following properties:*

(i) $A \geq 2c^2\varepsilon^2 q^2$ *and* $\sigma_{ess}(A) = [q^2, \infty)$.

(ii) *A has at least one eigenvalue in* $[2c^2\varepsilon^2 q^2, q^2)$.

(iii) *Any eigenvalue of A is simple. If* $\mu < q^2$ *is an eigenvalue and* u_μ *is a corresponding eigenvector, then for any* $\epsilon > 0$*, there exist* $C_1, C_2, M > 0$ *such that*

$$C_1 e^{-\sqrt{q^2-\mu+\epsilon}|x|} \leq |u_\mu^{(m)}(x)| \leq C_2 e^{-\sqrt{q^2-\mu-\epsilon}|x|} \qquad \text{if } |x| \geq M, \quad m = 0, 1, 2. \tag{2.8}$$

(iv) *Let* N_q *be the number of eigenvalues of A in* $[2c^2\varepsilon^2 q^2, q^2)$*. We have* $N_q < 1 + (2\ln 2)q^2$*. In particular, if* $q \leq \frac{1}{\sqrt{2\ln 2}}$*, then A has exactly one eigenvalue less than* q^2.

(v) *We have* $N_q \longrightarrow \infty$ *as* $q \longrightarrow \infty$.

It can be proved that there exist $c_1, c_2, q_0 > 0$ such that $c_1 q \leq N_q \leq c_2 q$ for any $q \geq q_0$, but we will not make use of this result in what follows.

Proof. Recall that $r_{2c\varepsilon}$ is given by (1.12). We have $A = -\frac{d^2}{dx^2} + q^2 V(x) + q^2$, where the function V given by $V(x) = (1+r_{2c\varepsilon}(x))^2 - 1 = (1-2c^2\varepsilon^2)(-1+\tanh^2(\sqrt{\frac{1-2c^2\varepsilon^2}{2}}x))$ is even, negative, tends exponentially to zero as $x \longrightarrow \pm\infty$, and $\inf_{x\in\mathbf{R}} V(x) = 2c^2\varepsilon^2 - 1$. Obviously, μ is an eigenvalue of A if and only if $\mu - q^2$ is an eigenvalue of $-\frac{d^2}{dx^2} + q^2 V(x)$, so (i), (ii), and (iii) follow at once from Proposition 2.1.

An easy computation gives

$$\begin{aligned} \int_0^\infty x|V(x)|dx &= (1 - 2c^2\varepsilon^2)\int_0^\infty x\left(1 - \tanh^2\left(\sqrt{\frac{1-2c^2\varepsilon^2}{2}}x\right)\right)dx \\ &= 2\int_0^\infty y(1-\tanh^2 y)dy = 2\int_0^\infty y(\tanh y - 1)' dy = 2\ln 2. \end{aligned}$$

Now (iv) is a direct consequence of Proposition 2.1(vi).

(v) Fix $n \in \mathbf{N}$, $n \geq 1$ and take n symmetric functions $\varphi_1, \ldots, \varphi_n \in C_0^\infty(\mathbf{R})$, $\varphi_i \not\equiv 0$ such that $\text{supp}(\varphi_i) \cap \text{supp}(\varphi_j) = \emptyset$ if $i \neq j$. Clearly,

$$\langle A\varphi_i, \varphi_i\rangle_{\mathbf{L}} - q^2\langle \varphi_i, \varphi_i\rangle_{\mathbf{L}} = \int_{\mathbf{R}} |\nabla\varphi_i|^2 dx + q^2\int_{\mathbf{R}} V(x)|\varphi_i(x)|^2 dx \longrightarrow -\infty \quad \text{ as } q \longrightarrow \infty.$$

Hence, there exists $q_0 > 0$ such that for any $q \geq q_0$ and any $i = 1, \ldots, n$ we have $\langle A\varphi_i, \varphi_i\rangle_{\mathbf{L}} - q^2\langle \varphi_i, \varphi_i\rangle_{\mathbf{L}} < 0$. Since the φ_i's have disjoint supports, we get

$$\begin{aligned} &\left\langle A\left(\sum_{i=1}^n \alpha_i\varphi_i\right), \sum_{i=1}^n \alpha_i\varphi_i\right\rangle_{\mathbf{L}} - q^2\left\|\sum_{i=1}^n \alpha_i\varphi_i\right\|_{\mathbf{L}}^2 \\ &= \sum_{i=1}^n |\alpha_i|^2\left(\int_{\mathbf{R}} |\nabla\varphi_i|^2 dx + q^2\int_{\mathbf{R}} V(x)|\varphi_i(x)|^2 dx\right) < 0. \end{aligned}$$

Therefore we have found an n-dimensional subspace of $\mathbf{H}$, $V_n = Span\{\varphi_1, \ldots, \varphi_n\}$, such that $\langle Au, u\rangle_{\mathbf{L}} - q^2||u||_{\mathbf{L}} < 0$ for any $u \in V_n$ and any $q \geq q_0$. By the min-max principle (see, e.g., [13, Theorem XIII.1, p. 76]) it follows that for $q \geq q_0$, A has at least n eigenvalues less than q^2; that is, $N_q \geq n$ if $q \geq q_0$. This proves (v). ∎

We have the following result concerning the existence of nontrivial solitary waves.

THEOREM 2.3. *Let $\lambda_* < q^2$ be an eigenvalue of A and let u_* be a corresponding eigenvector. There exists $\eta > 0$ and C^∞ functions*

$$s \longmapsto (\lambda(s), r(s), u(s)) \in \mathbf{R} \times \mathbf{H} \times (u_*^\perp \cap \mathbf{H})$$

defined on $(-\eta, \eta)$ such that $\lambda(0) = \lambda_*$, $r(0) = 0$, $u(0) = 0$, and

$$S(r_{2c\varepsilon} + sr(s), s(u_* + u(s))) = 0, \quad T(\lambda(s), r_{2c\varepsilon} + sr(s), s(u_* + u(s))) = 0.$$

Moreover, there exists a neighborhood U of $(\lambda_*, r_{2c\varepsilon}, 0)$ in $\mathbf{R} \times \mathbf{H} \times \mathbf{H}$ such that any solution of $S(r,u) = 0$, $T(\lambda, r, u) = 0$ in U is either of the form $(\lambda(s), r_{2c\varepsilon} + sr(s), s(u_* + u(s)))$ or of the form $(\lambda, r_{2c\varepsilon}, 0)$.

That is, $r = r_{2c\varepsilon} + sr(s)$, $u = s(u_* + u(s))$ are nontrivial solutions of (1.9)–(1.10) for $\lambda = \lambda(s)$.

Let $g_{2c\varepsilon} : (-1, \infty) \longrightarrow \mathbf{R}$, $g_{2c\varepsilon}(x) = (1+x)^3 - (1+x) - c^2\varepsilon^2(1 + x - \frac{1}{(1+x)^3})$. Then $S(r,u)$ can be written as $S(r,u) = -r'' + g_{2c\varepsilon}(r) + (1+r)u^2$. It is easily seen that $d_r S(r_{2c\varepsilon}, 0) = -\frac{d^2}{dx^2} + g'_{2c\varepsilon}(r_{2c\varepsilon})$.

For the proof of Theorem 2.3, we need the following lemmas.

LEMMA 2.4. *The linear operator $J := -\frac{d^2}{dx^2} + g'_{2c\varepsilon}(r_{2c\varepsilon}) : \mathbf{H} \longrightarrow \mathbf{L}$ has the following properties:*

(i) *J is self-adjoint, invertible, and has the essential spectrum $\sigma_{ess}(J) = [2 - 4c^2\varepsilon^2, \infty)$.*

(ii) *J has exactly one negative eigenvalue and any eigenvalue of J is simple.*

Proof. (i) The linear operator $B = -\frac{d^2}{dx^2} + g'_{2c\varepsilon}(r_{2c\varepsilon})$ with domain $D(B) = H^2(\mathbf{R})$ is self-adjoint in $L^2(\mathbf{R})$. We claim that $\mathrm{Ker}(B) = Span\{\frac{d}{dx} r_{2c\varepsilon}\}$. Indeed, we have

$$\frac{d^2}{dx^2} r_{2c\varepsilon} = g_{2c\varepsilon}(r_{2c\varepsilon}). \tag{2.9}$$

Thus $r''_{2c\varepsilon} \in C^1(\mathbf{R})$. Differentiating (2.9) with respect to x we get $\frac{d}{dx} r_{2c\varepsilon} \in \mathrm{Ker}(B)$.

Conversely, let $h \in \mathrm{Ker}(B)$. Then $h'' = g'_{2c\varepsilon}(r_{2c\varepsilon})h$, so that

$$(h' r'_{2c\varepsilon})' = h'' r'_{2c\varepsilon} + h' r''_{2c\varepsilon} = h g'_{2c\varepsilon}(r_{2c\varepsilon}) r'_{2c\varepsilon} + h' g_{2c\varepsilon}(r_{2c\varepsilon}) = (h g_{2c\varepsilon}(r_{2c\varepsilon}))'.$$

Hence $h' r'_{2c\varepsilon} = h g_{2c\varepsilon}(r_{2c\varepsilon}) + C$ on $\mathbf{R}$. Taking the limits as $|x| \longrightarrow \infty$, we get $C = 0$, so $h' r'_{2c\varepsilon} = h g_{2c\varepsilon}(r_{2c\varepsilon}) = h r''_{2c\varepsilon}$. Since $r'_{2c\varepsilon} \neq 0$ on $(-\infty, 0)$ and on $(0, \infty)$, on each of these intervals we have $(\frac{h}{r'_{2c\varepsilon}})' = \frac{h' r'_{2c\varepsilon} - h r''_{2c\varepsilon}}{(r'_{2c\varepsilon})^2} = 0$. Thus there exist constants C_1, C_2 such that $h(x) = C_1 r'_{2c\varepsilon}(x)$ on $(-\infty, 0)$ and $h(x) = C_2 r'_{2c\varepsilon}(x)$ on $(0, \infty)$. Consequently, $h'(x) = C_1 r''_{2c\varepsilon}(x) = C_1 g(r_{2c\varepsilon}(x))$ on $(-\infty, 0)$ and $h'(x) = C_2 r''_{2c\varepsilon}(x) = C_2 g_{2c\varepsilon}(r_{2c\varepsilon}(x))$ on $(0, \infty)$. But h' is continuous because $h \in H^2(\mathbf{R})$ and therefore $C_1 = C_2$, which proves our claim.

Since $r'_{2c\varepsilon} \notin \mathbf{H}$, it is clear that the restriction of B to $\mathbf{H}$ is one-to-one from $\mathbf{H}$ into $\mathbf{L}$. It remains to prove that $B\mathbf{H} = \mathbf{L}$. It is well known that $\mathrm{Im}(B) = \mathrm{Ker}(B)^\perp = (r'_{2c\varepsilon})^\perp$ since B is self-adjoint. We have $\mathbf{L} \subset \mathrm{Im}(B)$ because $r'_{2c\varepsilon}$ is an odd function. Let $f \in \mathbf{L}$. Clearly there exists $r \in H^2(\mathbf{R})$ such that $Br = f$. Let $\tilde{r}(x) = r(-x)$. It is easy to see that $B\tilde{r} = f$, hence there exists C such that $r - \tilde{r} = C r'_{2c\varepsilon}$. Then $r - \frac{1}{2} C r'_{2c\varepsilon} = \frac{1}{2}(r + \tilde{r}) \in \mathbf{H}$ and $B(r - \frac{1}{2} C r'_{2c\varepsilon}) = f$.

Now it is clear that J, which is the restriction of B to $\mathbf{H}$, is self-adjoint in $\mathbf{L}$ and invertible. The function $g'_{2c\varepsilon}(r_{2c\varepsilon})$ tends (exponentially) to $g'_{2c\varepsilon}(0) = 2 - 4c^2\varepsilon^2$ as $x \longrightarrow \infty$. It follows from Weyl's theorem that $\sigma_{ess}(J) = \sigma_{ess}(B) = [2 - 4c^2\varepsilon^2, \infty)$. This completes the proof of (i).

(ii) It follows from Proposition 2.1(iii) and (v) that any eigenvalue of J is simple and the number of negative eigenvalues of J is exactly the number of zeroes of u in $(0,\infty)$, where u is the solution of the Cauchy problem

$$\left\{ \begin{array}{l} -u'' + g'_{2c\varepsilon}(r_{2c\varepsilon})u = 0 \quad \text{ in } [0,\infty), \\ u(0) = 1, \quad u'(0) = 0. \end{array} \right. \tag{2.10}$$

We use the following simplified version of the well-known Sturm oscillation lemma (this is also a particular case of Lemma 5 in [8]).

STURM OSCILLATION LEMMA. *Let Y and Z be nontrivial solutions of the differential equation*

$$-\varphi'' + h(x)\varphi = 0$$

on some interval (μ,ν), where h is continuous on (μ,ν). If Y and Z are linearly independent and $Y(\mu) = Y(\nu) = 0$, then Z has at least one zero in (μ,ν).

From this lemma it follows at once that J has at most one negative eigenvalue. Indeed, suppose that J has at least two negative eigenvalues. Then the solution u of (2.10) has at least two zeroes in $(0,\infty)$, say $x_1 < x_2$. But the function $r'_{2c\varepsilon}$ also satisfies the differential equation in (2.10), and obviously u and $r'_{2c\varepsilon}$ are linearly independent (because $r'_{2c\varepsilon}(0) = 0$). Using Sturm's oscillation lemma, we infer that $r'_{2c\varepsilon}$ must have a zero on (x_1,x_2), which is absurd because $r'_{2c\varepsilon}(x) > 0$ on $(0,\infty)$.

Now let us prove that J has (at least) one negative eigenvalue. We argue again by contradiction and we suppose that J has no negative eigenvalues. Then the solution u of (2.10) has no zeroes in $[0,\infty)$, consequently $u(x) > 0$ for any $x \in [0,\infty)$. Since $g'_{2c\varepsilon}(r_{2c\varepsilon}(x)) \longrightarrow 2 - 4c^2\varepsilon^2 > 0$ as $x \longrightarrow \infty$, repeating the argument used in the proof of Proposition 2.1(iv) we infer that either $u(x) \longrightarrow \infty$ or $u(x) \longrightarrow 0$ as $x \longrightarrow \infty$. In the latter case we have also

$$|u^{(m)}(x)| \le Ce^{-\sqrt{2-4c^2\varepsilon^2-\delta}|x|}, \qquad m = 0,1,2,$$

for some constant $C > 0$, $\delta \in (0, 2 - 4c^2\varepsilon^2)$, and x sufficiently large. Consequently, $u \in \mathbf{H}$ and 0 is an eigenvalue of J. But this is excluded by (i). Therefore we must have $u(x) \longrightarrow \infty$ as $x \longrightarrow \infty$.

Since $u(0) = 1$, we have $u > 0$ in a neighborhood of 0. Note that $g'_{2c\varepsilon}(r_{2c\varepsilon}(0)) = (5 + \frac{3}{2c^2\varepsilon^2})(c^2\varepsilon^2 - \frac{1}{2}) < 0$, hence $g'_{2c\varepsilon}(r_{2c\varepsilon}) < 0$ near 0. From (2.10) we get $u''(x) < 0$ for $x > 0$ and x close to 0. We have $u'(0) = 0$, so there exists $\delta > 0$ such that $u'(x) < 0$ on $(0,\delta]$. We may choose δ so small that $u(\delta) > 0$ and $r''_{2c\varepsilon}(\delta) > 0$ (note that $r''_{2c\varepsilon}(0) = g_{2c\varepsilon}(r_{2c\varepsilon}(0)) = \frac{(1-2c^2\varepsilon^2)^2}{2\sqrt{2}c\varepsilon} > 0$). Let $\beta = \frac{u(\delta)}{r'_{2c\varepsilon}(\delta)} > 0$ and let $h(x) = \beta r'_{2c\varepsilon}(x) - u(x)$. Clearly, h is a solution of the differential equation in (2.10) and $h(\delta) = 0$, $h'(\delta) = \beta r''_{2c\varepsilon}(\delta) - u'(\delta) > 0$. Hence $h(x) > 0$ for $x > \delta$ and x close to δ. On the other hand, we have $\lim_{x\longrightarrow\infty} h(x) = -\infty$, so there exists $\eta > \delta$ such that $h(\eta) = 0$. Since both $r'_{2c\varepsilon}$ and h satisfy the differential equation in (2.10), by the Sturm oscillation lemma we infer that $r'_{2c\varepsilon}$ must have a zero in (δ,η), which is absurd. This finishes the proof of Lemma 2.4. □

LEMMA 2.5. *We have the following:*

(i) $\operatorname{Ker}(T(\lambda_*, r_{2c\varepsilon}, \cdot)) = Span(u_*)$.

(ii) $\operatorname{Im}(T(\lambda_*, r_{2c\varepsilon}, \cdot)) = u_*^{\perp} \cap \mathbf{L}$.

The proof is obvious.

Proof of Theorem 2.3. Let $\tilde{V} = \{r \in \mathbf{H} \mid \sup_{x\in\mathbf{R}} |r(x)| < 1\}$ and $I = (-\sqrt{2}c\varepsilon, \sqrt{2}c\varepsilon)$. Clearly $\tilde{V}$ is open in $\mathbf{H}$. We define $F : I \times \mathbf{R} \times \tilde{V} \times (\mathbf{H} \cap u_*^\perp) \longrightarrow \mathbf{L} \times \mathbf{L}$ by

$$F(s,\lambda,r,u) = \begin{cases} \begin{pmatrix} \frac{1}{s} S(r_{2c\varepsilon} + sr, s(u_* + u)) \\ \\ \frac{1}{s} T(\lambda, r_{2c\varepsilon} + sr, s(u_* + u)) \end{pmatrix} & \text{if } s \neq 0, \\ \\ \begin{pmatrix} d_r S(r_{2c\varepsilon}, 0).r \\ T(\lambda, r_{2c\varepsilon}, u_* + u) \end{pmatrix} & \text{if } s = 0. \end{cases}$$

It is easily seen that F is C^∞ because

$$F_1(s,\lambda,r,u) = \tfrac{1}{s}\left(S(r_{2c\varepsilon} + sr, s(u_* + u)) - S(r_{2c\varepsilon}, 0)\right)$$

$$= \tfrac{1}{s}\int_0^1 \frac{d}{dt} S(r_{2c\varepsilon} + tsr, ts(u_* + u))dt$$

$$= \tfrac{1}{s}\int_0^1 d_r S(r_{2c\varepsilon} + tsr, ts(u_* + u)).sr + d_u S(r_{2c\varepsilon} + tsr, ts(u_* + u)).s(u_* + u)dt$$

$$= \int_0^1 d_r S(r_{2c\varepsilon} + tsr, ts(u_* + u)).r + d_u S(r_{2c\varepsilon} + tsr, ts(u_* + u)).(u_* + u)dt$$

and $F_2(s,\lambda,r,u) = T(\lambda, r_{2c\varepsilon} + sr, u_* + u)$.

It is also clear that $F(0,\lambda_*,0,0) = \binom{0}{0}$ and

$$d_{(\lambda,r,u)}F(0,\lambda_*,0,0)(\tilde{\lambda},\tilde{r},\tilde{u}) = \begin{pmatrix} 0 \\ -\tilde{\lambda}u_* \end{pmatrix} + \begin{pmatrix} d_r S(r_{2c\varepsilon},0).\tilde{r} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ T(\lambda_*, r_{2c\varepsilon}, \tilde{u}) \end{pmatrix}.$$

In view of Lemmas 2.4 and 2.5, $d_{(\lambda,r,u)}F(0,\lambda_*,0,0)$ is invertible. By the implicit function theorem, there exist $\eta > 0$ and C^∞ functions defined on $(-\eta,\eta)$,

$$s \longmapsto (\lambda(s), r(s), u(s)) \in \mathbf{R} \times \mathbf{H} \times (\mathbf{H} \cap u_*^\perp),$$

such that $\lambda(0) = \lambda_*$, $r(0) = 0$, $u(0) = 0$, and $F(s,\lambda(s),u(s),r(s)) = (0,0)$. It is obvious that for $s \neq 0$, $(\lambda(s), (r_{2c\varepsilon} + sr(s), s(u_0 + u(s))))$ satisfy the system (1.9)–(1.10). Finally, the uniqueness part in Theorem 2.3 is proved exactly in the same way as in the bifurcation from a simple eigenvalue theorem (see [4]). ∎

Remark 2.6. Let $\lambda(s)$, $r(s)$, $u(s)$ be given by Theorem 2.3. We have $\dot{\lambda}(0) = 0$, $\dot{u}(0) = 0$, and

$$\ddot{\lambda}(0) = -\frac{4q^2}{||u_*||^2_{\mathbf{L}}}\langle(1 + r_{2c\varepsilon})u_*^2, J^{-1}((1 + r_{2c\varepsilon})u_*^2)\rangle_{\mathbf{L}}, \tag{2.11}$$

where the dots denote derivatives with respect to s and J is the operator in Lemma 2.4.

To see this, we differentiate with respect to s the equation $T(\lambda(s), r_{2c\varepsilon} + sr(s), u_* + u(s)) = 0$ and then take $s = 0$ to obtain

$$-\frac{d^2}{dx^2}\dot{u}(0) + [q^2(1 + r_{2c\varepsilon})^2 - \lambda_*]\dot{u}(0) - \dot{\lambda}(0)u_* = 0, \tag{2.12}$$

that is, $(A-\lambda_*)\dot{u}(0)-\dot{\lambda}(0)u_*=0$. But $\mathrm{Im}(A-\lambda_*)$ and $\mathrm{Ker}(A-\lambda_*)=Span\{u_*\}$ are orthogonal (because A is self-adjoint), and thus (2.12) implies that $\dot{\lambda}(0)=0$ and $\dot{u}(0)=0$.

We differentiate twice with respect to s the equation $T(\lambda(s), r_{2c\varepsilon}+sr(s), u_*+u(s))=0$, then we take $s=0$ to get

$$(A-\lambda_*)\ddot{u}(0)+4q^2(1+r_{2c\varepsilon})\dot{r}(0)u_*-\ddot{\lambda}(0)u_*=0. \tag{2.13}$$

Substracting the equation $-r''_{2c\varepsilon}+g_{2c\varepsilon}(r_{2c\varepsilon})=0$ from the equation $S(r_{2c\varepsilon}+sr(s), s(u_*+u(s)))=0$ and then dividing by s we get

(2.14)
$$-\frac{d^2}{dx^2}r(s)+\int_0^1 g'_{2c\varepsilon}(r_{2c\varepsilon}+tsr(s))\,dt\cdot r(s)+s(1+r_{2c\varepsilon}+sr(s))(u_*+u(s))^2=0.$$

We differentiate (2.14) with respect to s, then we take $s=0$ to obtain

$$-\frac{d^2}{dx^2}\dot{r}(0)+g'_{2c\varepsilon}(r_{2c\varepsilon})\dot{r}(0)+(1+r_{2c\varepsilon})u_*^2=0;$$

that is, $J\dot{r}(0)+(1+r_{2c\varepsilon})u_*^2=0$, which can still be written as

$$\dot{r}(0)=-J^{-1}((1+r_{2c\varepsilon})u_*^2). \tag{2.15}$$

Taking the scalar product of (2.13) with u_* we find $\ddot{\lambda}(0)||u_*||^2_{\mathbf{L}}=4q^2\langle(1+r_{2c\varepsilon})u_*^2, \dot{r}(0)\rangle_{\mathbf{L}}$. We replace $\dot{r}(0)$ from (2.15) in the last equality to obtain (2.11).

3. Global branches of solutions. Our purpose is to obtain information about the global structure of the set of nontrivial solutions of (1.9)–(1.10). We give a nonexistence result first.

PROPOSITION 3.1. (i) *The system* (1.9)–(1.10) *does not admit solutions* $(\lambda, r, u)\in\mathbf{R}\times V\times H^1(\mathbf{R})$ *with* $(r,u)\neq(0,0)$ *if* $c\geq\frac{1}{\varepsilon\sqrt{2}}$.

(ii) *Suppose that* $c<\frac{1}{\varepsilon\sqrt{2}}$ *and let* $(\lambda, r, u)\in\mathbf{R}\times V\times H^1(\mathbf{R})$ *be a nontrivial solution of the system* (1.9)–(1.10). *Then* $2c^2\varepsilon^2q^2<\lambda\leq q^2$ *and* $-1+\sqrt{2}c\varepsilon<r(x)\leq 0$ *for any* $x\in\mathbf{R}$.

Proof. Let $(\lambda, r, u)\in\mathbf{R}\times V\times H^1(\mathbf{R})$ be a solution of (1.9)–(1.10). Since $H^1(\mathbf{R})\subset C_b(\mathbf{R})$, (1.9)–(1.10) imply that r'' and u'' are continuous; hence $r, u\in C^2(\mathbf{R})$.

If $u\equiv 0$ and $c\geq\frac{1}{\varepsilon\sqrt{2}}$, the only solution of (1.9) which tends to zero at $\pm\infty$ is $r\equiv 0$ (this was proved in [7], but can be easily deduced from the arguments below). From now on we suppose that $u\not\equiv 0$. Multiplying (1.10) by u and integrating we find

$$\int_{\mathbf{R}}|u'|^2dx+q^2\int_{\mathbf{R}}(1+r)^2|u|^2dx=\lambda\int_{\mathbf{R}}|u|^2dx. \tag{3.1}$$

Since $u\not\equiv 0$, we have necessarily $\lambda>0$. Let $G_{2c\varepsilon}(s)=\int_0^s g_{2c\varepsilon}(\tau)d\tau=\frac{1}{4}((1+s)^2-1)^2(1-\frac{2c^2\varepsilon^2}{(1+s)^2})$. Multiplying (1.9) by r' gives

$$-\frac{1}{2}[(r')^2]'+[G_{2c\varepsilon}(r)]'+\frac{1}{2}[(1+r)^2]'u^2=0, \tag{3.2}$$

and multiplying (1.10) by u' leads to

$$-\frac{1}{2}[(u')^2]'+\frac{1}{2}q^2(1+r)^2(u^2)'-\frac{\lambda}{2}(u^2)'=0. \tag{3.3}$$

From (3.2) and (3.3) we get

$$-\frac{1}{2}[(r')^2]' - \frac{1}{2q^2}[(u')^2]' + [G_{2c\varepsilon}(r)]' + \frac{1}{2}[(1+r)^2u^2]' - \frac{\lambda}{2q^2}(u^2)' = 0. \tag{3.4}$$

Integrating (3.4) from $-\infty$ to x and taking into account that $r(x) \longrightarrow 0$, $r'(x) \longrightarrow 0$, $u(x) \longrightarrow 0$, and $u'(x) \longrightarrow 0$ as $x \longrightarrow \pm\infty$, we obtain

(3.5)
$$|r'|^2(x) + \frac{1}{q^2}|u'|^2(x) + \left(\frac{\lambda}{q^2} - (1+r(x))^2\right)u^2(x) = 2G_{2c\varepsilon}(r(x)) \qquad \text{for any } x \in \mathbf{R}.$$

Suppose that there exists $x_0 \in \mathbf{R}$ such that $r(x_0) < \min(-1 + \frac{\sqrt{\lambda}}{q}, -1 + \sqrt{2}c\varepsilon)$. Then $\frac{\lambda}{q^2} - (1 + r(x_0))^2 > 0$ and the left-hand side of (3.5) is positive at x_0 (because $u(x_0) = u'(x_0) = 0$ and (1.10) would imply $u \equiv 0$) while $G_{2c\varepsilon}(r(x_0)) < 0$, a contradiction. Thus $r(x) \geq \min(-1 + \frac{\sqrt{\lambda}}{q}, -1 + \sqrt{2}c\varepsilon)$ for any $x \in \mathbf{R}$.

Suppose that $\lambda \leq 2c^2\varepsilon^2q^2$ (that is, $\frac{\sqrt{\lambda}}{q} \leq \sqrt{2}c\varepsilon$). Then we have $(1 + r(x))^2 \geq \frac{\lambda}{q^2}$ for any $x \in \mathbf{R}$ and (3.1) gives

$$\int_{\mathbf{R}} |u'|^2 dx + q^2 \int_{\mathbf{R}} \left((1+r)^2 - \frac{\lambda}{q^2}\right) u^2 dx = 0,$$

which implies $u \equiv 0$, again a contradiction. Therefore we have $\lambda > 2c^2\varepsilon^2q^2$ and $r(x) \geq -1 + \sqrt{2}c\varepsilon$ for any $x \in \mathbf{R}$. This is impossible if $\sqrt{2}c\varepsilon > 1$ because $r(x) \longrightarrow 0$ as $x \longrightarrow \pm\infty$.

Hence we cannot have solutions other than $(\lambda, 0, 0)$ if $\sqrt{2}c\varepsilon > 1$. From now on we suppose that $\sqrt{2}c\varepsilon \leq 1$. In this case we have $r \leq 0$ on $\mathbf{R}$ by the maximum principle. Indeed, the function $g_{2c\varepsilon}$ is strictly increasing and positive on $(0, \infty)$. Suppose that r achieves a positive maximum at x_0. Then $r''(x_0) \leq 0$. On the other hand, from (1.9) we infer that $r''(x_0) = g_{2c\varepsilon}(r(x_0)) + (1 + r(x_0))u^2(x_0) > 0$, which is absurd.

If $\sqrt{2}c\varepsilon = 1$ we have seen that $0 \geq r(x) \geq -1 + \sqrt{2}c\varepsilon = 0$; hence $r \equiv 0$. Then (1.10) becomes $u'' = (q^2 - \lambda)u$; together with the boundary condition $u(x) \longrightarrow 0$ as $x \longrightarrow \pm\infty$, this gives $u \equiv 0$. Thus (i) is proved.

From now on we suppose throughout that $2c^2\varepsilon^2 < 1$. Clearly, if $r(x_0) = -1+\sqrt{2}c\varepsilon$ for some $x_0 \in \mathbf{R}$, then (3.5) would imply $u(x_0) = u'(x_0) = 0$ (because $\lambda > 2c^2\varepsilon^2q^2$), hence $u \equiv 0$ by (1.10), which is impossible. Hence $0 \geq r(x) > -1 + \sqrt{2}c\varepsilon$ for any $x \in \mathbf{R}$.

It remains only to show that we cannot have nontrivial solutions with $\lambda > q^2$. Suppose that (λ, r, u) is such a solution. First, observe that r cannot vanish because (3.5) would give a contradiction. We prove that r decays sufficiently fast at infinity. Take $0 < \epsilon < \frac{\lambda}{q^2} - 1$. There exists $M_\epsilon > 0$ such that $(1 + r(x))^2 \leq 1 + \epsilon$ on $[M_\epsilon, \infty)$ (because $r(x) \longrightarrow 0$ as $x \longrightarrow \infty$). Using (3.5), we have on $[M_\epsilon, \infty)$

$$0 \leq \left(\frac{\lambda}{q^2} - 1 - \epsilon\right) u^2(x) \leq 2G_{2c\varepsilon}(r(x)),$$

hence $0 \leq (\frac{\lambda}{q^2} - 1 - \epsilon)\frac{u^2(x)}{|r(x)|} \leq 2\frac{|G_{2c\varepsilon}(r(x))|}{|r(x)|}$. Passing to the limit as $x \longrightarrow \infty$ we obtain $\lim_{x\to\infty} \frac{u^2(x)}{r(x)} = 0$. Dividing (1.9) by r we get

$$\frac{r''(x)}{r(x)} = \frac{g_{2c\varepsilon}(r(x))}{r(x)} + (1 + r(x))\frac{u^2(x)}{r(x)} \longrightarrow g'_{2c\varepsilon}(0) > 0 \qquad \text{as } x \longrightarrow \infty. \tag{3.6}$$

Since r'' must have at least one zero between two zeroes of r', (3.6) shows that r' has no zeroes in some neighborhood of infinity. In that neighborhood we have

$$\frac{(|r'(x)|^2)'}{(r^2(x))'} = \frac{r''(x)}{r(x)} \longrightarrow g'_{2c\varepsilon}(0) > 0 \qquad \text{as } x \longrightarrow \infty.$$

Since $r(x) \longrightarrow 0$ and $r'(x) \longrightarrow 0$ at infinity, we may apply l'Hôpital's rule to get $\lim_{x\to\infty}(\frac{r'(x)}{r(x)})^2 = g'_{2c\varepsilon}(0)$. We know that r and r' have constant signs in a neighborhood of infinity and that they cannot have the same sign because r tends to 0 at infinity, so necessarily $\lim_{x\to\infty}\frac{r'(x)}{r(x)} = -\sqrt{g'_{2c\varepsilon}(0)}$. The argument, already used in the proof of Proposition 2.1, shows that for any $\epsilon > 0$, there exists $C_\epsilon > 0$ such that

$$|r(x)| \le C_\epsilon e^{-\sqrt{g'_{2c\varepsilon}(0)-\epsilon}\, x} \qquad \text{for any } x \in [0,\infty).$$

Of course a similar estimate is valid on $(-\infty, 0]$. In particular, r^2+2r is a continuous, bounded function on $\mathbf{R}$ and $\lim_{x\to\pm\infty}|x|(r^2(x)+2r(x)) = 0$. Moreover, multiplication by r^2+2r is a bounded operator on $L^2(\mathbf{R})$; hence it is also bounded with respect to $-\frac{d^2}{dx^2}$ with relative bound zero. Consequently, by the Kato–Agmon–Simon theorem (see, e.g., [13, Theorem XIII.58, p. 226]), the operator $-\frac{d^2}{dx^2} + q^2(r^2+2r)$ (with domain $H^2(\mathbf{R})$ and range $L^2(\mathbf{R})$) cannot have eigenvalues embedded in the continuous spectrum $(0,\infty)$. This means exactly that the operator $-\frac{d^2}{dx^2} + q^2(1+r)^2$ has no eigenvalues in (q^2,∞) and contradicts the existence of a nontrivial solution (λ, r, u) with $\lambda > q^2$. $\square$

We will use the following variant of the global bifurcation theorem of Rabinowitz.

PROPOSITION 3.2. *Let E be a real Banach space and $\Omega \subset \mathbf{R}\times E$ an open set. Suppose that $G : \Omega \longrightarrow E$ is compact on closed, bounded subsets $\omega \subset \Omega$ such that $dist(\omega, \partial\Omega) > 0$ and is of the form $G(a,u) = L(a,u)+H(a,u)$, where L and H satisfy the following assumptions:*

(a) *$L(a,\cdot)$ is linear, compact for any fixed a and $(a,u) \longmapsto L(a,u)$ is continuous and compact on closed, bounded subsets $\omega \subset \Omega$ such that $dist(\omega, \partial\Omega) > 0$.*

(b) *For any closed, bounded subset $\omega \subset \Omega$ such that $dist(\omega, \partial\Omega) > 0$, there exists a function h_ω such that $h_\omega(s) \longrightarrow 0$ as $s \longrightarrow 0$ and*

$$||H(a,u)|| \le ||u||h_\omega(||u||) \qquad \textit{for any } (a,u) \in \omega.$$

(c) *There exists a_0 and $\epsilon > 0$ such that*
- *$(a_0, 0) \in \Omega$,*
- *for any $a \in [a_0-\epsilon, a_0+\epsilon] \setminus \{a_0\}$ we have $\mathrm{Ker}(Id - L(a,\cdot)) = \{0\}$,*
- *if $a_1 \in [a_0-\epsilon, a_0)$ and $a_2 \in (a_0, a_0+\epsilon]$, then*
$$ind(Id - L(a_1,\cdot), 0) \ne ind(Id - L(a_2,\cdot), 0).$$

Let

$$\mathcal{S} = \{(a,u) \in \Omega \mid u \ne 0 \textit{ and } u = G(a,u)\}$$

be the set of nontrivial solutions of the equation $u = G(a,u)$. Then $\mathcal{S}\cup\{(a_0,0)\}$ possesses a maximal subcontinuum (i.e., a maximal closed connected subset) $\mathcal{C}_{a_0}$ which contains $(a_0,0)$ and has at least one of the following properties:

(i) *$\mathcal{C}_{a_0}$ is unbounded.*

(ii) *$dist(\mathcal{C}_{a_0}, \partial\Omega) = 0$.*

(iii) *$\mathcal{C}_{a_0}$ meets $(a_1, 0)$, where $a_1 \ne a_0$ and $\mathrm{Ker}(Id - L(a_1,\cdot)) \ne \{0\}$.*

From the first assertion in (c) it follows that the index $ind(Id - L(a,\cdot),0) = deg(Id - L(a,\cdot), B(0,\rho), 0)$ is well defined for any $a \in [a_0 - \epsilon, a_0 + \epsilon] \setminus \{a_0\}$. By (a) and the homotopy invariance of the Leray–Schauder degree, it is a continuous function of a. Thus we have necessarily $\mathrm{Ker}(Id - L(a_0,\cdot)) \neq \{0\}$ (since otherwise $ind(Id - L(a_0,\cdot),0)$ would be defined and $ind(Id - L(a,\cdot)\ 0)$ would be constant for $a \in [a_0 - \epsilon, a_0 + \epsilon]$, contradicting the last assertion in (c)).

The proof of Proposition 3.2 is similar to that of Theorem 1.3, p. 490 in [12] (see also Corollary 1.12 in [12]).

Next, we give a reformulation of problem (1.9)–(1.10) suitable for the use of Proposition 3.2.

Equation (1.9) can be written as $-r'' + g_{2c\varepsilon}(r) + (1+r)u^2 = 0$, where $g_{2c\varepsilon}(x) = (1+x)^3 - (1+x) - c^2\varepsilon^2(1 + x - \frac{1}{(1+x)^3})$. We will seek solutions of the form $r(x) = r_{2c\varepsilon}(x) + w(x)$. Taking into account that $r_{2c\varepsilon}$ satisfies $-r''_{2c\varepsilon} + g_{2c\varepsilon}(r_{2c\varepsilon}) = 0$, (1.9) becomes

$$-w'' + g_{2c\varepsilon}(r_{2c\varepsilon} + w) - g_{2c\varepsilon}(r_{2c\varepsilon}) + (1 + r_{2c\varepsilon} + w)u^2 = 0. \tag{3.7}$$

Note that $g'_{2c\varepsilon}(0) = 2 - 4c^2\varepsilon^2 > 0$, thus the linear operator $-\frac{d^2}{dx^2} + g'_{2c\varepsilon}(0)$ (with domain $\mathbf{H}$ and range $\mathbf{L}$) is invertible, and thus (3.7) is equivalent to

(3.8)
$$\begin{aligned} w = &-\Big(-\tfrac{d^2}{dx^2} + g'_{2c\varepsilon}(0)\Big)^{-1}[g_{2c\varepsilon}(r_{2c\varepsilon} + w) - g_{2c\varepsilon}(r_{2c\varepsilon}) - g'_{2c\varepsilon}(r_{2c\varepsilon})w + (1 + r_{2c\varepsilon} + w)u^2] \\ &-\Big(-\tfrac{d^2}{dx^2} + g'_{2c\varepsilon}(0)\Big)^{-1}[(g'_{2c\varepsilon}(r_{2c\varepsilon}) - g'_{2c\varepsilon}(0))w]. \end{aligned}$$

In the same way, (1.10) can be written as

$$-u'' + (q^2 - \lambda)u = q^2(1 - (1 + r_{2c\varepsilon} + w)^2)u.$$

For $\lambda < q^2$, the linear operator $-\frac{d^2}{dx^2} + q^2 - \lambda$ is invertible and (1.10) becomes

$$\begin{aligned} u = &-q^2\left(-\frac{d^2}{dx^2} + q^2 - \lambda\right)^{-1}[(r^2_{2c\varepsilon} + 2r_{2c\varepsilon})u] \\ &-q^2\left(-\frac{d^2}{dx^2} + q^2 - \lambda\right)^{-1}[(w^2 + 2wr_{2c\varepsilon} + 2w)u]. \end{aligned} \tag{3.9}$$

We denote

$$\begin{aligned} H_1(w,u) &= \left(-\frac{d^2}{dx^2} + g'_{2c\varepsilon}(0)\right)^{-1}[g_{2c\varepsilon}(r_{2c\varepsilon} + w) - g_{2c\varepsilon}(r_{2c\varepsilon}) - g'_{2c\varepsilon}(r_{2c\varepsilon})w \\ &\qquad + (1 + r_{2c\varepsilon} + w)u^2], \\ H_2(\lambda, w, u) &= q^2\left(-\frac{d^2}{dx^2} + q^2 - \lambda\right)^{-1}[(w^2 + 2wr_{2c\varepsilon} + 2w)u], \\ A_\lambda(u) = A(\lambda, u) &= q^2\left(-\frac{d^2}{dx^2} + q^2 - \lambda\right)^{-1}[(r^2_{2c\varepsilon} + 2r_{2c\varepsilon})u], \\ B(w) &= \left(-\frac{d^2}{dx^2} + g'_{2c\varepsilon}(0)\right)^{-1}[(g'_{2c\varepsilon}(r_{2c\varepsilon}) - g'_{2c\varepsilon}(0))w]. \end{aligned}$$

It is easy to see that $A_\lambda,\ B : \mathbf{L} \longrightarrow \mathbf{H}$ are linear and continuous. Denote $V_{2c\varepsilon} = \{r \in \mathbf{H} \mid r + r_{2c\varepsilon} \in V\}$. It is obvious that $V_{2c\varepsilon}$ is open in $\mathbf{H}$. Since $\mathbf{H} \subset C_b^1(\mathbf{R})$ and $\mathbf{H}$ is an algebra, H_1 and H_2 are well defined and continuous from $V_{2c\varepsilon} \times \mathbf{H}$ and $(-\infty, q^2) \times \mathbf{H} \times \mathbf{H}$, respectively, to $\mathbf{H}$.

If $\lambda < q^2$, then (λ, r, u) satisfies the system (1.9)–(1.10) if and only if (λ, w, u) (where $w = r - r_{2c\varepsilon}$) satisfies the system (3.8)–(3.9), which is equivalent to

$$\begin{pmatrix} w \\ u \end{pmatrix} = -\begin{pmatrix} B & 0 \\ 0 & A_\lambda \end{pmatrix}\begin{pmatrix} w \\ u \end{pmatrix} - \begin{pmatrix} H_1(w,u) \\ H_2(\lambda, w, u) \end{pmatrix}. \tag{3.10}$$

We have already shown in the introduction that we cannot expect to have properness for problem (1.9)–(1.10). The counterexample that we have seen is essentially due to the invariance by translations of the system and to the fact that we have localized solutions. Of course passing from (1.9)–(1.10) to (3.10) should not prevent the same problems from appearing. To overcome this difficulty, we shall work on some weighted Sobolev space. As a "weight," we take a function $W : \mathbf{R} \longrightarrow \mathbf{R}$ which satisfies the following properties:

(W1) W is continuous and even, i.e., $W(x) = W(-x)$;

(W2) $W \geq 1$ and $\lim\limits_{x\to\infty} W(x) = \infty$;

(W3) There exists $C_W > 0$ such that $W(a+b) \leq C_W(W(a) + W(b))$.

It follows easily from (W1) and (W3) that there exist $K, s > 0$ such that $W(x) \leq K|x|^s$ for $|x| \geq 1$. Indeed, from (W3) we infer that for all $a \in \mathbf{R}$, $W(2^n a) \leq (2C_W)^n W(a)$. If $x \in [2^{n-1}, 2^n]$ and $M = \max_{x\in[0,1]} W(x)$, then

$$\begin{aligned} W(x) &\leq (2C_W)^n W\left(\frac{x}{2^n}\right) \leq 2C_W M(2C_W)^{n-1} \\ &= 2C_W M 2^{(n-1)(1+\log_2 C_W)} \leq 2C_W M x^{1+\log_2 C_W}. \end{aligned}$$

In particular, we get

$$e^{-a|\cdot|}W(\cdot) \in L^1 \cap L^\infty(\mathbf{R}) \quad \forall a > 0. \tag{W4}$$

For a function W satisfying (W1)–(W3) we consider the spaces

$$\mathbf{L}_W = \{\varphi \in \mathbf{L} \mid W\varphi \in \mathbf{L}\},$$

$$\mathbf{H}_W = \{\varphi \in \mathbf{H} \mid W\varphi, W\varphi', W\varphi'' \in \mathbf{L}\},$$

endowed with the norms $||\varphi||_{\mathbf{L}_W} = ||W\varphi||_{L^2}$, respectively $||\varphi||^2_{\mathbf{H}_W} = ||W\varphi||^2_{L^2} + ||W\varphi'||^2_{L^2} + ||W\varphi''||^2_{L^2}$. Equipped with these norms, $\mathbf{L}_W$ and $\mathbf{H}_W$ are Hilbert spaces. It is clear that $||\varphi||_{L^2} \leq ||\varphi||_{\mathbf{L}_W}$, $||\varphi||_{H^2} \leq ||\varphi||_{\mathbf{H}_W}$, and $\mathbf{L}_W$ (respectively, $\mathbf{H}_W$) is a dense subspace of $\mathbf{L}$ (respectively, of $\mathbf{H}$).

LEMMA 3.3. *The embedding $\mathbf{H}_W \subset C_b^1(\mathbf{R})$ is compact.*

Proof. It is clear that the embeddings $\mathbf{H}_W \subset H^2(\mathbf{R}) \subset C_b^1(\mathbf{R})$ are continuous. To prove compactness, consider an arbitrary sequence $u_n \rightharpoonup 0$ in $\mathbf{H}_W$ and let us show that $u_n \longrightarrow 0$ in $C_b^1(\mathbf{R})$. Fix $\epsilon > 0$. Let $K = \sup_n ||u_n||_{\mathbf{H}_W}$. There exists $M > 0$ such that $W(x) \geq \frac{K}{\epsilon}$ if $|x| \geq M$. It follows that $||u_n||_{H^2((-\infty,M)\cup(M,\infty))} \leq \epsilon$. By the Sobolev embedding theorem, we have $||u_n||_{L^\infty((-\infty,M]\cup[M,\infty))} + ||u_n'||_{L^\infty((-\infty,M]\cup[M,\infty))} \le C_S\epsilon$. On the other hand $u_{n|[-M,M]} \rightharpoonup 0$ in $H^2(-M,M)$. Since the embedding $H^2(-M,M) \subset C^1([-M,M])$ is compact, it follows that $u_n \longrightarrow 0$ in $C^1([-M,M])$, so $||u_n||_{L^\infty([-M,M])} + ||u_n'||_{L^\infty([-M,M])} \le \epsilon$ if n is sufficiently big. Thus $||u_n||_{L^\infty(\mathbf{R})} + ||u_n'||_{L^\infty(\mathbf{R})} \le (C_S+1)\epsilon$ for n sufficiently big. As ϵ was arbitrary, we infer that $u_n \longrightarrow 0$ in $C_b^1(\mathbf{R})$ and the lemma is proved. □

Lemma 3.4. *Let W satisfy* (W1)–(W3). *For any $a>0$, the operator $-\frac{d^2}{dx^2}+a : \mathbf{H}_W \longrightarrow \mathbf{L}_W$ is bounded and invertible. Moreover, the norm of $(-\frac{d^2}{dx^2}+a)^{-1}$ is uniformly bounded in $\mathcal{L}(\mathbf{L}_W,\mathbf{H}_W)$ when the parameter a remains in a compact subinterval of* $(0,\infty)$.

Proof. It is clear that

$$\left\|\left(-\frac{d^2}{dx^2}+a\right)v\right\|_{\mathbf{L}_W} = ||-v''+av||_{\mathbf{L}_W} \le C||v||_{\mathbf{H}_W},$$

and thus the operator is bounded. Since $-\frac{d^2}{dx^2}+a : \mathbf{H} \longrightarrow \mathbf{L}$ is bounded and invertible, it is clear that the restriction of $-\frac{d^2}{dx^2}+a$ to $\mathbf{H}_W$ is one-to-one and for any $f \in \mathbf{L}_W \subset \mathbf{L}$ there exists a unique $v \in \mathbf{H}$ such that $(-\frac{d^2}{dx^2}+a)v = f$. It remains only to prove that $v \in \mathbf{H}_W$ and $||v||_{\mathbf{H}_W} \le ||f||_{\mathbf{L}_W}$. Using the Fourier transform we get $(\xi^2+a)\widehat{v}(\xi) = \widehat{f}(\xi)$ or, equivalently, $\widehat{v}(\xi) = \frac{1}{\xi^2+a}\widehat{f}(\xi)$. Since $\mathcal{F}(e^{-\sqrt{a}|\cdot|})(\xi) = \frac{2\sqrt{a}}{\xi^2+a}$, we infer that

$$v = \frac{1}{2\sqrt{a}}(e^{-\sqrt{a}|\cdot|}) * f. \tag{3.11}$$

From (3.11) we get

$$\begin{aligned}|v(x)W(x)| &= \frac{1}{2\sqrt{a}}W(x)\left|\int_{\mathbf{R}} e^{-\sqrt{a}|x-y|}f(y)dy\right| \\ &\le \frac{C_W}{2\sqrt{a}}\int_{\mathbf{R}} W(x-y)e^{-\sqrt{a}|x-y|}|f(y)| + e^{-\sqrt{a}|x-y|}W(y)|f(y)|dy \\ &\le C_1(a)[((We^{-\sqrt{a}|\cdot|}) * |f|)(x) + (e^{-\sqrt{a}|\cdot|}) * (|f|W)(x)],\end{aligned}$$

that is, $|vW| \le C_1(a)[(We^{-\sqrt{a}|\cdot|}) * |f| + e^{-\sqrt{a}|\cdot|} * (|f|W)]$. But

$$||(We^{-\sqrt{a}|\cdot|}) * |f|\,||_{L^2} \le ||We^{-\sqrt{a}|\cdot|}||_{L^1}||f||_{L^2} \le ||We^{-\sqrt{a}|\cdot|}||_{L^1}||f||_{\mathbf{L}_W}$$

and

$$||e^{-\sqrt{a}|\cdot|} * (|f|W)||_{L^2} \le ||e^{-\sqrt{a}|\cdot|}||_{L^1}||Wf||_{L^2},$$

and thus we obtain from (3.11) that

$$||v||_{\mathbf{L}_W} \le C_2(a)||f||_{\mathbf{L}_W}, \tag{3.12}$$

where $C_2(a)$ remains bounded if $a \in [d,e]$, $0<d<e<\infty$.

In the same way, we have $\widehat{v'}(\xi) = i\xi\widehat{v}(\xi) = \frac{i\xi}{\xi^2+a}\widehat{f}(\xi)$; hence $v'(x) = -\frac{1}{2}\zeta_a * f(x)$, where $\zeta_a(x) = \mathrm{sgn}(x)e^{-\sqrt{a}|x|}$. Repeating the above argument we find

$$||v'W||_{L^2} \le C_3(a)||f||_{\mathbf{L}_W}, \tag{3.13}$$

where $C_3(a)$ remains bounded if a is in a compact interval of $(0,\infty)$.

Finally, using the equation satisfied by v we get $v'' = -f + av$; hence

$$||v''W||_{L^2} \le ||f||_{\mathbf{L}_W} + a||v||_{\mathbf{L}_W} \le (1 + aC_2(a))||f||_{\mathbf{L}_W}. \tag{3.14}$$

Lemma 3.4 follows from (3.12), (3.13), and (3.14). □

Note that the operator $-\frac{d^2}{dx^2} + a : \mathbf{H}_W \longrightarrow \mathbf{L}_W$ is not invertible if the weight W increases too fast at infinity. Indeed, if $f \in C_0^\infty(\mathbf{R})$ and $f \ge 0$, it is easily seen (e.g., from (3.11)) that the solution v of $-v'' + av = f$ behaves like $e^{-\sqrt{a}|\cdot|}$ at $\pm\infty$. If we take $W(x) = e^{b|x|}$ and $a < b^2$, then v does not belong to $\mathbf{H}_W$, and thus $-\frac{d^2}{dx^2} + a : \mathbf{H}_W \longrightarrow \mathbf{L}_W$ is not surjective.

The next lemma shows that we do not lose solutions if we work in $\mathbf{H}_W$ instead of $\mathbf{H}$.

Lemma 3.5. *Let (λ, r, u) be a solution of* (1.9)–(1.10) *with $r \in \mathbf{H}$, $u \in \mathbf{H}$, and $\lambda < q^2$. Then r and u belong to $\mathbf{H}_W$.*

Proof. We have already seen in Proposition 3.1 that $-1+\sqrt{2}c\varepsilon < r \le 0$. Applying Proposition 2.1(iv) (see also Corollary 2.2(iii)) for $V(x) = q^2(r^2(x) + 2r(x))$, we infer that for any $\epsilon > 0$, u, u', and u'' decay at $\pm\infty$ faster than $e^{-\sqrt{q^2-\lambda-\epsilon}|x|}$; hence $u \in \mathbf{H}_W$.

Since $g'_{2c\varepsilon}(0) > 0$ and $r(x) \longrightarrow 0$ as $|x| \longrightarrow \infty$, there exists $M > 0$ such that $r(x)g_{2c\varepsilon}(r(x)) \ge \frac{1}{2}g'_{2c\varepsilon}(0)r^2(x)$ if $|x| > M$.

Consider a symmetric function $\chi \in C_0^\infty(\mathbf{R})$ such that $\chi \equiv 1$ on $[-1,1]$, χ is nonincreasing on $[0,\infty)$, and $\operatorname{supp}(\chi) \subset [-2,2]$. We multiply (1.9) by $xr(x)\chi(\frac{x}{n})$ and integrate on $[0,\infty)$. Integrating by parts, we get

$$\begin{aligned}
&\int_0^\infty |r'|^2(x)x\chi\left(\frac{x}{n}\right)dx - \frac{1}{2}r^2(0) - \frac{1}{2}\int_0^\infty r^2(x)\left(\frac{2}{n}\chi'\left(\frac{x}{n}\right) + \frac{x}{n^2}\chi''\left(\frac{x}{n}\right)\right)dx\\
&+\int_0^M g_{2c\varepsilon}(r(x))r(x)x\chi\left(\frac{x}{n}\right)dx + \int_M^\infty g_{2c\varepsilon}(r(x))r(x)x\chi\left(\frac{x}{n}\right)dx\\
&+\int_0^\infty (1+r(x))u^2(x)r(x)x\chi\left(\frac{x}{n}\right)dx = 0.
\end{aligned} \tag{3.15}$$

By the monotone convergence theorem, the first integral in (3.15) tends to $\int_0^\infty |r'(x)|^2 xdx$ as $n \longrightarrow \infty$, while the fourth integral tends to $\int_M^\infty g_{2c\varepsilon}(r(x))r(x)xdx$. The other three integrals converge as $n \longrightarrow \infty$ by Lebesgue's theorem on dominated convergence. Letting $n \longrightarrow \infty$ in (3.15) we obtain

$$\begin{aligned}
&\int_0^\infty |r'|^2(x)xdx - \frac{1}{2}r^2(0) + \int_0^M g_{2c\varepsilon}(r(x))r(x)xdx\\
&+\int_M^\infty g_{2c\varepsilon}(r(x))r(x)xdx + \int_0^\infty r(x)(1+r(x))xu^2(x)dx = 0.
\end{aligned} \tag{3.16}$$

Since the second and the last integral in (3.16) are finite (because u decays exponentially at $\pm\infty$), we infer that $\int_0^\infty |r'|^2(x)xdx < \infty$ and $\int_M^\infty g_{2c\varepsilon}(r(x))r(x)xdx < \infty$. Consequently, $|x|^{\frac{1}{2}}r'(x)$ and $|x|^{\frac{1}{2}}r(x)$ belong to $L^2(\mathbf{R})$.

We have $g_{2c\varepsilon}(s) = g'_{2c\varepsilon}(0)s + h(s)s^2$, where h is continuous on $(-1,\infty)$; hence $h(r(x))$ is bounded. Equation (1.9) can be written as

$$-r'' + g'_{2c\varepsilon}(0)r = -(1+r)u^2 - h(r)r^2, \tag{3.17}$$

which gives, as in the proof of Lemma 3.4,

$$r = -\frac{1}{2\sqrt{g'_{2c\varepsilon}(0)}} e^{-\sqrt{g'_{2c\varepsilon}(0)}|\cdot|} * ((1+r)u^2 + h(r)r^2). \tag{3.18}$$

Suppose that $|x|^\alpha r(x) \in L^2(\mathbf{R})$ for some $\alpha > 0$. Since $|x|^\beta u(x) \in L^p(\mathbf{R})$ for any $\beta > 0$ and $1 \le p \le \infty$ we have

$$\begin{aligned} |x|^{2\alpha}|r(x)| \le C[(| \cdot |^{2\alpha} e^{-\sqrt{g'_{2c\varepsilon}(0)}|\cdot|}) * ((1+r)u^2 + h(r)r^2)(x) \\ + e^{-\sqrt{g'_{2c\varepsilon}(0)}|\cdot|} * ((1+r)u^2 | \cdot |^{2\alpha} + h(r)(| \cdot |^\alpha r)^2)](x), \end{aligned} \tag{3.19}$$

and we infer that $| \cdot |^{2\alpha} r \in L^p(\mathbf{R})$ for $1 \le p \le \infty$.

We have already proved that $|x|^{\frac{1}{2}} r(x) \in L^2(\mathbf{R})$, so it follows easily by induction that $|x|^\sigma r(x) \in L^p(\mathbf{R})$ for any $\sigma > 0$ and $1 \le p \le \infty$. Since $W(x) \le K|x|^s$ for some $K, s > 0$, we infer that $(1+r)u^2 + h(r)r^2 \in \mathbf{L}_W$. Now it follows from (3.17) and Lemma 3.4 that $r \in \mathbf{H}_W$, and Lemma 3.5 is proved. □

Now we turn our attention to the operators A, B, H_1, and H_2 appearing in (3.10).

LEMMA 3.6. *We have the following:*

(i) *For any* $\lambda \in (-\infty, q^2)$, $A_\lambda : \mathbf{H}_W \longrightarrow \mathbf{H}_W$ *is linear, compact and the mapping* $(\lambda, u) \longmapsto A_\lambda(u)$ *is continuous from* $(-\infty, q^2) \times \mathbf{H}_W$ *to* $\mathbf{H}_W$ *and compact on closed bounded subsets of* $[d, e] \times \mathbf{H}_W$ *for* $-\infty < d < e < q^2$.

(ii) *The linear operator* $B : \mathbf{H}_W \longrightarrow \mathbf{H}_W$ *is compact.*

(iii) $H_1 : ((V - r_{2c\varepsilon}) \cap \mathbf{H}_W) \times \mathbf{H}_W \longrightarrow \mathbf{H}_W$ *is continuous, compact on closed bounded subsets* ω_1 *of* $((V - r_{2c\varepsilon}) \cap \mathbf{H}_W) \times \mathbf{H}_W$ *such that* $dist(\omega_1, (\mathbf{H}_W \setminus (V - r_{2c\varepsilon})) \times \mathbf{H}_W) > 0$ *and*

$$||H_1(w, u)||_{\mathbf{H}_W} \le C_{\omega_1}(||w||^2_{\mathbf{H}_W} + ||u||^2_{\mathbf{H}_W}). \tag{3.20}$$

(iv) $H_2 : (-\infty, q^2) \times \mathbf{H}_W \times \mathbf{H}_W \longrightarrow \mathbf{H}_W$ *is continuous, compact on closed bounded subsets of* $[d, e] \times \mathbf{H}_W \times \mathbf{H}_W$ *for* $-\infty < d < e < q^2$ *and*

$$||H_2(\lambda, w, u)||_{\mathbf{H}_W} \le C_{d,e}(||w||^2_{\mathbf{H}_W} + ||w||^4_{\mathbf{H}_W} + ||u||^2_{\mathbf{H}_W}) \qquad \text{for any } \lambda \in [d, e]. \tag{3.21}$$

Proof. It is easy to see that $u_n \rightharpoonup u_*$ in $\mathbf{H}_W$ and $v_n \rightharpoonup v_*$ in $\mathbf{H}_W$ imply that $u_n v_n \longrightarrow u_* v_*$ in $\mathbf{L}_W$. Indeed, (u_n) and (v_n) are bounded in $\mathbf{H}_W$, and by Lemma 3.3 we have

$$||u_n v_n - u_* v_*||_{\mathbf{L}_W} \le ||v_n - v_*||_{L^\infty} ||u_n||_{\mathbf{L}_W} + ||u_n - u_*||_{L^\infty} ||v_*||_{\mathbf{L}_W} \longrightarrow 0 \quad \text{as } n \longrightarrow \infty. \tag{3.22}$$

(i) It is now clear that $u \longmapsto (r^2_{2c\varepsilon} + 2r^2_{2c\varepsilon})u$ is a linear compact mapping from $\mathbf{H}_W$ to $\mathbf{L}_W$, and we get (i) by using Lemma 3.4 and the resolvent formula

$$\begin{aligned} &\Big(-\frac{d^2}{dx^2} + q^2 - \lambda_1\Big)^{-1} - \Big(-\frac{d^2}{dx^2} + q^2 - \lambda_2\Big)^{-1} \\ &\quad = (\lambda_1 - \lambda_2)\Big(-\frac{d^2}{dx^2} + q^2 - \lambda_1\Big)^{-1}\Big(-\frac{d^2}{dx^2} + q^2 - \lambda_2\Big)^{-1}. \end{aligned}$$

(ii) This is obvious.

(iii) Let ω_1 be as in Lemma 3.6. We claim that there exists $\eta > 0$ such that for any $(w,u) \in \omega_1$ we have $\inf_{x\in\mathbf{R}}(w(x)+r_{2c\varepsilon}(x)) \geq -1+\eta$. We argue by contradiction and suppose that there exists a sequence $(w_n,u_n) \in \omega_1$ such that $a_n := \inf_{x\in\mathbf{R}}(w_n(x) + r_{2c\varepsilon}(x)) = (w_n + r_{2c\varepsilon})(x_n)$ tends to -1. The sequence (w_n) is bounded in $\mathbf{H}_W$; hence we may assume (passing to a subsequence if necessary) that $w_n \rightharpoonup w_*$ in $\mathbf{H}_W$. By Lemma 3.3, $w_n + r_{2c\varepsilon} \longrightarrow w_* + r_{2c\varepsilon}$ in $C^1_b(\mathbf{R})$. Since $w_*(x) + r_{2c\varepsilon}(x) \longrightarrow 0$ as $x \longrightarrow \infty$, the sequence (x_n) is bounded, say, $x_n \in [-M,M]$. Take $\chi \in C_0^\infty(\mathbf{R})$ such that $\mathrm{supp}(\chi) \subset [-M-1,M+1]$ and $\chi \equiv 1$ on $[-M,M]$. Then $\inf_{x\in\mathbf{R}}(w_n(x)+r_{2c\varepsilon}(x)-(a_n+1)\chi(x)) = w_n(x_n)+r_{2c\varepsilon}(x_n)-(a_n+1)\chi(x_n) = -1$, so that $w_n+r_{2c\varepsilon}-(a_n+1)\chi \notin V$ and

$$dist(w_n, \mathbf{H}_W \setminus (V - r_{2c\varepsilon})) \leq dist(w_n, w_n - (a_n+1)\chi) = |1+a_n|\ ||\chi||_{\mathbf{H}_W} \longrightarrow 0$$

as $n \longrightarrow \infty$, contradicting the fact that $(w_n,u_n) \in \omega_1$. This proves the claim.

For a given $w \in V - r_{2c\varepsilon}$, we have

$$\begin{aligned}
&(g_{2c\varepsilon}(r_{2c\varepsilon}+w) - g_{2c\varepsilon}(r_{2c\varepsilon}) - g'_{2c\varepsilon}(r_{2c\varepsilon})w)(x)\\
&\quad = \int_0^1 g'_{2c\varepsilon}(r_{2c\varepsilon}+tw)w(x)dt - g'_{2c\varepsilon}(r_{2c\varepsilon})w(x)\\
&\quad = w^2(x)\int_0^1\int_0^1 g''_{2c\varepsilon}(r_{2c\varepsilon}+tsw)(x)ds\ t\ dt = w^2(x)h_1(w)(x),
\end{aligned}$$

where $h_1(w)(x) = \int_0^1\int_0^1 g''_{2c\varepsilon}(r_{2c\varepsilon}+tsw)(x)ds\ t\ dt$.

To prove (iii) it suffices to show that for any sequence $(w_n,u_n) \in \omega_1$ such that $w_n \rightharpoonup w_*$ and $u_n \rightharpoonup u_*$ in $\mathbf{H}_W$, we have $H_1(w_n,u_n) \longrightarrow H_1(w_*,u_*)$ in $\mathbf{H}_W$. In view of Lemma 3.4, it suffices to show that

(3.23)
$$h_1(w_n)w_n^2 + (1+r_{2c\varepsilon}+w_n)u_n^2 \longrightarrow h_1(w_*)w_*^2 + (1+r_{2c\varepsilon}+w_*)u_*^2 \qquad \text{in } \mathbf{L}_W.$$

The sequence (w_n) being bounded in $\mathbf{H}_W$, there exists $K > 0$ such that $-1 + \min(\eta, \sqrt{2}c\varepsilon) \leq r_{2c\varepsilon}(x) + stw_n(x) \leq K$ for any $x \in \mathbf{R}$, $n \in \mathbf{N}$, and $s, t \in [0,1]$. Since $g''_{2c\varepsilon}$ is uniformly continuous on $[-1+\min(\eta,\sqrt{2}c\varepsilon), K]$, it is standard to prove that $h_1(w_n) \longrightarrow h_1(w_*)$ in $L^\infty(\mathbf{R})$, and then (3.23) follows from (3.22). Finally, using Lemma 3.4 we have for any $(w,u) \in \omega_1$

$$||H_1(w,u)||_{\mathbf{H}_W} \leq C||h_1(w)w^2 + (1+r_{2c\varepsilon}+w)u^2||_{\mathbf{L}_W} \leq C_{\omega_1}(||w||^2_{\mathbf{H}_W} + ||u||^2_{\mathbf{H}_W}).$$

(iv) From the preceding arguments it is easy to see that the mapping $(w,u) \longmapsto (w^2 + 2wr_{2c\varepsilon} + 2w)u$ is continuous from $\mathbf{H}_W \times \mathbf{H}_W$ to $\mathbf{L}_W$ and the image of any bounded set in $\mathbf{H}_W \times \mathbf{H}_W$ is precompact in $\mathbf{L}_W$, so (iv) follows from Lemma 3.4 and the resolvent formula above. The estimate (3.21) is straightforward. ∎

LEMMA 3.7. *For any $\lambda < q^2$ we have the following:*

(i) $\mathrm{Ker}(Id_{\mathbf{H}_W} + A_\lambda) \neq \{0\}$ *if and only if λ is an eigenvalue of the operator $A = -\frac{d^2}{dx^2} + q^2(1+r_{2c\varepsilon})^2$. In this case we have $\mathrm{Ker}(Id_{\mathbf{H}_W} + A_\lambda)^n = Span\{u_\lambda\}$ for any $n \in \mathbf{N}^*$.*

(ii) *If λ is not an eigenvalue of A, then $ind(Id_{\mathbf{H}_W} + A_\lambda, 0) = (-1)^{n(\lambda)}$ (where $n(\lambda)$ is the number of eigenvalues of A less than λ).*

Proof. (i) It is easy to see that $u \in \mathbf{L}$ and $u + A_\lambda u = 0$ is equivalent to $u \in \mathbf{H}$ and $Au = \lambda u$. Recall that if $\lambda < q^2$ is an eigenvalue of A in $\mathbf{L}$, then the corresponding eigenvector u_λ is in $\mathbf{H}_W$ by Corollary 2.2(iii). Consequently, we have $\mathrm{Ker}(Id_{\mathbf{H}_W} + A_\lambda) = \mathrm{Ker}(Id_{\mathbf{L}} + A_\lambda) = \mathrm{Ker}(\lambda Id_{\mathbf{H}} - A) = Span\{u_\lambda\}$.

To prove (i), it suffices to show that $u_\lambda \notin \mathrm{Im}(Id_{\mathbf{L}} + A_\lambda)$. Suppose by contradiction that there exists $v \in \mathbf{L}$ such that $v + A_\lambda v = u_\lambda$. This is equivalent to $v \in \mathbf{H}$ and $Av - \lambda v = -u''_\lambda + (q^2-\lambda)u_\lambda$, that is, $-u''_\lambda + (q^2-\lambda)u_\lambda \in \mathrm{Im}(A-\lambda)$. Since $A-\lambda$ is self-adjoint on $\mathbf{L}$, $-u''_\lambda + (q^2-\lambda)u_\lambda$ must be orthogonal (in $\mathbf{L}$) to $\mathrm{Ker}(A-\lambda) = Span\{u_\lambda\}$, which gives $\int_{\mathbf{R}} |u'_\lambda|^2 dx + (q^2-\lambda)\int_{\mathbf{R}} |u_\lambda|^2 dx = 0$, a contradiction.

(ii) A well-known result of Leray and Schauder asserts that if K is a compact operator on a real Banach space X and 1 is not an eigenvalue of K, then

$$ind(Id - K, 0) = (-1)^\beta,$$

where β is the sum of all the (algebraic) multiplicities of eigenvalues of K greater than 1 (see, e.g., [6, Theorem 4.6, p. 133]).

Thus, for a given λ which is not an eigenvalue of A, we are interested in the eigenvalues $\mu > 1$ of $-A_\lambda$. Clearly, $-A_\lambda u = \mu u$ is equivalent to

$$q^2\left(-\frac{d^2}{dx^2} + q^2 - \lambda\right)^{-1}((r_{2c\varepsilon}^2 + 2r_{2c\varepsilon})u) + \mu u = 0,$$

that is,

$$-u'' + q^2(1+r_{2c\varepsilon})^2 u + q^2\left(1 - \frac{1}{\mu}\right)[1-(1+r_{2c\varepsilon})^2]u = \lambda u.$$

In other words, $\mu > 1$ is an eigenvalue of $-A_\lambda$ if and only if λ is an eigenvalue of the operator

$$\begin{aligned} M_\mu &= -\frac{d^2}{dx^2} + q^2(1+r_{2c\varepsilon})^2 + q^2\left(1-\frac{1}{\mu}\right)[1-(1+r_{2c\varepsilon})^2] \\ &= A + q^2\left(1-\frac{1}{\mu}\right)[1-(1+r_{2c\varepsilon})^2]. \end{aligned}$$

Note that $M_\mu \geq A$ for any $\mu \geq 1$ and $\sigma_{ess}(M_\mu) = [q^2, \infty)$ by Weyl's theorem. By Proposition 2.1(iv), $\lambda \in (-\infty, q^2)$ is an eigenvalue of M_μ considered as an operator on $\mathbf{L}_W$ if and only if λ is an eigenvalue of M_μ considered as an operator on $\mathbf{L}$. We will work on $\mathbf{L}$ because on this space M_μ is self-adjoint.

If $\lambda < q^2$ is not an eigenvalue of A, we will prove that there are exactly $n(\lambda)$ values $\mu \in (1,\infty)$ such that λ is an eigenvalue of M_μ.

For $\mu \in [1,\infty)$, we define

$$\alpha_n(\mu) = \sup_{\varphi_1,\dots,\varphi_{n-1}\in\mathbf{H}} \inf_{\psi\in\{\varphi_1,\dots,\varphi_{n-1}\}^\perp} \frac{\langle M_\mu\psi,\psi\rangle_{\mathbf{L}}}{||\psi||^2_{\mathbf{L}}}. \tag{3.24}$$

By the min-max principle ([13, Theorem XIII.1, p. 76]), either $\alpha_n(\mu)$ is the nth eigenvalue of M_μ (counted with multiplicity) or $\alpha_n(\mu) = q^2$. By Proposition 2.1(iii), the eigenvalues of M_μ are simple, thus we have $\alpha_p(\mu) < \alpha_n(\mu)$ if $p < n$ and $\alpha_p(\mu) < q^2$.

It is obvious that the functions $\mu \longmapsto \alpha_n(\mu)$ are increasing on $[1,\infty)$ because $M_{\mu_1} \leq M_{\mu_2}$ if $1 \leq \mu_1 < \mu_2$. In fact, α_n is strictly increasing on $\{\mu \in [1,\infty) \mid \alpha_n(\mu) < q^2\}$. To see this, consider $\mu_1 < \mu_2$ such that $\alpha_n(\mu_2) < q^2$. Then $\alpha_1(\mu_2), \dots, \alpha_n(\mu_2)$ are eigenvalues of M_{μ_2}. Let $u_1, \dots, u_n \in \mathbf{H}$ be corresponding eigenvectors with $||u_i||_{\mathbf{L}} = 1$.

Clearly, $u_1, \dots, u_n$ are mutually orthogonal in $\mathbf{L}$ and it is easily seen from the definition of M_μ that $\langle M_{\mu_1} u_i, u_i \rangle_{\mathbf{L}} < \langle M_{\mu_2} u_i, u_i \rangle_{\mathbf{L}} = \alpha_i(\mu_2)$, $i = 1, \dots, n$. Note that the quantity $N(u) = (\int_{\mathbf{R}} [1 - (1 + r_{2c\varepsilon})^2] |u|^2 dx)^{\frac{1}{2}}$ is a norm on $\mathbf{L}$. Since $Span\{u_1, \dots, u_n\}$ is finite dimensional, there exists $N_1 > 0$ such that $N(u) \geq N_1 ||u||_{\mathbf{L}}$ for any $u \in Span\{u_1, \dots, u_n\}$. Therefore

(3.25)
$$\begin{aligned}
&\left\langle M_{\mu_1} \left(\sum_{i=1}^{n} a_i u_i \right), \left(\sum_{i=1}^{n} a_i u_i \right) \right\rangle_{\mathbf{L}} \\
&= \left\langle M_{\mu_2} \left(\sum_{i=1}^{n} a_i u_i \right), \left(\sum_{i=1}^{n} a_i u_i \right) \right\rangle_{\mathbf{L}} - \left\langle (M_{\mu_2} - M_{\mu_1}) \left(\sum_{i=1}^{n} a_i u_i \right), \left(\sum_{i=1}^{n} a_i u_i \right) \right\rangle_{\mathbf{L}} \\
&= \sum_{i=1}^{n} \alpha_i(\mu_2) |a_i|^2 - q^2 \left(\frac{1}{\mu_1} - \frac{1}{\mu_2} \right) \int_{\mathbf{R}} [1 - (1 + r_{2c\varepsilon})^2] \left| \sum_{i=1}^{n} a_i u_i \right|^2 dx \\
&\leq \alpha_n(\mu_2) \left\| \sum_{i=1}^{n} a_i u_i \right\|_{\mathbf{L}}^2 - q^2 \left(\frac{1}{\mu_1} - \frac{1}{\mu_2} \right) N_1^2 \left\| \sum_{i=1}^{n} a_i u_i \right\|_{\mathbf{L}}^2 .
\end{aligned}$$

Thus for any u in the n-dimensional subspace $Span\{u_1, \dots, u_n\}$ we have

$$\langle M_{\mu_1} u, u \rangle_{\mathbf{L}} \leq \left(\alpha_n(\mu_2) - q^2 \left(\frac{1}{\mu_1} - \frac{1}{\mu_2} \right) N_1^2 \right) ||u||_{\mathbf{L}}^2 .$$

By the min-max principle it follows that $\alpha_n(\mu_1) \leq \alpha_n(\mu_2) - q^2 (\frac{1}{\mu_1} - \frac{1}{\mu_2}) N_1^2$.

A standard argument shows that each α_n is continuous. Indeed, suppose by contradiction that $\mu_* \in (1, \infty)$ is a discontinuity point. Then necessarily $l_1 := \sup_{\mu < \mu_*} \alpha_n(\mu) < \inf_{\mu > \mu_*} \alpha_n(\mu) := l_2$. Take $0 < \epsilon < \frac{l_2 - l_1}{4}$ and $\mu_1 < \mu_*$, $\mu_2 > \mu_*$ such that $q^2 (\frac{1}{\mu_1} - \frac{1}{\mu_2}) < \epsilon$. Since $\alpha_n(\mu_2) > l_2 - \epsilon$, there exist $\varphi_1, \dots, \varphi_{n-1} \in \mathbf{H}$ such that $\langle M_{\mu_2} \psi, \psi \rangle_{\mathbf{L}} > l_2 - \epsilon$ for any $\psi \in \{\varphi_1, \dots, \varphi_{n-1}\}^\perp$ with $||\psi||_{\mathbf{L}} = 1$. We have

$$\begin{aligned}
&\langle M_{\mu_2} \psi, \psi \rangle_{\mathbf{L}} - \langle M_{\mu_1} \psi, \psi \rangle_{\mathbf{L}} \\
&= q^2 \left(\frac{1}{\mu_1} - \frac{1}{\mu_2} \right) \int_{\mathbf{R}} [1 - (1 + r_{2c\varepsilon})^2] |\psi|^2 dx \leq q^2 \left(\frac{1}{\mu_1} - \frac{1}{\mu_2} \right) ||\psi||_{\mathbf{L}}^2 < \epsilon;
\end{aligned}$$

thus $\langle M_{\mu_1} \psi, \psi \rangle_{\mathbf{L}} > l_2 - 2\epsilon$ for any $\psi \in \{\varphi_1, \dots, \varphi_{n-1}\}^\perp$ with $||\psi||_{\mathbf{L}} = 1$. Therefore $\alpha_n(\mu_1) > l_2 - 2\epsilon$, which is a contradiction.

We also have for any $u \in \mathbf{H}$

$$\langle M_\mu u, u \rangle_{\mathbf{L}} = ||u'||_{L^2}^2 + q^2 ||u||_{\mathbf{L}}^2 - \frac{q^2}{\mu} \int_{\mathbf{R}} [1 - (1 + r_{2c\varepsilon})^2] |u|^2 dx \geq q^2 ||u||_{\mathbf{L}}^2 - \frac{C}{\mu} ||u||_{\mathbf{L}}^2 ;$$

hence $\alpha_1(\mu) \geq q^2 - \frac{C}{\mu} \longrightarrow q^2$ as $\mu \longrightarrow \infty$. Consequently, $\alpha_n(\mu) \longrightarrow q^2$ as $\mu \longrightarrow \infty$ for any $n \geq 1$.

Note that $\lambda < q^2$ is an eigenvalue of M_μ if and only if $\lambda = \alpha_n(\mu)$ for some $n \in \mathbf{N}^*$. We know that there are exactly $n(\lambda)$ eigenvalues of A less than λ, say, $\lambda_1 < \lambda_2 < \dots < \lambda_{n(\lambda)} < \lambda$. We have $\alpha_i(1) = \lambda_i$ because $M_1 = A$, and the functions α_i are strictly increasing (until they reach the value q^2, if this happens) and continuous, and they tend to q^2 at infinity. We infer that for each $i \in \{1, \dots, n(\lambda)\}$, there exists exactly one value μ_i such that $\alpha_i(\mu_i) = \lambda$. Moreover, $\mu_1 > \mu_2 > \cdots > \mu_{n(\lambda)} > 1$. For any $n > n(\lambda)$, we have $\alpha_n(1) > \lambda$, hence $\alpha_n(\mu) > \lambda$ for $\mu \in (0, \infty)$ because α_n is increasing.

Thus we have shown that the operator $-A_\lambda$ has exactly $n(\lambda)$ eigenvalues greater than 1, $\mu_1 > \mu_2 > \cdots > \mu_{n(\lambda)}$. Moreover, $\mathrm{Ker}(\mu_i + A_\lambda) = \mathrm{Ker}(M_{\mu_i} - \lambda)$. We know by Proposition 2.1(iii) that $\mathrm{Ker}(M_{\mu_i} - \lambda)$ is one dimensional. If this kernel is spanned by a function v_i, then $v_i \notin \mathrm{Im}(\mu_i + A_\lambda)$. Indeed, $\mu_i u + A_\lambda u = v_i$ would imply $(M_{\mu_i} - \lambda)u = \frac{1}{\mu_i}(-v_i'' + (q^2 - \lambda)v_i)$. Since M is self-adjoint, $-v_i'' + (q^2 - \lambda)v_i$ would be orthogonal to $\mathrm{Ker}(M_{\mu_i} - \lambda) = Span\{v_i\}$, which gives a contradiction. Consequently, we have $\mathrm{Ker}(\mu_i + A_\lambda)^n = Span\{v_i\}$ for any $n \in \mathbf{N}^*$; that is, μ_i is a simple eigenvalue of $-A_\lambda$.

As a consequence, we have $ind(Id_{\mathbf{H}_W} + A_\lambda, 0) = (-1)^{n(\lambda)}$ and Lemma 3.7 is proved. □

We are now in position to state the main result of this paper.

THEOREM 3.8. *Let $\mathcal{S}$ be the set of nontrivial solutions of the system* (1.9)–(1.10) *in $\mathbf{R} \times (\mathbf{H} \cap V) \times \mathbf{H}$. For any eigenvalue $\lambda_m < q^2$ of $A = -\frac{d^2}{dx^2} + (1 + r_{2c\varepsilon})^2$, the set $\mathcal{S} \cup \{(\lambda_m, r_{2c\varepsilon}, 0)\}$ contains a maximal closed connected subset $\mathcal{C}_m$ in $(-\infty, q^2) \times \mathbf{H}_W \times \mathbf{H}_W$ such that $\mathcal{C}_m \cap \mathcal{C}_p = \emptyset$ if $m \neq p$ and $\mathcal{C}_m$ satisfies at least one of the two following properties:*

(i) *$\mathcal{C}_m$ is unbounded in $\mathbf{R} \times \mathbf{H}_W \times \mathbf{H}_W$.*

(ii) *There exists a sequence $(\lambda_n, r_n, u_n) \in \mathcal{C}_m$ such that $\lambda_n \longrightarrow q^2$ as $n \longrightarrow \infty$.*

Proof. We have already seen that $(\lambda, r, u) \in (-\infty, q^2) \times (\mathbf{H} \cap V) \times \mathbf{H}$ is a nontrivial solution of (1.9)–(1.10) if and only if $(\lambda, r - r_{2c\varepsilon}, u)$ belongs to $(-\infty, q^2) \times (\mathbf{H}_W \cap (V - r_{2c\varepsilon})) \times \mathbf{H}_W$ and satisfies the system (3.8)–(3.9) (or, equivalently, (3.10)).

Let $E = \mathbf{H}_W \times \mathbf{H}_W$, $\Omega = (-\infty, q^2) \times (\mathbf{H}_W \cap (V - r_{2c\varepsilon})) \times \mathbf{H}_W$, $L_\lambda = \begin{pmatrix} -B & 0 \\ 0 & -A_\lambda \end{pmatrix}$, and $H(\lambda, w, u) = \begin{pmatrix} -H_1(w,u) \\ -H_2(\lambda,w,u) \end{pmatrix}$. Let $G(\lambda, w, u) = L_\lambda(w, u) + H(\lambda, w, u)$. It is obvious that on Ω, (3.10) is equivalent to the equation $(w, u) = G(\lambda, w, u)$. It follows easily from Lemma 3.6 that L and H satisfy assumptions (a) and (b) in Proposition 3.2.

We claim that $Id_{\mathbf{H}_W} + B : \mathbf{H}_W \longrightarrow \mathbf{H}_W$ is invertible. Indeed, $(Id_{\mathbf{H}_W} + B)u = v$ is equivalent to $-u'' + g_{2c\varepsilon}'(r_{2c\varepsilon})u = (-\frac{d^2}{dx^2} + g_{2c\varepsilon}'(0))v$. By Lemma 2.4, there exists a unique $u \in \mathbf{H}$ satisfying this equation. We have

$$-u'' + g_{2c\varepsilon}'(0)u = \left(-\frac{d^2}{dx^2} + g_{2c\varepsilon}'(0)\right)v + (g_{2c\varepsilon}'(0) - g_{2c\varepsilon}'(r_{2c\varepsilon}))u \in \mathbf{L}_W$$

(recall that $v \in \mathbf{H}_W$ and $g_{2c\varepsilon}'(0) - g_{2c\varepsilon}'(r_{2c\varepsilon})$ decays exponentially at infinity). Using Lemma 3.4, we infer that $u \in \mathbf{H}_W$.

For $\lambda < q^2$, it is clear that $Id_{\mathbf{H}_W \times \mathbf{H}_W} - L_\lambda$ is not invertible if and only if $Id_{\mathbf{H}_W} + A_\lambda$ is not invertible, i.e., if and only if λ is an eigenvalue of A. Let $\lambda_1 < \lambda_2 < \cdots < \lambda_{N_q} < q^2$ be the eigenvalues of A below q^2. If λ is not an eigenvalue of A, we infer using Lemma 3.7 that $i(\lambda) := ind(Id_{\mathbf{H}_W \times \mathbf{H}_W} - L_\lambda, 0) = ind(Id_{\mathbf{H}_W} + A_\lambda, 0) \cdot ind(Id_{\mathbf{H}_W} + B, 0) = (-1)^{n(\lambda)} ind(Id_{\mathbf{H}_W} + B, 0)$ is constant on each of the intervals $(-\infty, \lambda_1), (\lambda_i, \lambda_{i+1}), (\lambda_{N_q}, q^2)$ and changes sign at each λ_i. Consequently, L_λ also satisfies assumption (c) in Proposition 3.2 at any point $(\lambda_i, 0, 0)$. Let $\tilde{\mathcal{S}}_0 = \{(\lambda, w, u) \in \Omega \mid (w, u) \neq (0, 0)$ and (λ, w, u) satisfies (3.10)$\}$ and let $\tilde{\mathcal{S}} = \tilde{\mathcal{S}}_0 \setminus \{(\lambda, -r_{2c\varepsilon}, 0) \mid \lambda \in (-\infty, q^2)\}$. Note that the solutions $(\lambda, -r_{2c\varepsilon}, 0)$ of (3.10) correspond to the solutions $(\lambda, 0, 0)$ of (1.9)–(1.10) and $\mathcal{S} \cap ((-\infty, q^2) \times (V \cap \mathbf{H}_W) \times \mathbf{H}_W) = \tilde{\mathcal{S}} + (0, r_{2c\varepsilon}, 0)$. We may apply Proposition 3.2 to infer that for any $1 \leq m \leq N_q$, there exists a maximal closed connected subset $\mathcal{D}_m$ (in Ω) of $\tilde{\mathcal{S}}_0 \cup \{(\lambda_m, 0, 0)\}$ which contains $(\lambda_m, 0, 0)$ and satisfies at least one of the following properties:

1. $\mathcal{D}_m$ is unbounded.
2. There exists a sequence $(\lambda_n, w_n, u_n) \in \mathcal{D}_m$ such that $\lambda_n \longrightarrow q^2$ as $n \longrightarrow \infty$.
3. There exists a sequence $(\lambda_n, w_n, u_n) \in \mathcal{D}_m$ such that $dist(w_n, \partial((V - r_{2c\varepsilon}) \cap \mathbf{H}_W)) \longrightarrow 0$, that is, $\inf_{x\in\mathbf{R}}(w_n(x) + r_{2c\varepsilon}(x)) \longrightarrow -1$ as $n \longrightarrow \infty$.
4. The closure in Ω of $\mathcal{D}_m$ contains a point $(\lambda_i, 0, 0)$ with $i \neq m$.

Let us show first that $\mathcal{D}_m$ cannot meet $\{(\lambda, -r_{2c\varepsilon}, 0) \mid \lambda \in (-\infty, q^2)\}$. A straightforward computation gives $d_{(w,u)}(Id_E - G)(\lambda, -r_{2c\varepsilon}, 0) = Id_E$ for any $\lambda < q^2$. By the implicit function theorem, there exists a neighborhood N_λ of $(\lambda, -r_{2c\varepsilon}, 0)$ in $\mathbf{R} \times E$ such that the only solutions of the equation $(w, u) = G(\lambda, w, u)$ in N_λ are $(\mu, -r_{2c\varepsilon}, 0)$. Hence $\cup_\lambda N_\lambda$ is a neighborhood of $\{(\lambda, -r_{2c\varepsilon}, 0) \mid \lambda < q^2\}$ in Ω which contains no other solutions of (3.10). Consequently, we have $\mathcal{D}_m \subset \tilde{\mathcal{S}}$.

By Proposition 3.1, for any $(\lambda, w, u) \in \tilde{\mathcal{S}}_0$ we have $\inf_{x\in\mathbf{R}}(w(x) + r_{2c\varepsilon}(x)) > -1 + \sqrt{2}c\varepsilon$; hence $\mathcal{D}_m$ cannot satisfy property 3.

We will also eliminate the alternative 4. Observe that if $(\lambda, r, u) \in (-\infty, q^2) \times \mathbf{H} \times \mathbf{H}$ is a nontrivial solution of (1.9)–(1.10), then, in particular, u is an eigenvector of the linear operator $-\frac{d^2}{dx^2} + q^2(1 + r)^2$ corresponding to the eigenvalue λ. It is easily checked that this operator is a compact perturbation of $-\frac{d^2}{dx^2} + q^2$, so it has the essential spectrum $[q^2, \infty)$. Since $\lambda < q^2$, the operator $-\frac{d^2}{dx^2} + q^2(1 + r)^2$ has only a finite number of eigenvalues less than λ, say p. We define $z(\lambda, r, u) = p$. By Proposition 2.1(v), we know that u has exactly p zeroes in $(0, \infty)$. We also define $z(\lambda_i, r_{2c\varepsilon}, 0) = i - 1$. We have the following lemma.

LEMMA 3.9. *The function z is continuous on* $(\mathcal{S} \cup \{(\lambda_i, r_{2c\varepsilon}, 0) \mid i = 1, \ldots, N_q\}) \cap ((-\infty, q^2) \times \mathbf{H} \times \mathbf{H})$.

Assume for the moment that Lemma 3.9 holds. Obviously, the function z is also continuous for the $\mathbf{R} \times E$ topology. Since z takes values in $\mathbf{N}$, it must be constant on each connected component of $(\mathcal{S} \cup \{(\lambda_i, r_{2c\varepsilon}, 0) \mid i = 1, \ldots, N_q\}) \cap ((-\infty, q^2) \times \mathbf{H} \times \mathbf{H}) = (\tilde{\mathcal{S}} + (0, r_{2c\varepsilon}, 0)) \cup \{(\lambda_i, r_{2c\varepsilon}, 0) \mid i = 1, \ldots, N_q\}$. In particular, it is constant on $\mathcal{D}_m + (0, r_{2c\varepsilon}, 0)$ and we find $z(\mathcal{D}_m + (0, r_{2c\varepsilon}, 0)) = z(\lambda_m, r_{2c\varepsilon}, 0) = m - 1$. We have also $z(\mathcal{D}_i + (0, r_{2c\varepsilon}, 0)) = i - 1$; hence $\mathcal{D}_m$ and $\mathcal{D}_i$ are disjoint if $i \neq m$ (in fact, we see that the closures of $\mathcal{D}_m$ and $\mathcal{D}_i$ in $(-\infty, q^2) \times \mathbf{H} \times \mathbf{H}$ are disjoint if $i \neq m$). Thus $\mathcal{D}_m$ cannot satisfy alternative 4 above; hence it necessarily satisfies one of alternatives 1 or 2. Let $\mathcal{C}_m = \mathcal{D}_m + (0, r_{2c\varepsilon}, 0)$. It is now clear that $\mathcal{C}_m$ satisfies (i) or (ii) in Theorem 3.8. $\square$

Proof of Lemma 3.9. Let $(\lambda, r, u), (\nu_n, r_n, u_n) \in (\mathcal{S} \cup \{(\lambda_i, r_{2c\varepsilon}, 0) \mid i = 1, \ldots, N_q\}) \cap ((-\infty, q^2) \times \mathbf{H} \times \mathbf{H})$ be such that $z(\lambda, r, u) = p$ and $(\nu_n, r_n, u_n) \longrightarrow (\lambda, r, u)$ as $n \longrightarrow \infty$. Let $\mu_1 < \mu_2 < \cdots < \mu_{p+1} = \lambda$ be the eigenvalues of the operator $B = -\frac{d^2}{dx^2} + q^2(1 + r)^2$ in $\mathbf{L}$ and let $u_1^*, \ldots, u_{p+1}^* = u$ be corresponding eigenvectors. Denote $B_n = -\frac{d^2}{dx^2} + q^2(1 + r_n)^2$.

We prove that $z(\nu_n, r_n, u_n) \geq p$ if n is sufficiently big. There is nothing to do if $p = 0$. Suppose that $p \geq 1$. Take $0 < \epsilon < \frac{\mu_{p+1} - \mu_p}{4}$ and let n_0 be sufficiently big, so that $||(r_n - r)(2 + r_n + r)||_{L^\infty} < \frac{\epsilon}{q^2}$ and $\lambda - \epsilon < \nu_n < \lambda + \epsilon$ for any $n \geq n_0$. For $n \geq n_0$ and $v \in Span\{u_1^*, \ldots u_p^*\}$ we have

$$\langle B_n v, v\rangle_{\mathbf{L}} = \langle Bv, v\rangle_{\mathbf{L}} + \langle (B_n - B)v, v\rangle_{\mathbf{L}}$$
$$\leq \mu_p ||v||_{\mathbf{L}}^2 + q^2 \int_{\mathbf{R}} (r_n - r)(2 + r_n + r)|v|^2 dx \leq (\mu_p + \epsilon)||v||_{\mathbf{L}}^2 < (\nu_n - \epsilon)||v||_{\mathbf{L}}^2.$$

By the min-max principle, B_n has at least p eigenvalues less than or equal to $\nu_n - \epsilon$, and thus $z(\nu_n, r_n, u_n) \geq p$.

Let $\mu_{p+2} = \sup_{\varphi_1,\ldots,\varphi_{p+1}\in\mathbf{H}} \inf_{\psi\in\{\varphi_1,\ldots,\varphi_{p+1}\}^\perp} \frac{\langle B\psi,\psi\rangle_\mathbf{L}}{||\psi||^2_\mathbf{L}}$. Since $\lambda = \mu_{p+1} < q^2$ and λ is a simple eigenvalue of B by Proposition 2.1(iii), we know by the min-max principle that either $\mu_{p+2} = q^2$ or μ_{p+2} is an eigenvalue of B and $\mu_{p+2} > \mu_{p+1}$. Let $\epsilon \in (0, \frac{\mu_{p+2}-\mu_{p+1}}{4})$. Take n_0 as above and $\varphi_1, \ldots, \varphi_{p+1} \in \mathbf{H}$ such that $\inf_{\psi\in\{\varphi_1,\ldots,\varphi_{p+1}\}^\perp} \frac{\langle B\psi,\psi\rangle_\mathbf{L}}{||\psi||^2_\mathbf{L}} \geq \mu_{p+2} - \epsilon$. For any $\psi \in \{\varphi_1, \ldots, \varphi_{p+1}\}^\perp$, $\psi \neq 0$ we have

$$\langle B_n\psi,\psi\rangle_\mathbf{L} = \langle B\psi,\psi\rangle_\mathbf{L} + \langle(B_n - B)\psi,\psi\rangle_\mathbf{L} \geq (\mu_{p+2} - \epsilon)||\psi||^2_\mathbf{L} - \epsilon||\psi||^2_\mathbf{L} \geq (\nu_n + \epsilon)||\psi||^2_\mathbf{L}.$$

It follows from the min-max principle that for $n \geq n_0$, either B_n has at most $p+1$ eigenvalues or the $(p+2)$th eigenvalue is greater than $\nu_n + \epsilon$. Since ν_n is an eigenvalue of B_n, there are at most p eigenvalues of B_n less than ν_n, hence $z(\nu_n, r_n, u_n) \leq p$ for any $n \geq n_0$. This finishes the proof of Lemma 3.9 and that of Theorem 3.8. ◻

We were not able to eliminate either of the alternatives in Theorem 3.8.

Up to now, we have proved the existence of branches of nontrivial *symmetric* solutions (λ, r, u) to the system (1.9)–(1.10). For any such solution, $(\tilde\psi, \tilde\varphi)$ is a travelling-wave of (1.1) for $\varepsilon^2(c^2\delta^2 + k^2) = \lambda$ and satisfies the boundary condition (1.2), where $\tilde\varphi(x) = \frac{1}{\varepsilon}u(\frac{x}{\varepsilon})e^{ic\delta x}$ and $\tilde\psi(x) = (1 + r(\frac{x}{\varepsilon}))e^{i\psi_0(x)}$ (with $\psi_0(x) = c\int_0^x [1 - \frac{1}{(1+r(\frac{s}{\varepsilon}))^2}]ds = c\varepsilon\int_0^{\frac{x}{\varepsilon}} \frac{2r(\tau)+r^2(\tau)}{(1+r(\tau))^2}d\tau$). Note also that $\tilde\psi(-x) = \overline{\tilde\psi(x)}$, $\tilde\varphi(-x) = \overline{\tilde\varphi(x)}$, $|\tilde\psi| > \sqrt{2}c\varepsilon$ by Proposition 2.1, and the phase ψ_0 of $\tilde\psi$ remains bounded because r decays at infinity faster than $|x|^\beta$ for any $\beta > 0$ (see the end of the proof of Lemma 3.5). Since $2c^2\varepsilon^2q^2 < \lambda \leq q^2$, we have bounds on the single-particle impurity energy: $c^2(2q^2 - \delta^2) < k^2 \leq \frac{q^2}{\varepsilon^2} - c^2\delta^2$.

Remark 3.10. It follows from Corollary 2.2(iv)–(v) that there is exactly one branch of travelling-waves bifurcating from the trivial solutions if $q \leq \frac{1}{\sqrt{2\ln 2}}$. The number of these branches is the same as the number of eigenvalues of A, so it tends to infinity as $q \longrightarrow \infty$.

It is natural to ask how the branches $\mathcal{C}_m$ given by Theorem 3.8 behave in $\mathbf{R}\times\mathbf{H}\times\mathbf{H}$. The topology of $\mathbf{H}_W$ being stronger than that of $\mathbf{H}$, any of the sets $\mathcal{C}_m$ is also connected in $\mathbf{R}\times\mathbf{H}\times\mathbf{H}$. Roughly speaking, either $\mathcal{C}_m$ approaches $\{q^2\}\times(\mathbf{H}\cap V)\times\mathbf{H}$, or $\mathcal{C}_m$ is unbounded in $\mathbf{R}\times\mathbf{H}\times\mathbf{H}$, or it remains bounded in $\mathbf{R}\times\mathbf{H}\times\mathbf{H}$ but the norm in $\mathbf{R}\times\mathbf{H}_W\times\mathbf{H}_W$ tends to infinity along $\mathcal{C}_m$, i.e., "there is some mass moving to infinity."

Remark 3.11. The importance of Theorem 2.3 is that it gives a precise description of $\mathcal{C}_m$ in a neighborhood of $(\lambda_m, r_{2c\varepsilon}, 0)$ in $\mathbf{R}\times\mathbf{H}\times\mathbf{H}$. Let $\mathcal{C}_m^+$ (respectively $\mathcal{C}_m^-$) be the maximal subcontinuum in $\mathbf{R}\times\mathbf{H}_W\times\mathbf{H}_W$ of $\mathcal{C}_m\setminus\{(\lambda(s), r_{2c\varepsilon}+sr(s), s(u_m+u(s))) \mid s \in (-\eta, 0)\}$, (respectively of $\mathcal{C}_m\setminus\{(\lambda(s), r_{2c\varepsilon}+sr(s), s(u_m+u(s))) \mid s \in (0, \eta)\}$), where the curve $s \longmapsto (\lambda(s), r(s), u(s))$ is given by Theorem 2.3. It can be proved by using a variant of a classical result of Rabinowitz (Theorem 1.40, p. 500 in [12]) that each of $\mathcal{C}_m^+$ and $\mathcal{C}_m^-$ satisfies (i) or (ii) in Theorem 3.8.

Remark 3.12. It is not hard to prove that in dimension $N = 1$, 2, or 3 the Cauchy problem for the system (1.1) is globally well posed in $(1+H^1(\mathbf{R}^N))\times H^1(\mathbf{R}^N)$. However, the dynamics associated to (1.1) and the asymptotic behavior of solutions are not yet understood.

Remark 3.13. The existence of solitary waves for (1.1) in dimension greater than 1 is an open problem. Even the existence of "trivial" solitary waves (i.e., solutions of the form $(\psi(x_1 - ct, x_2, \ldots, x_N), 0)$ is a difficult problem. Note that if $\varphi \equiv 0$, the system (1.1) reduces to the Gross–Pitaevskii equation

$$2i\frac{\partial\psi}{\partial t} = -\Delta\psi + (|\psi|^2 - 1)\psi, \quad |\psi| \longrightarrow 1 \text{ as } |x| \longrightarrow \infty.$$

The existence of travelling-waves moving with small speed for this equation was proved, for instance, in [2] (in dimension $N = 2$) and in [1], [3] (in dimension $N \geq 3$).

Acknowledgments. I am indebted to Professor Jean-Claude Saut, who brought the subject to my attention, for interesting and stimulating discussions. I am also very grateful to Professor Louis Jeanjean for constant help and encouragement and to Professor Patrick J. Rabier for useful remarks.

REFERENCES

[1] F. Bethuel, G. Orlandi, and D. Smets, *Vortex rings for the Gross-Pitaevskii equation*, J. Eur. Math. Soc. (JEMS), 6 (2004), pp. 17–94.

[2] F. Bethuel and J.-C. Saut, *Travelling waves for the Gross-Pitaevskii equation* I, Ann. Inst. H. Poincaré Phys. Théor., 70 (1999), pp. 147–238.

[3] D. Chiron, *Étude mathématique de modèles issus de la physique de la matière condensée*, Ph.D. thesis, Université de Paris VI, Paris, France, 2004.

[4] M. G. Crandall and P. H. Rabinowitz, *Bifurcation from simple eigenvalues*, J. Funct. Anal., 8 (1971), pp. 321–340.

[5] J. Grant and P. H. Roberts, *Motions in a Bose condensate* III. *The structure and effective masses of charged and uncharged impurities*, J. Phys. A: Math., Nucl. Gen., 7 (1974), pp. 260–279.

[6] M. A. Krasnoselskii, *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon Press, Oxford, UK, 1964.

[7] M. Mariş, *Stationary solutions to a nonlinear Schrödinger equation with potential in one dimension*, Proc. Roy. Soc. Edinburgh Sect. A, 133 (2003), pp. 409–437.

[8] K. McLeod, *Uniqueness of positive radial solutions of* $\Delta u + f(u) = 0$ *in* $\mathbf{R}^N$, II, Trans. Amer. Math. Soc., 339 (1993), pp. 495–505.

[9] J. Pejsachowicz and P. J. Rabier, *Degree theory for* C^1 *Fredholm mappings of index* 0, J. Anal. Math., 76 (1998), pp. 289–319.

[10] P. J. Rabier and C. A. Stuart, *Global bifurcation for quasilinear elliptic equations on* $\mathbf{R}^N$, Math. Z., 237 (2001), pp. 85–124.

[11] P. J. Rabier and C. A. Stuart, *Fredholm and properness properties of quasilinear elliptic operators on* $\mathbf{R}^N$, Math. Nachr., 231 (2001), pp. 129–168.

[12] P. H. Rabinowitz, *Some global results for nonlinear eigenvalue problems*, J. Funct. Anal., 7 (1971), pp. 487–513.

[13] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Vol. 4, Academic Press, New York, London, 1978.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1560–1588

-DIMENSIONAL ELLIPTIC INVARIANT TORI FOR THE PLANAR (+ 1)-BODY PROBLEM*

LUCA BIASCO†, LUIGI CHIERCHIA†, AND ENRICO VALDINOCI‡

Abstract. For any $N \geq 2$ we prove the existence of quasi-periodic orbits lying on N-dimensional invariant elliptic tori for the planetary planar $(N+1)$-body problem. For small planetary masses, such orbits are close to the limiting solutions given by the N planets revolving around the sun on planar circles. The eigenvalues of the linearized secular dynamics are also computed asymptotically. The proof is based on an appropriate averaging and KAM theory which overcomes the difficulties caused by the intrinsic degeneracies of the model. For concreteness, we focus on a caricature of the outer solar system.

Key words. N-body problem, nearly integrable Hamiltonian systems, lower-dimensional elliptic tori

AMS subject classifications. 70F10, 34C27, 37J40, 70K43

DOI. 10.1137/S0036141004443646

1. Introduction and results.

1.1. Quasi-periodic motions in the many-body problem. The existence of stable trajectories of the many-body problem viewed as a model for the solar system has been the subject of researches of many distinguished scientists both in the past and in recent years; see, for example, the theoretical work of Poincaré [Poi1905], Arnold [A63], Herman [H95], and the numerical investigations of Laskar [L96]. Only recently, a complete proof, based on [H95], of the existence of quasi-periodic motions (corresponding to maximal invariant tori of dimension $3N-1$) for the $(N+1)$-body problem for arbitrary N has been produced in [F04]. We recall that the main difficulties that one encounters in the application of general tools (such as averaging and KAM theory) to particular cases of interest in celestial mechanics, are related to the strong degeneracies of the analytical models.

The scope of this paper is to show the existence of quasi-periodic orbits lying on N-dimensional invariant elliptic tori for the planar $(N+1)$-body problem. The main difference from [H95] and [F04], besides the dimension of the constructed tori, relies on the explicit evaluations of the eigenvalues of the linearized secular dynamics (which allow us to apply more standard KAM methods).

Though the method exposed here is quite general, for concreteness we will focus our attention on a caricature of the outer solar system. More precisely, our model will be given by a Sun and N planets with relatively small masses (say, of order ε). All these $(N+1)$ bodies are considered as point masses in mutual gravitational interaction. Two planets (such as Jupiter and Saturn in the real world) will be assumed to have mass considerably bigger than the other planets. The bodies lie in a given plane and we assume that the initial configuration is far from collisions. We

*Received by the editors April 26, 2004; accepted for publication (in revised form) April 14, 2005; published electronically January 27, 2006. This work was supported by MIUR *Variational Methods and Nonlinear Differential Equations*.
http://www.siam.org/journals/sima/37-5/44364.html

†Dipartimento di Matematica, Università "Roma Tre," Largo S. L. Murialdo 1, I-00146 Roma, Italy (biasco@mat.uniroma3.it, luigi@mat.uniroma3.it).
‡Dipartimento di Matematica, Università di Roma Tor Vergata, Via della Ricerca Scientifica, 1, I-00133 Roma, Italy (valdinoci@mat.uniroma2.it).

also assume, mimicking the case of the outer solar system, that the two big planets have an orbit which is internal with respect to the orbits of the small planets. We will establish, for a large set of semiaxes, the existence of quasi-periodic orbits with small eccentricities filling up N-dimensional invariant elliptic tori. Such orbits can be seen as continuations of "limiting" circular trajectories of the system obtained by neglecting the mutual interactions among the planets. A more precise statement is given in Theorem 1.1 below.

The above "outer model," which roughly mimics some traits of physically relevant cases, has also the nice feature of providing particularly simple expressions in the related perturbing functions, as we will see in section 3 below. We stress, however, that many other situations (such as one large planet plus $N-1$ small planets; "inner" or "mixed" models, etc.) may be easily dealt with using the techniques and results presented in this paper.

The proof of our result is based on techniques developed in [BCV03] and on the explicit computation of the eigenvalues of the quadratic part of the so-called principal part of the perturbation for the planar many-body problem.

The first result on quasi-periodic orbits of interest in celestial mechanics goes back to [A63], where quasi-periodic orbits lying on 4-dimensional tori are shown to exist for the planar three-body problem (the general case was discussed there, but no complete proof was given). Related results were given in [JM66], which found linearly unstable quasi-periodic orbits lying on 2-dimensional tori for the nonplanar three-body problem. More recently, [LR95] and [R95] and [BCV03] proved the existence of quasi-periodic orbits for the nonplanar three-body problem, lying on 4-dimensional and linearly stable 2-dimensional tori, respectively. Two-dimensional invariant tori for the planar three-body problem have been found in [F02]. Periodic orbits of the nonplanar three-body problem winding around invariant tori have been constructed in [BBV04]. Finally, the existence of a positive measure set of initial data giving rise to maximal invariant tori for the planetary $(N+1)$-body problem has been established in [F04].

The paper is organized as follows. In section 1.2, we give a more precise statement of our main result. In section 2 we write down the $(N+1)$-body problem Hamiltonian in Delaunay–Poincaré variables. In section 3 (which, in a sense, is the crucial part of the paper) we discuss degeneracies. In section 4 we give the proof of the main result. The scheme of proof is similar to the one presented in [BCV03] (see also [BBV04]) in the three-body case and it is based on a "general" averaging theorem and on KAM theory for lower-dimensional tori (see [P96], [BCV03], [BBV04]). For completeness, we include a classical (but not easy to find) description of analytical properties of the Delaunay–Poincaré variables (see section 2 and Appendix A); in Appendix B we collect some simple linear algebra lemmata that are used in the arguments given in section 3.

1.2. Statement of results. We denote the $N+1$ massive points ("bodies") by $P_0, \dots, P_N$ and let $m_0, \dots, m_N$ be their masses interacting through gravity (with constant of gravitation 1). Fix $m_0 > 0$ and assume that

$$m_i = \varepsilon \mu_i, \qquad i = 1, \dots, N, \qquad 0 < \varepsilon < 1. \tag{1.1}$$

Here, ε is regarded as a small parameter and μ_i is of order 1 in ε. The point P_0 represents the "Sun" and the points P_i, $i = 1, \dots, N$, the "planets." We assume that all the bodies lie on a fixed plane, that will be identified with $\mathbf{R}^2$. The phase space of this dynamical system—the planetary, planar $(N+1)$-body system—has dimension $4N$ (after reduction by the symmetries of translations).

We will state the result in terms of orbital elements of the "osculating ellipses" of the two-body problems associated to (P_0, P_j). Let $u^{(0)}$ and $u^{(j)}$ denote the coordinates of P_0 and P_j (at a given time) and let $\dot{u}^{(0)}$ and $\dot{u}^{(j)}$ denote the corresponding velocities. By definition, the "osculating ellipse" is the ellipse described by the solution of the two-body problem (P_0, P_j) with initial data given by $(u^{(0)}, u^{(j)}, \dot{u}^{(0)}, \dot{u}^{(j)})$. Of course, such ellipses describe the motions of the full $(N+1)$-body problem only approximately; nevertheless, they provide a nice set of coordinates allowing, for example, to describe the true motions in terms of the eccentricities e_j and the major semiaxes a_j of the osculating ellipses. For further details and pictures of the orbital elements, we refer the reader to [Ch88] and [BCV03].

In this paper we consider a planetary (planar) model with planets evolving from phase points corresponding to well-separated nearly circular ellipses ($e_i \ll 1$); here "well-separated" means that

$$0 < a_i < \theta\, a_{i+1}, \qquad 1 \le i \le N-1. \tag{1.2}$$

for a suitable constant $0 < \theta < 1$. For concreteness, we shall focus on a caricature of the outer solar system; i.e., we will assume that, for some $m_0 < \bar{\mu}_i < 4m_0$,

$$\begin{aligned} &\mu_i = \bar{\mu}_i && \text{for} \quad i = 1, 2, \\ &\mu_i = \delta\bar{\mu}_i && \text{for} \quad i = 3, \dots, N, \qquad 0 < \delta < 1. \end{aligned} \tag{1.3}$$

In this setting, P_1 and P_2 imitate (in a very rough way, of course) the physical[1] features of the giant planets Jupiter and Saturn, while P_3 and P_4 represent Uranus and Neptune.[2]

A rough description of our main result is given in the following theorem; a more precise and quantitative version is given in Theorem 4.2 below.

THEOREM 1.1. *Consider a planar, planetary $(N+1)$-body system satisfying (if $N \ge 3$)* (1.1) *and* (1.3). *Let $A \subset \mathbf{R}^N$ be a compact set of semiaxes where* (1.2) *holds for a suitable $0 < \theta < 1$. Then, there exists $\delta^\star > 0$ and for any $0 < \delta < \delta^\star$ there exists $\varepsilon^\star > 0$ so that the following holds. For any $0 < \varepsilon < \varepsilon^\star$, the planetary, planar $(N+1)$-body system possesses a family of N-dimensional elliptic invariant Diophantine quasi-periodic tori; such family is parametrized by the osculating major semiaxes varying in a subset of A of density*[3] $1 - C_1\varepsilon^{c_1}$. *These motions correspond to orbits with osculating eccentricities bounded by $C_2\varepsilon^{c_2}$ and the variation in time of the osculating major semiaxes of these orbits is bounded by $C_3\varepsilon^{c_3}$.*

We have the following few comments.

- The numbers $\delta^\star$ and θ can be easily computed in the course of the proof and are not "very small"; in fact θ is a "universal" constant while $\delta^\star$ depends only on N and A. On the other hand, $\varepsilon^\star$, which depends on N, A, and δ, is related to a KAM smallness condition and rough estimates lead, as is well known, to ridiculously small quantities (for somewhat more serious KAM estimates, we refer the reader to [CC03]). Finally, the positive constants C_i's depend on N, A, and δ, while the c_i's depend only on N (and could also be easily calculated; see (4.43)).
- The assumptions (1.2) and (1.3) in the theorem are used to check *explicitly* suitable "nondegeneracy" conditions. However, giving explicit constants and

[1] A mathematical motivation for considering *two* dominant planets is given in Remark 3.2(iii).

[2] The Jupiter/Saturn mass ratio is approximately 3.34, while the Neptune/Uranus mass ratio is about 1.18 (to have it all, the Jupiter/Uranus mass ratio is ~ 21.78).

[3] Here and in what follows, the "density" is intended with respect to Lebesgue measure.

estimates, one can show that the thesis of the theorem holds, essentially, with no hypotheses on the semiaxes a_j and the rescaled masses μ_j (provided $a_i \neq a_j > 0$ and $\mu_j > 0$); a rigorous argument, based on analytic continuation of the eigenvalues, could be given along the lines discussed in [F04].

- The invariant tori found in Theorem 1.1 are *lower-dimensional elliptic tori* meaning that the dimension of the tori is strictly smaller than (in fact, half of) the dimension of the Lagrangian (maximal) tori, which have dimension $2N$. "Elliptic" means that the tori are linearly stable. It is not difficult to show that such elliptic tori are surrounded by a set of positive measure of maximal tori.
- The proof given below is based on a well-known elliptic KAM theorem, which works under "nondegeneracy" (or Melnikov) conditions. To check these conditions one has to study the eigenvalues of the "secular" (or averaged) quadratic part of the Newtonian many-body interaction, which will be denoted $\overline{\mathcal{H}}_{1,2}$; "quadratic" here refers to the symplectic Cartesian variables measuring the eccentricity and the orientation of the osculating ellipses. The diagonalization of $\overline{\mathcal{H}}_{1,2}$ is trivial (under the only assumption that $a_i \neq a_j$), while conditions (1.2) and (1.3) will be used to check that the associated eigenvalues are nonzero, simple, and distinct so that Melnikov conditions are satisfied. The proof is noninductive on N.

2. Poincaré Hamiltonian setting. The results described in this section are classical (even if not easy to find) and go back to Delaunay and Poincaré; the reader not familiar with Delaunay and Poincaré variables will find a self-contained exposition in Appendix A.

Consider $N+1$ bodies $P_0, \dots, P_N$, in a fixed (ecliptic) plane, of masses $m_0, \dots, m_N$ interacting through gravity (with constant of gravitation 1). We assume that the mass of P_0 (the "star") is much larger than the mass of the other bodies (the "planets"); i.e., we assume (1.1). In heliocentric planar (suitably rescaled) variables, the dynamics of the planar $(N+1)$-body problem is governed (as explained in Appendix A) by the Hamiltonian

$$\mathcal{H}^{(N)}(X,x) := \mathcal{H}_0^{(N)}(X,x) + \varepsilon \mathcal{H}_1^{(N)}(X,x), \tag{2.1}$$

where $X := (X^{(1)}, \dots, X^{(N)}) \in \mathbf{R}^{2N}$ and $x := (x^{(1)}, \dots, x^{(N)}) \in \mathbf{R}^{2N}$ are conjugated Cartesian symplectic variables and

$$\begin{aligned} \mathcal{H}_0^{(N)} &:= \sum_{i=1}^{N} \left(\frac{1}{2\mathrm{m_i}} |X^{(i)}|^2 - \frac{\mathrm{m}_i \mathrm{M}_i}{|x^{(i)}|} \right), \\ \mathcal{H}_1^{(N)} &:= \sum_{1 \le i < j \le N} \left(X^{(i)} \cdot X^{(j)} - \frac{\mu_i \mu_j}{m_0^2} \, \frac{1}{|x^{(i)} - x^{(j)}|} \right); \end{aligned} \tag{2.2}$$

here we have introduced the dimensionless masses[4]

$$\mathrm{M}_i := 1 + \varepsilon \frac{\mu_i}{m_0}, \qquad \mathrm{m}_i := \frac{\mu_i}{m_0 + \varepsilon \mu_i} = \frac{\mu_i}{m_0} \, \frac{1}{\mathrm{M}_i}. \tag{2.3}$$

The Hamiltonian $\mathcal{H}_0^{(N)}$ is simply the sum of N uncoupled planar Kepler problems (formed by the star and the ith planet). Being interested in phase region where

[4]Beware not to confuse the dimensionless masses m_i with the real masses m_i introduced at the beginning of section 1.2.

the uncoupled Kepler problem describes nearly circular orbits, we introduce *planar Poincaré variables*, the construction of which is based on the classical 4-dimensional symplectic map (2.8) below. Let

$$F_1(t) := \left(1-\frac{t}{4}\right)^{\frac{1}{2}}, \qquad F_2(t) := \frac{1}{2}\left(1-\frac{t}{4}\right)^{-1}, \qquad \left(|t|<\frac{1}{4}\right); \tag{2.4}$$

let $G_0(s,t) = t + st + \cdots$ be the function analytic in a neighborhood of $(0,0)$ implicitly defined by

$$G_0(0,0) = 0, \qquad G_0 = s\sin G_0 + t\cos G_0; \tag{2.5}$$

define the following four functions of three variables $(\hat\eta, \hat\xi, \lambda)$ real-analytic in a neighborhood of the set $\{(\hat\eta,\hat\xi) = (0,0)\} \times \mathbf{T}$:

$$\begin{aligned}
G(\hat\eta,\hat\xi,\lambda) &:= G_0\Big((\hat\eta\cos\lambda - \hat\xi\sin\lambda)\ F_1(t),\ (\hat\xi\cos\lambda - \hat\eta\sin\lambda)\ F_1(t)\Big),\\
\mathcal{E}_{\mathrm{s}}(\hat\eta,\hat\xi,\lambda) &:= \Big(\hat\xi\cos(\lambda+G) + \hat\eta\sin(\lambda+G)\Big)\ F_1(t),\\
\mathcal{C}(\hat\eta,\hat\xi,\lambda) &:= \cos(\lambda+\mathcal{E}_{\mathrm{s}}) - \hat\eta\ F_1(t) - \hat\xi\,\mathcal{E}_{\mathrm{s}}\ F_1(t)\,F_2(t),\\
\mathcal{S}(\hat\eta,\hat\xi,\lambda) &:= \sin(\lambda+\mathcal{E}_{\mathrm{s}}) + \hat\xi\ F_1(t) - \hat\eta\,\mathcal{E}_{\mathrm{s}}\ F_1(t)\,F_2(t),
\end{aligned} \tag{2.6}$$

where t is short for $t = \hat\eta^2 + \hat\xi^2$, G is short for $G(\hat\eta,\hat\xi,\lambda)$, and $\mathcal{E}_{\mathrm{s}}$ is short for $\mathcal{E}_{\mathrm{s}}(\hat\eta,\hat\xi,\lambda)$.

LEMMA 2.1 (planar Poincaré variables). *Fix $\varepsilon, \mu, m_0 > 0$ and let*

$$\begin{aligned}
&\mathrm{M} := 1 + \varepsilon\frac{\mu}{m_0}, \qquad \mathrm{m} := \frac{\mu}{m_0}\ \frac{1}{\mathrm{M}}, \qquad \overline{\mathrm{m}} := \frac{\mu}{m_0}\ \frac{1}{\sqrt{\mathrm{M}}},\\
&\sigma := \left(\frac{\mu}{m_0}\right)^3\frac{1}{\mathrm{M}}, \qquad a = a(\Lambda;\mu,\varepsilon) := \frac{\Lambda^2}{\overline{\mathrm{m}}^2}.
\end{aligned} \tag{2.7}$$

Then, for any $\Lambda_+ > \Lambda_- > 0$, there exists a ball B around the origin in $\mathbf{R}^2$ such that the 4-dimensional map

$$\Psi_{\mathrm{P}} : (\Lambda,\lambda,\eta,\xi) \in D := (\Lambda_-,\Lambda_+)\times\mathbf{T}\times B \to (X,x) \in \mathbf{R}^4,$$

where

$$\begin{aligned}
x_1 &= x_1(\Lambda,\lambda,\eta,\xi;\mu,\varepsilon) := a(\Lambda;\mu,\varepsilon)\ \mathcal{C}\left(\frac{\eta}{\sqrt\Lambda},\frac{\xi}{\sqrt\Lambda},\lambda\right),\\
x_2 &= x_2(\Lambda,\lambda,\eta,\xi;\mu,\varepsilon) := a(\Lambda;\mu,\varepsilon)\ \mathcal{S}\left(\frac{\eta}{\sqrt\Lambda},\frac{\xi}{\sqrt\Lambda},\lambda\right),\\
X &= X(\Lambda,\lambda,\eta,\xi;\mu,\varepsilon) := \frac{\overline{\mathrm{m}}^4}{\Lambda^3}\frac{\partial x}{\partial\lambda}(\Lambda,\lambda,\eta,\xi;\mu,\varepsilon) = \frac{\overline{\mathrm{m}}}{a(\Lambda;\mu,\varepsilon)^{3/2}}\ \frac{\partial x}{\partial\lambda},
\end{aligned} \tag{2.8}$$

is real-analytic in D and symplectic:

$$d\Lambda\wedge d\lambda + d\eta\wedge d\xi = dX_1\wedge dx_1 + dX_2\wedge dx_2.$$

Furthermore, if $\mathcal{H}_0^{(1)}$ denotes the two-body Hamiltonian

$$\mathcal{H}_0^{(1)}(X,x) := \frac{1}{2\mathrm{m}}|X|^2 - \frac{\mathrm{mM}}{|x|},$$

then, on the phase region of negative energies $(\mathcal{H}_0^{(1)})^{-1}(-\frac{\sigma}{2\Lambda_-^2}, -\frac{\sigma}{2\Lambda_+^2})$*, one has*

$$\mathcal{H}_0^{(1)} \circ \Psi_{\mathrm{P}} = -\frac{\sigma}{2\Lambda^2};$$

in the planar coordinates $x \in \mathbf{R}^2$ *the corresponding motion describes an ellipse of major semiaxis* $a = a(\Lambda; \mu, \varepsilon)$ *and eccentricity*

$$e = \sqrt{\frac{\eta^2+\xi^2}{\Lambda}}\ F_1\left(\frac{\eta^2+\xi^2}{\Lambda}\right) = \sqrt{\frac{\eta^2+\xi^2}{\Lambda}\ \left(1 - \frac{\eta^2+\xi^2}{4\Lambda}\right)}. \tag{2.9}$$

The proof of this lemma can be found in Appendix A. Note that

$$\mathcal{C}(0,0,\lambda) = \cos\lambda, \qquad \mathcal{S}(0,0,\lambda) = \sin\lambda, \tag{2.10}$$

so that the $(a,\lambda) \to x$ transformation is, for $\eta = \xi = 0$, just polar coordinates. Let, now, $\Psi_{\mathrm{P}}^{(N)}$ be the $4N$-dimensional map, parametrized by $(\mu_1, \ldots, \mu_N, \varepsilon)$, defined by

$$\Psi_{\mathrm{P}}^{(N)} : \Big((\Lambda_1, \lambda_1, \eta_1, \xi_1), \ldots, (\Lambda_N, \lambda_N, \eta_N, \xi_N)\Big) \in \Big((0,\infty) \times \mathbf{T} \times \mathbf{R}^2\Big)^N \to (X, x) \tag{2.11}$$

with

$$(X, x) = \Big((X^{(1)}, \ldots, X^{(N)}), (x^{(1)}, \ldots, x^{(N)})\Big),$$
$$(X^{(i)}, x^{(i)}) = \Psi_{\mathrm{P}}(\Lambda_i, \lambda_i, \eta_i, \xi_i; \mu_i, \varepsilon). \tag{2.12}$$

Then, $\Psi_{\mathrm{P}}^{(N)}$ is symplectic and

$$\mathcal{H}_0^{(N)} \circ \Psi_{\mathrm{P}}^{(N)} = -\frac{1}{2}\sum_{i=1}^N \frac{\sigma_i}{\Lambda_i^2} =: \mathcal{H}_0(\Lambda), \qquad \sigma_i := \left(\frac{\mu_i}{m_0}\right)^3 \frac{1}{\mathrm{M}_i}. \tag{2.13}$$

In such Poincaré variables the full planar $(N+1)$-body Hamiltonian $\mathcal{H}^{(N)}$ becomes

$$\mathcal{H}(\Lambda, \lambda, \eta, \xi) = \mathcal{H}_0(\Lambda) + \varepsilon\mathcal{H}_1(\Lambda, \lambda, \eta, \xi), \qquad \mathcal{H}_1 := \mathcal{H}_1^{(N)} \circ \Psi_{\mathrm{P}}^{(N)} =: \mathcal{H}_1^{\mathrm{compl}} + \mathcal{H}_1^{\mathrm{princ}}, \tag{2.14}$$

where the so-called "complementary part" $\mathcal{H}_1^{\mathrm{compl}}$ and the "principal part" $\mathcal{H}_1^{\mathrm{princ}}$ of the perturbation are, respectively, the functions

$$\sum_{1 \le i < j \le N} X^{(i)} \cdot\ X^{(j)} \qquad \text{and} \qquad \sum_{1 \le i < j \le N} \frac{\mu_i \mu_j}{m_0^2} \frac{1}{|x^{(i)} - x^{(j)}|} \tag{2.15}$$

expressed in Poincaré variables:[5]

$$X^{(i)} = X(\Lambda_i, \lambda_i, \eta_i, \xi_i; \mu_i, \varepsilon) \qquad \text{and} \qquad x^{(i)} = x(\Lambda_i, \lambda_i, \eta_i, \xi_i; \mu_i, \varepsilon).$$

[5] $X = (X_1, X_2)$ and $x = (x_1, x_2)$ denote here the functions defined in (2.8).

Notice that, since $X^{(i)} = (\overline{\mathrm{m}}_i^4/\Lambda_i^3)\partial_{\lambda_i} x^{(i)}$ the λ-average of $\mathcal{H}_1^{\mathrm{compl}}$ vanishes. Moreover, as it is well known, the λ-average of $\mathcal{H}_1$ is an even function of (η, ξ); see, also, Appendix A. Hence, we may split the perturbation function as

$$\mathcal{H}_1 = \overline{\mathcal{H}}_1 + \widetilde{\mathcal{H}}_1 \tag{2.16}$$

with

$$\overline{\mathcal{H}}_1(\Lambda, \eta, \xi) := \int_{\mathbf{T}^N} \mathcal{H}_1 \frac{d\lambda}{(2\pi)^N}, \qquad \int_{\mathbf{T}^N} \widetilde{\mathcal{H}}_1 \, d\lambda = 0. \tag{2.17}$$

Furthermore, $\overline{\mathcal{H}}_1$ may be written as

$$\overline{\mathcal{H}}_1(\Lambda, \eta, \xi) = \overline{\mathcal{H}}_{1,0}(\Lambda) + \overline{\mathcal{H}}_{1,2}(\Lambda, \eta, \xi) + \overline{\mathcal{H}}_{1,*}(\Lambda, \eta, \xi), \tag{2.18}$$

where $\overline{\mathcal{H}}_{1,0} := \overline{\mathcal{H}}_1(\Lambda, 0, 0)$, $\overline{\mathcal{H}}_{1,2}$ is the (η, ξ)-quadratic part of $\overline{\mathcal{H}}_1$ while $\overline{\mathcal{H}}_{1,*}$ is the "remainder of order four":

$$|\overline{\mathcal{H}}_{1,*}(\Lambda, \eta, \xi)| \le \mathrm{const}\, |(\eta, \xi)|^4.$$

3. The averaged quadratic potential $\overline{\mathcal{H}}_{1,2}$. In this section we analyze the function $\overline{\mathcal{H}}_{1,2}$ (i.e., the (η, ξ)-quadratic part of the λ-average of the perturbation) defined in (2.18), which may be written as

$$\overline{\mathcal{H}}_{1,2} = \frac{1}{2} \sum_{1 \le i,j \le N} Q_{ij} \begin{pmatrix} \eta_j \\ \xi_j \end{pmatrix} \cdot \begin{pmatrix} \eta_i \\ \xi_i \end{pmatrix}, \tag{3.1}$$

where Q_{ij} are (2×2) matrices defined as

$$Q_{ij} := \left. \begin{pmatrix} \partial^2_{\eta_i, \eta_j} \overline{\mathcal{H}}_{1,2} & \partial^2_{\eta_i, \xi_j} \overline{\mathcal{H}}_{1,2} \\ \partial^2_{\xi_i, \eta_j} \overline{\mathcal{H}}_{1,2} & \partial^2_{\xi_i, \xi_j} \overline{\mathcal{H}}_{1,2} \end{pmatrix} \right|_{(\Lambda, 0, 0)}.$$

The aim of this section is to prove that there exists a symplectic linear change of variables $(p, q) \to (\eta, \xi)$ putting the quadratic part (3.1) in the normal form

$$\frac{1}{2} \sum_{i=1}^N \bar{\Omega}_i (p_i^2 + q_i^2); \tag{3.2}$$

see Remark 3.1(i). A crucial fact, in order to apply KAM theory, consists in proving that such $\bar{\Omega}_i$'s are nondegenerate[6] in the sense that they are nonvanishing and distinct. Such nondegeneracy is proved in Proposition 3.2 in which we manage to compute explicitly the asymptotics of the $\bar{\Omega}_i$'s.

In view of the definition of the Poincaré variables, we look at the rescaled variables $(\hat{\eta}, \hat{\xi})$ rather than (η, ξ). Therefore, we define

$$\bar{f}_{ij}(\Lambda, \hat{\eta}, \hat{\xi}) := \frac{1}{(2\pi)^N} \int_{\mathbf{T}^N} \frac{d\lambda}{|x^{(i)}(\Lambda_i, \lambda_i, \sqrt{\Lambda_i}\hat{\eta}_i, \sqrt{\Lambda_i}\hat{\xi}_i; \mu_i, \varepsilon) - x^{(j)}(\Lambda_j, \lambda_j \sqrt{\Lambda_j}\hat{\eta}_j, \sqrt{\Lambda_j}\hat{\xi}_j; \mu_j, \varepsilon)|}. \tag{3.3}$$

[6]See the "Melnikov condition" (4.33).

Thus, letting[7]

$$a_i := a(\Lambda_i;\mu_i,\varepsilon), \qquad\qquad c_{ij} := \frac{1}{m_0}\left(\frac{\mathrm{M}_i\mathrm{M}_j}{a_ia_j}\right)^{1/4},$$

$$A_{ij} := \begin{pmatrix} \partial^2_{\hat\eta_i\hat\eta_j}\bar f_{ij} & \partial^2_{\hat\eta_i\hat\xi_j}\bar f_{ij} \\ \partial^2_{\hat\xi_i\hat\eta_j}\bar f_{ij} & \partial^2_{\hat\xi_i\hat\xi_j}\bar f_{ij}\end{pmatrix}\Bigg|_{\hat\eta=\hat\xi=0}, \quad B_{ij} := \begin{pmatrix} \partial^2_{\hat\eta_j\hat\eta_j}\bar f_{ij} & \partial^2_{\hat\eta_j\hat\xi_j}\bar f_{ij} \\ \partial^2_{\hat\xi_j\hat\eta_j}\bar f_{ij} & \partial^2_{\hat\xi_j\hat\xi_j}\bar f_{ij}\end{pmatrix}\Bigg|_{\hat\eta=\hat\xi=0}, \tag{3.4}$$

we find

$$Q_{ij} = \begin{cases} \sqrt{\mu_i\mu_j}\,c_{ij}\;A_{ij} & \text{if } i\neq j,\\ \displaystyle\sum_{k\neq j}\sqrt{\mu_k\mu_j}\,c_{kj}\;B_{kj} & \text{if } i=j.\end{cases}$$

It is a remarkable fact that, *for the planar planetary* $(N+1)$*-body problem the matrices* A_{ij} *and* B_{ij} *are proportional to the* (2×2) *identity matrix* $\mathbf{1}_2 = \left(\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right)$ and have simple integral representation. In fact, define, for $a\neq b$,

$$\mathcal{J}(a,b) := \frac{1}{2\pi}\int_0^{2\pi}\frac{-17ab\cos t + 8(a^2+b^2)\cos(2t) + ab\cos(3t)}{(a^2+b^2-2ab\cos t)^{5/2}}\,dt,$$

$$\mathcal{I}(a,b) := \frac{1}{2\pi}\int_0^{2\pi}\frac{-7ab + 4(a^2+b^2)\cos t - ab\cos(2t)}{(a^2+b^2-2ab\cos t)^{5/2}}\,dt,$$

and denote, for $a_i\neq a_j$,

$$\alpha_{ij} := \frac{a_ia_j}{8}\mathcal{J}(a_i,a_j), \qquad \beta_{ij} := \frac{a_ia_j}{4}\mathcal{I}(a_i,a_j). \tag{3.5}$$

Then, the following "algebraic" result holds.

PROPOSITION 3.1. *Assume* $a_i\neq a_j$ *for* $i\neq j$. *Then* $A_{ij} = \alpha_{ij}\mathbf{1}_2$ *and* $B_{ij} = \beta_{ij}\mathbf{1}_2$.

Remark 3.1. (i) An immediate corollary of this result is that, in the collisionless domain $\{a_i\neq a_j\}$, $\overline{\mathcal{H}}_{1,2}$ has the simple form

$$\overline{\mathcal{H}}_{1,2} = \frac12\Big(M\eta\cdot\eta + M\xi\cdot\xi\Big), \tag{3.6}$$

M being the real, symmetric $(N\times N)$ matrix with entries

$$M_{ij} = \begin{cases} \sqrt{\mu_i\mu_j}\,c_{ij}\;\alpha_{ij} & \text{if } i\neq j,\\ \displaystyle\sum_{k\neq j}\sqrt{\mu_k\mu_j}\,c_{kj}\;\beta_{kj} & \text{if } i=j.\end{cases} \tag{3.7}$$

The Hamiltonian (3.6) can be immediately put in symplectic normal form: if U is the real orthogonal matrix ($U^T = U^{-1}$) which diagonalizes M ($U^TMU = \mathrm{diag}\,(\bar\Omega_1,\dots,\bar\Omega_N)$), then the map $p = U^T\eta$, $q = U^T\xi$ is symplectic and, in such variables, the new Hamiltonian takes the form (3.2).

(ii) The functions $\mathcal{J}$ and $\mathcal{I}$ (which admit simple representations in terms of Gauss hypergeometric functions) are symmetric ($\mathcal{J}(a,b) = \mathcal{J}(b,a)$ and $\mathcal{I}(a,b) = \mathcal{I}(b,a)$) and satisfy

$$\mathcal{J}(a,b) = b^{-3}\mathcal{J}(a/b,1), \qquad \mathcal{I}(a,b) = b^{-3}\mathcal{I}(a/b,1), \qquad a<b.$$

[7]Recall (2.7) and (2.3).

The functions of one real variable $s \in (-1,1) \to \mathcal{J}(s,1)$ and $s \in (-1,1) \to \mathcal{I}(s,1)$ are, respectively, even and odd in s, and satisfy, for small s, the following asymptotics:

$$\mathcal{J}(s,1) = -\frac{15}{8}s^2 - \frac{105}{8}s^4 + O(s^6), \qquad \mathcal{I}(s,1) = 3s + \frac{45}{8}s^3 + O(s^5). \tag{3.8}$$

(iii) Proposition 3.1 is a suitable version of a well-known result which can be found, e.g., in [Poi1905]; see also [LR95].

(iv) The asymptotics of the α_{ij}'s and β_{ij}'s may be also computed in terms of the Laplace coefficients (see, e.g., [LR95]); for our purposes it is simpler to derive the needed asymptotics directly from the integral representations given before (3.5).

Proof of Proposition 3.1. The computations we are going to perform are algebraic in character and it is therefore enough to consider real variables. Fix $i \neq j$ and define

$$\mathcal{R}_{ij}(\Lambda,\lambda,\hat\eta,\hat\xi) := \left| x^{(i)}(\Lambda_i,\lambda_i,\sqrt{\Lambda_i}\hat\eta_i,\sqrt{\Lambda_i}\hat\xi_i;\mu_i,\varepsilon) - x^{(j)}(\Lambda_j,\lambda_j\sqrt{\Lambda_j}\hat\eta_j,\sqrt{\Lambda_j}\hat\xi_j;\mu_j,\varepsilon)\right|^2, \tag{3.9}$$

so that (recall (3.3))

$$\bar f_{ij}(\Lambda,\hat\eta,\hat\xi) = \frac{1}{(2\pi)^N}\int_{\mathbf{T}^N}\frac{d\lambda}{\sqrt{\mathcal{R}_{ij}}}. \tag{3.10}$$

By (2.8) we find

$$\mathcal{R}_{ij} = a_i^2\chi_i^2 + a_j^2\chi_j^2 - 2a_ia_j\Big(\mathcal{C}_i\mathcal{C}_j - \mathcal{S}_i\mathcal{S}_j\Big), \tag{3.11}$$

where $\mathcal{C}_k$, $\mathcal{S}_k$, and χ_k are short for, respectively,

$$\mathcal{C}_k = \mathcal{C}(\hat\eta_k,\hat\xi_k,\lambda_k), \quad \mathcal{S}_k = \mathcal{S}(\hat\eta_k,\hat\xi_k,\lambda_k), \quad \text{and} \quad \chi_k = \sqrt{\mathcal{C}_k^2 + \mathcal{S}_k^2}.$$

The proof will consist in computing explicitly λ-averages of quantities of the form

$$\rho_{\zeta_i,\zeta_j}(\lambda_i,\lambda_j) := \partial^2_{\zeta_i\zeta_j}\frac{1}{\sqrt{\mathcal{R}_{ij}}}\Bigg|_{\hat\eta=\hat\xi=0} = \frac{3(\partial_{\zeta_i}\mathcal{R}_{ij})(\partial_{\zeta_j}\mathcal{R}_{ij}) - 2\mathcal{R}_{ij}(\partial_{\zeta_i\zeta_j}\mathcal{R}_{ij})}{4\mathcal{R}_{ij}^{5/2}}\Bigg|_{\hat\eta=\hat\xi=0}, \tag{3.12}$$

where ζ_k denotes either of the variables $\hat\eta_k$ or $\hat\xi_k$. Thus, what we need to do is to compute suitable orders in the variables $(\hat\eta_k,\hat\xi_k)$ of the function $\mathcal{R}_{ij}$. For this purpose the following lemma will be useful.

LEMMA 3.1. *Define the following elementary functions:*

$$\begin{aligned}
C_+(\lambda) &:= 1+\cos^2\lambda \;=\; \frac{3+\cos(2\lambda)}{2},\\
C_-(\lambda) &:= 1+\sin^2\lambda \;=\; \frac{3-\cos(2\lambda)}{2},\\
S_0(\lambda) &:= \cos\lambda\sin\lambda \;=\; \frac{1}{2}\sin(2\lambda),\\
\bar\chi(x,y,\lambda) &:= 1-2y\cos\lambda+2x\sin\lambda,\\
S(x,y,\lambda) &:= \sin\lambda + xC_+(\lambda) + yS_0(\lambda),\\
C(x,y,\lambda) &:= \cos\lambda - yC_-(\lambda) - xS_0(\lambda),
\end{aligned}$$

and denote by $O_p(z_1,\dots,z_n)$ *a function of the variables* $(z_1,\dots,z_n)$ *(depending possibly on other variables) analytic in a neighborhood of* $(0,\dots,0)$ *and starting with a homogeneous polynomial of degree p in* $(z_1,\dots,z_n)$. *Then,*

$$\begin{aligned}
&\chi_k^2 = \mathcal{C}(\hat\eta_k,\hat\xi_k,\lambda_k)^2+\mathcal{S}(\hat\eta_k,\hat\xi_k,\lambda_k)^2=\bar\chi(\hat\eta_k,\hat\xi_k,\lambda_k)+O_2(\hat\eta_k,\hat\xi_k),\\
&\mathcal{C}_k=\mathcal{C}(\hat\eta_k,\hat\xi_k,\lambda_k)=C(\hat\xi_k,\hat\eta_k,\lambda_k)+O_2(\hat\eta_k,\hat\xi_k),\\
&\mathcal{S}_k=\mathcal{S}(\hat\eta_k,\hat\xi_k,\lambda_k)=S(\hat\xi_k,\hat\eta_k,\lambda_k)+O_2(\hat\eta_k,\hat\xi_k).
\end{aligned}\tag{3.13}$$

The proof of this lemma follows at once from the explicit expressions for $\mathcal{C}$ and $\mathcal{S}$ given in Lemma 2.1 and is left to the reader.

We consider first the matrices A_{ij} (which allow to compute Q_{ij} for $i\neq j$) and then we turn to the matrices B_{ij} (which allow to compute Q_{jj}).

Computation of the matrices A_{ij}. First, observe that the two derivatives involved in the definition of A_{ij} are *always* mixed in the variables with indexes i and j. Thus, we can neglect the terms of third order in $(\hat\eta_i,\hat\xi_i,\hat\eta_j,\hat\xi_j)$ and the terms of *second order* of the type $O_2(\hat\eta_i,\hat\xi_i)$ and $O_2(\hat\eta_j,\hat\xi_j)$.

By Lemma 3.1, the function $\mathcal{R}_{ij}$ in (3.11) has the form

$$\begin{aligned}
\mathcal{R}_{ij}&=a_i^2(1-2\hat\eta_i\cos\lambda_i+2\hat\xi_i\sin\lambda_i)+a_j^2(1-2\hat\eta_j\cos\lambda_j+2\hat\xi_j\sin\lambda_j)\\
&\quad-2a_ia_j\Big[(\cos\lambda_i-\hat\xi_iS_0(\lambda_i)-\hat\eta_iC_-(\lambda_i))(\cos\lambda_j-\hat\xi_jS_0(\lambda_j)-\hat\eta_jC_-(\lambda_j))\\
&\quad+(\sin\lambda_i+\hat\xi_iC_+(\lambda_i)+\hat\eta_iS_0(\lambda_i))(\sin\lambda_j+\hat\xi_jC_+(\lambda_j)+\hat\eta_jS_0(\lambda_j))\Big]\\
&\quad+O_2(\hat\eta_i,\hat\xi_i)+O_2(\hat\eta_j,\hat\xi_j)+O_3(\hat\eta_i,\hat\xi_i,\hat\eta_j,\hat\xi_j).
\end{aligned}\tag{3.14}$$

Therefore, letting $(\cdot)|_0$ be short for $(\cdot)|_{\hat\eta_i=\hat\xi_i=\hat\eta_j=\hat\xi_j=0}$, one finds

$$\begin{aligned}
\mathcal{R}_{ij}\big|_0&=a_i^2+a_j^2-2a_ia_j\cos(\lambda_i-\lambda_j),\\
\partial_{\hat\eta_i}\mathcal{R}_{ij}\big|_0&=-2a_i^2\cos\lambda_i-2a_ia_j[-C_-(\lambda_i)\cos\lambda_j+S_0(\lambda_i)\sin\lambda_j],\\
\partial_{\hat\eta_j}\mathcal{R}_{ij}\big|_0&=-2a_j^2\cos\lambda_j-2a_ia_j[-C_-(\lambda_j)\cos\lambda_i+S_0(\lambda_j)\sin\lambda_i],\\
\partial_{\hat\xi_i}\mathcal{R}_{ij}\big|_0&=2a_i^2\sin\lambda_i-2a_ia_j[C_+(\lambda_i)\sin\lambda_j-S_0(\lambda_i)\cos\lambda_j],\\
\partial_{\hat\xi_j}\mathcal{R}_{ij}\big|_0&=2a_j^2\sin\lambda_j-2a_ia_j[C_+(\lambda_j)\sin\lambda_i-S_0(\lambda_j)\cos\lambda_i],\\
\partial^2_{\hat\eta_i\hat\eta_j}\mathcal{R}_{ij}\big|_0&=-2a_ia_j[C_-(\lambda_i)C_-(\lambda_j)+S_0(\lambda_i)S_0(\lambda_j)],\\
\partial^2_{\hat\xi_i\hat\xi_j}\mathcal{R}_{ij}\big|_0&=-2a_ia_j[C_+(\lambda_i)C_+(\lambda_j)+S_0(\lambda_i)S_0(\lambda_j)],\\
\partial^2_{\hat\eta_i\hat\xi_j}\mathcal{R}_{ij}\big|_0&=-2a_ia_j[C_-(\lambda_i)S_0(\lambda_j)+S_0(\lambda_i)C_+(\lambda_j)],\\
\partial^2_{\hat\xi_i\hat\eta_j}\mathcal{R}_{ij}\big|_0&=-2a_ia_j[S_0(\lambda_i)C_-(\lambda_j)+C_+(\lambda_i)S_0(\lambda_j)].
\end{aligned}\tag{3.15}$$

In particular, $\mathcal{R}_{ij}$ and $\partial_{\hat\eta_i}\mathcal{R}_{ij}$ are even in $(\lambda_i,\lambda_j)\in\mathbf{T}^2$, while $\partial_{\hat\xi_j}\mathcal{R}_{ij}$ and $\partial^2_{\hat\eta_i\hat\xi_j}\mathcal{R}_{ij}$ are odd. Thus, recalling the definition of $\rho_{\zeta_i,\zeta_j}(\lambda_i,\lambda_j)$ in (3.12), we find that $\rho_{\hat\eta_i,\hat\xi_j}(\lambda_i,\lambda_j)$ is odd in (λ_i,λ_j) and it has therefore zero average. For the same reasons, also $\rho_{\hat\xi_i,\hat\eta_j}(\lambda_i,\lambda_j)$ has zero average. Hence, the off-diagonal terms of A_{ij} are zero. We now compute the diagonal terms of A_{ij}. We begin with $\rho_{\hat\eta_i,\hat\eta_j}(\lambda_i,\lambda_j)$. By (3.12) and the list in (3.15), we find

$$\rho_{\hat\eta_i,\hat\eta_j}(\lambda_i,\lambda_j) := \partial^2_{\hat\eta_i\hat\eta_j}\frac{1}{\sqrt{\mathcal{R}_{ij}}}\bigg|_0 = \frac{\rho_1(\lambda_i,\lambda_j)}{\rho_2(\lambda_i,\lambda_j)} \tag{3.16}$$

with

(3.17)

$$\begin{aligned}\rho_1(\lambda_i,\lambda_j) := a_ia_j\cdot\Big[&-24a_i^2\cos(2\lambda_i)-24a_j^2\cos(2\lambda_j)+8(a_i^2+a_j^2)\cos(2(\lambda_i-\lambda_j))\\ &-3a_ia_j\cos(\lambda_i-3\lambda_j)-17a_ia_j\cos(\lambda_i-\lambda_j)\\ &+a_ia_j\cos(3(\lambda_i-\lambda_j))-3a_ia_j\cos(3\lambda_i-\lambda_j)\\ &+54a_ia_j\cos(\lambda_i+\lambda_j)\Big],\\ \rho_2(\lambda_i,\lambda_j) := 8\Big(&a_i^2+a_j^2-2a_ia_j\cos(\lambda_i-\lambda_j)\Big)^{5/2}.\end{aligned}$$

Thus, changing the variable of integration, one finds

$$\begin{aligned}&\frac{1}{(2\pi)^2}\int_{\mathbf{T}^2}\partial^2_{\hat\eta_i\hat\eta_j}\frac{1}{\sqrt{\mathcal{R}_{ij}}}\bigg|_0 d\lambda_i\,d\lambda_j\\ &\qquad=\frac{1}{2\pi}\int_{\mathbf{T}}a_ia_i\cdot\frac{-17a_ia_j\cos t+8(a_i^2+a_j^2)\cos(2t)+a_ia_j\cos(3t)}{8(a_i^2+a_j^2-2a_ia_j\cos t)^{5/2}}\,dt\\ &\qquad=\frac{a_ia_j}{8}\,\mathcal{J}(a_i,a_j)=:\alpha_{ij}.\end{aligned}$$

The case $\rho_{\hat\xi_i,\hat\xi_j}(\lambda_i,\lambda_j)$ is very similar (and will yield the same result). In place of (3.16) one finds

$$\rho_{\hat\xi_i,\hat\xi_j}(\lambda_i,\lambda_j) := \partial^2_{\hat\xi_i\hat\xi_j}\frac{1}{\sqrt{\mathcal{R}_{ij}}}\bigg|_0 = \frac{\rho_3(\lambda_i,\lambda_j)}{\rho_2(\lambda_i,\lambda_j)} \tag{3.18}$$

with

$$\begin{aligned}\rho_3(\lambda_i,\lambda_j) := a_ia_j\cdot\Big[&24a_i^2\cos(2\lambda_i)+24a_j^2\cos(2\lambda_j)+8(a_i^2+a_j^2)\cos(2(\lambda_i-\lambda_j))\\ &+3a_ia_j\cos(\lambda_i-3\lambda_j)-17a_ia_j\cos(\lambda_i-\lambda_j)\\ &+a_ia_j\cos(3(\lambda_i-\lambda_j))+3a_ia_j\cos(3\lambda_i-\lambda_j)\\ &-54a_ia_j\cos(\lambda_i+\lambda_j)\Big].\end{aligned} \tag{3.19}$$

Integrating, one finds again

$$\frac{1}{(2\pi)^2}\int_{\mathbf{T}^2}\partial^2_{\hat\xi_i\hat\xi_j}\frac{1}{\sqrt{\mathcal{R}_{ij}}}\bigg|_0 d\lambda_i\,d\lambda_j=\alpha_{ij}.$$

This proves Proposition 3.1 in the case of Q_{ij}, with $i\neq j$.

Computation of the matrices B_{ij}. Observe that the derivatives involved in the definition of B_{ij} are two derivatives with the same index j. We can, therefore, neglect the third order terms and set $\hat\eta_i=\hat\xi_i=0$.

Recalling (2.10) we see that $\chi_i|_{\hat\eta_i=\hat\xi_i=0}=1$ and

$$\mathcal{R}_{ij}\Big|_{\hat\eta_i=\hat\xi_i=0}=a_i^2+a_j^2\chi_j^2-2a_ia_j\chi_j\left((\cos\lambda_i)\,\frac{\mathcal{C}_j}{\chi_j}-(\sin\lambda_i)\,\frac{\mathcal{S}_j}{\chi_j}\right). \tag{3.20}$$

Defining $\varphi_j=\varphi_j(\Lambda_j,\lambda_j,\hat\eta_j,\hat\xi_j)$ through the relations[8]

$$\cos\varphi_j=\frac{\mathcal{C}_j}{\chi_j},\qquad \sin\varphi_j=\frac{\mathcal{S}_j}{\chi_j}, \tag{3.21}$$

we find

$$\mathcal{R}_{ij}\Big|_{\hat\eta_i=\hat\xi_i=0}=a_i^2+a_j^2\chi_j^2-2a_ia_j\chi_j\cos(\varphi_j-\lambda_i). \tag{3.22}$$

Denote by $\langle\, f\,\rangle_{\theta,\tau}$ the average of a function f over the angles θ and τ. Integrating *first* with respect to λ_i and changing variable of integration ($t=\lambda_i-\varphi_j$), one gets

$$\left\langle\,\frac{1}{\sqrt{\mathcal{R}_{ij}}}\Big|_{\hat\eta_i=\hat\xi_i=0}\right\rangle_{\lambda_i,\lambda_j}=\left\langle\,\frac{1}{\sqrt{\tilde{\mathcal{R}}_{ij}}}\,\right\rangle_{t,\lambda_j} \tag{3.23}$$

with

$$\tilde{\mathcal{R}}_{ij}:=a_i^2+a_j^2\chi_j^2-2a_ia_j\chi_j\cos t. \tag{3.24}$$

At this point, the argument is completely analogous to that used above. First, we observe that

$$\left\langle\,\partial^2_{\zeta_h\zeta_k}\frac{1}{\sqrt{\mathcal{R}_{ij}}}\Big|_0\right\rangle_{\lambda_i,\lambda_j}=\left\langle\,\frac{3(\partial_{\zeta_h}\tilde{\mathcal{R}}_{ij})(\partial_{\zeta_k}\tilde{\mathcal{R}}_{ij})-2\tilde{\mathcal{R}}_{ij}(\partial^2_{\zeta_h\zeta_k}\tilde{\mathcal{R}}_{ij})}{4\tilde{\mathcal{R}}_{ij}^{5/2}}\Big|_0\right\rangle_{t,\lambda_j}, \tag{3.25}$$

where ζ_ℓ denotes here any of the variables $\hat\eta_j,\hat\xi_j$. From Lemma 3.1 it follows that $\tilde{\mathcal{R}}_{ij}$ can be written as

$$\tilde{\mathcal{R}}_{ij}=f(t)+g(t)(h_1-h_2)+a_j^2h_1^2+O_3(\hat\eta_j,\hat\xi_j)$$

with

$$f(t):=a_i^2+a_j^2-2a_ia_j\cos t,\qquad g(t):=-2a_j^2+2a_ia_j\cos t,$$
$$h_1:=\hat\eta_j\cos\lambda_j-\hat\xi_j\sin\lambda_j,\qquad h_2:=\hat\xi_j^2\cos^2\lambda_j+\hat\eta_j^2\sin^2\lambda_j+\hat\eta_j\hat\xi_j\sin(2\lambda_j).$$

Thus, since h_1 is of order one in $(\hat\eta_j,\hat\xi_j)$ and h_2 is of order two in $(\hat\eta_j,\hat\xi_j)$,

$$\begin{aligned}
\tilde{\mathcal{R}}_{ij}\Big|_0&=f(t),\\
\partial_{\eta_j}\tilde{\mathcal{R}}_{ij}\Big|_0&=g(t)\,\cos\lambda_j,\\
\partial_{\hat\xi_j}\tilde{\mathcal{R}}_{ij}\Big|_0&=-g(t)\,\sin\lambda_j,\\
\partial^2_{\hat\eta_j\hat\eta_j}\tilde{\mathcal{R}}_{ij}\Big|_0&=-2g(t)\,\sin^2\lambda_j+2a_j^2\cos^2\lambda_j,\\
\partial^2_{\hat\eta_j\hat\xi_j}\tilde{\mathcal{R}}_{ij}\Big|_0&=-(g(t)+a_j)^2\,\sin(2\lambda_j),\\
\partial^2_{\hat\xi_j\hat\xi_j}\tilde{\mathcal{R}}_{ij}\Big|_0&=-2g(t)\,\cos^2\lambda_j+2a_j^2\sin^2\lambda_j.
\end{aligned}$$

[8]Physically, φ_j coincides with v_j+g_j where v_j and g_j are, respectively, the true anomaly and the argument of the perihelion of the osculating ellipse associated to the star and the jth planet; compare to Appendix A.

Therefore, using (3.25), one finds

$$\left\langle \partial^2_{\hat\eta_j\hat\xi_j} \frac{1}{\sqrt{\tilde{\mathcal{R}}_{ij}}}\bigg|_0 \right\rangle_{\lambda_i,\lambda_j} = \left\langle \frac{-\left(\frac{3}{2}g^2+2(a_j^2-g)f\right)\sin(2\lambda_j)}{4f^{5/2}}\bigg|_0 \right\rangle_{t,\lambda_j} = 0$$

(since the integrand is odd in λ_j), showing that also B_{ij} is a diagonal matrix. To compute the diagonal elements we calculate

$$\partial^2_{\hat\eta_j\hat\eta_j}\frac{1}{\sqrt{\tilde{\mathcal{R}}_{ij}}}\bigg|_0 = \frac{\tilde\rho_1(\lambda_j,t)}{\tilde\rho_2(\lambda_j,t)} \qquad \text{and} \qquad \partial^2_{\hat\xi_j\hat\xi_j}\frac{1}{\sqrt{\tilde{\mathcal{R}}_{ij}}}\bigg|_0 = \frac{\tilde\rho_3(\lambda_j,t)}{\tilde\rho_2(\lambda_j,t)} \tag{3.26}$$

with

$$\begin{aligned}
\tilde\rho_1 &= -7\,a_i^2\,a_j^2 + (9\,a_i^2\,a_j^2 + 8\,{a_j}^4)\,\cos(2\,\lambda_j) + \frac{7}{2}\,a_i^2\,a_j^2\,\cos(2\,\lambda_j - 2\,t)\\
&\quad - (2\,{a_i}^3\,a_j + 10\,a_i\,{a_j}^3)\,\cos(2\,\lambda_j - t) + 4({a_i}^3\,a_j + a_i\,{a_j}^3)\,\cos(t)\\
&\quad - a_i^2\,a_j^2\,\cos(2\,t) - (2\,{a_i}^3\,a_j + 10\,a_i\,{a_j}^3)\,\cos(2\,\lambda_j + t) + \frac{7}{2}\,a_i^2\,a_j^2\,\cos(2\,\lambda_j + 2\,t);\\
\tilde\rho_2 &= 4(a_i^2 + a_j^2 - 2a_ia_j\cos t)^{5/2};\\
\tilde\rho_3 &= -7\,a_i^2\,a_j^2 - (9\,a_i^2\,a_j^2 + 8\,{a_j}^4)\,\cos(2\,\lambda_j) - \frac{7}{2}\,a_i^2\,a_j^2\,\cos(2\,\lambda_j - 2\,t)\\
&\quad + (2\,{a_i}^3\,a_j + 10\,a_i\,{a_j}^3)\,\cos(2\,\lambda_j - t) + 4({a_i}^3\,a_j + a_i\,{a_j}^3)\,\cos(t)\\
&\quad - a_i^2\,a_j^2\,\cos(2\,t) + (2\,{a_i}^3\,a_j + 10\,a_i\,{a_j}^3)\,\cos(2\,\lambda_j + t) - \frac{7}{2}\,a_i^2\,a_j^2\,\cos(2\,\lambda_j + 2\,t);
\end{aligned}$$

taking the λ_j-average, one finds immediately

$$\left\langle \partial^2_{\hat\eta_j\hat\eta_j}\frac{1}{\sqrt{\tilde{\mathcal{R}}_{ij}}}\bigg|_0\right\rangle_{\lambda_i,\lambda_j} = \left\langle \partial^2_{\hat\xi_j\hat\xi_j}\frac{1}{\sqrt{\tilde{\mathcal{R}}_{ij}}}\bigg|_0\right\rangle_{\lambda_i,\lambda_j} = \frac{a_ia_j}{4}\mathcal{I}(a_i,a_j) =: \beta_{ij}. \qquad \square$$

The next result shows that, for δ and ε small, generically the eigenvalues of M in (3.6)–(3.7) are nonvanishing, simple, and distinct. We formulate the result regarding the semiaxis a_j as *independent variables*. Recall the definitions of α_{ij} and β_{ij} in (3.5) and let (if $N \geq 3$)

$$\beta_j := \sum_{k=1,2} \frac{\sqrt{\bar\mu_k\bar\mu_j}}{m_0}\,\frac{1}{\sqrt[4]{a_ka_j}}\,\beta_{kj}, \qquad j \geq 3. \tag{3.27}$$

PROPOSITION 3.2. *Assume that a_j and $\bar\mu_j$ verify*[9]

$$\alpha_{12} \neq 0 \quad \text{and} \quad \beta_{12} \neq \pm\alpha_{12}, \qquad \beta_i \neq 0 \quad \text{and} \quad \beta_i \neq \beta_j \quad \text{for} \quad i \neq j. \tag{3.28}$$

Then, there exist $0 < \delta^\star < 1$ and $0 < \varepsilon_0 < 1$ such that for all $0 < \delta < \delta^\star$ and $0 \leq \varepsilon < \varepsilon_0$ the eigenvalues $\{\bar\Omega_1, \dots, \bar\Omega_N\}$ of the matrix M are nonvanishing, simple,

[9]Clearly, if $N = 2$, the statements regarding the β_j and the eigenvalues $\bar\Omega_j$ for $j \geq 3$ have to be omitted.

and distinct. Furthermore the following asymptotics hold:[10]

$$\bar\Omega_1 = \frac{\sqrt{\bar\mu_1\bar\mu_2}}{m_0}\ \frac{\beta_{12}+\alpha_{12}}{\sqrt[4]{a_1 a_2}}\ + O(\sqrt{\delta},\varepsilon),$$

$$\bar\Omega_2 = \frac{\sqrt{\bar\mu_1\bar\mu_2}}{m_0}\ \frac{\beta_{12}-\alpha_{12}}{\sqrt[4]{a_1 a_2}}\ + O(\sqrt{\delta},\varepsilon),$$

$$\bar\Omega_j = \sqrt{\delta}\,\beta_j\ +\ \sqrt{\delta}\ O(\sqrt{\delta},\varepsilon),\qquad 3\le j\le N. \tag{3.29}$$

As mentioned above (see Remark 3.1(iv)) the asymptotic of the α_{ij}'s and β_{ij}'s may be evaluated in terms of the Laplace coefficients (see, e.g., [L91]). For completeness we give a detailed proof.

Proof. First of all, from the definition of c_{ij} (see (3.4) and (2.3)) it follows that

$$c_{ij} = \frac{1}{m_0}\ \frac{1}{\sqrt[4]{a_i a_j}} + O(\varepsilon). \tag{3.30}$$

Thus, by definition of M, by definition of β_j and α_{ij}, and by the hypothesis on the masses μ_i (see (1.3)) we find the following asymptotics:

$$\begin{aligned}
&\begin{pmatrix} M_{11} & M_{12}\\ M_{21} & M_{22}\end{pmatrix} = M^\star + O(\sqrt{\delta},\varepsilon), && \text{where}\quad M^\star := \frac{\sqrt{\bar\mu_1\bar\mu_2}}{m_0\sqrt[4]{a_1a_2}}\begin{pmatrix}\beta_{12} & \alpha_{12}\\ \alpha_{12} & \beta_{12}\end{pmatrix},\\
&M_{jj} = \sqrt{\delta}\beta_j + O(\delta,\varepsilon) && \text{for } j\ge 3,\\
&M_{ij} = O(\sqrt{\delta}) && \text{for } i=1,2 \text{ and } j\ge 3, \text{ or } j=1,2 \text{ and } i\ge 3,\\
&M_{ij} = O(\delta) && \text{for } i,j\ge 3 \text{ with } i\ne j.
\end{aligned}$$

Therefore[11]

$$M = \begin{pmatrix} M^\star + O(\sqrt{\delta},\varepsilon) & O(\sqrt{\delta})\\ O(\sqrt{\delta}) & \sqrt{\delta}M_\star + O(\delta,\varepsilon)\end{pmatrix},$$

where

$$M_\star := \operatorname{diag}(\beta_3,\dots,\beta_N) \in \mathrm{Mat}((N-2)\times(N-2)).$$

The eigenvalues of $M^\star$ are

$$\frac{\sqrt{\bar\mu_1\bar\mu_2}}{m_0\sqrt[4]{a_1a_2}}\ (\beta_{12}+\alpha_{12})\qquad\text{and}\qquad \frac{\sqrt{\bar\mu_1\bar\mu_2}}{m_0\sqrt[4]{a_1a_2}}\ (\beta_{12}-\alpha_{12}),$$

which, by the first two requirements in (3.28), are nonzero, simple, and distinct. The matrix $M_\star$ is diagonal and its eigenvalues β_j are also nonzero, simple, and distinct by (3.28). The claim now follows by elementary linear algebra (compare, e.g., Lemma B.2 in Appendix B). $\square$

[10] We use the standard notation $a = O(\varepsilon) \iff \exists$ a constant $c>0$ (independent of ε) and $0<\varepsilon_0<1$ s.t. $|a|\le c|\varepsilon|$ for all $|\varepsilon|\le\varepsilon_0$; $O(\sigma,\varepsilon) = O(\sigma)+O(\varepsilon)$.

[11] The $O(\sqrt{\delta})$ in the upper right part of M is a $(2\times(N-2))$ matrix while the $O(\sqrt{\delta})$ in the lower left part of M is an $((N-2)\times 2)$ matrix.

Remark 3.2. (i) The hypotheses (3.28) of Proposition 3.2 are easily checked, for example, if a_j verifies (1.2) for a suitable $\theta > 0$. In fact the asymptotics for $\mathcal{J}(s,1)$ and $\mathcal{I}(s,1)$ (see (3.8)) yield immediately

$$
\begin{aligned}
\alpha_{12} &= -\frac{15}{64}\,\frac{1}{a_2}\,\Big(\frac{a_1}{a_2}\Big)^3\ \Big[1+O\Big(\Big(\frac{a_1}{a_2}\Big)^2\Big)\Big],\\
\beta_{12} &= \frac{3}{4}\,\frac{1}{a_2}\,\Big(\frac{a_1}{a_2}\Big)^2\ \Big[1+O\Big(\Big(\frac{a_1}{a_2}\Big)^2\Big)\Big],\\
\beta_j &= \frac{3}{4}\,\frac{\sqrt{\bar\mu_j\bar\mu_2}}{m_0}\ \frac{1}{a_j^{3/2}}\ \Big(\frac{a_2}{a_j}\Big)^{7/4}\ \Big[1+O\Big(\Big(\frac{a_1}{a_2}\Big)^{7/4}\Big)+O\Big(\Big(\frac{a_2}{a_j}\Big)^2\Big)\Big], \qquad j\ge 3,\\
\beta_{12}\pm\alpha_{12} &= \frac{3}{4}\,\frac{1}{a_2}\,\Big(\frac{a_1}{a_2}\Big)^2\ \Big[1\mp\frac{5}{16}\,\frac{a_1}{a_2}+O\Big(\Big(\frac{a_1}{a_2}\Big)^2\Big)\Big],\\
\beta_j-\beta_i &= \frac{3}{4}\,\frac{\sqrt{\bar\mu_j\bar\mu_2}}{m_0}\ \frac{1}{a_j^{3/2}}\ \Big(\frac{a_2}{a_j}\Big)^{7/4}\\
&\qquad\times\Big[1+O\Big(\Big(\frac{a_1}{a_2}\Big)^{7/4}\Big)+O\Big(\Big(\frac{a_2}{a_j}\Big)^2\Big)+O\Big(\Big(\frac{a_i}{a_j}\Big)^{13/4}\Big)\Big], \qquad i>j\ge 3.
\end{aligned}
$$

Thus, if θ is small enough and if (1.2) holds, one sees that

$$
\begin{aligned}
&\alpha_{12} < 0,\\
&\beta_{12}\pm\alpha_{12} > 0,\\
&\beta_j > 0 \qquad \forall\ j\ge 3,\\
&\beta_j-\beta_i > 0 \qquad \forall\ i>j\ge 3,
\end{aligned}
$$

and the hypotheses (3.28) are verified as claimed.

(ii) The $O(\cdot)$'s appearing in (3.29) (and in the proof of Proposition 3.2) depend on the a_j's (and on[12] m_0). Thus, the order in fixing the various parameters is important. One way of proceeding is as follows. First determine θ as explained in the previous point (i). Then, let $\bar a_i$, $1\le i\le N$, be positive numbers such that (1.2) holds, i.e., $\bar a_i/\bar a_{i+1} < \theta$ for any $1\le i\le N-1$; (the $\bar a_i$ may be physically interpreted as observed mean major semiaxis). Now, consider a compact order-one neighborhood $A\subset\{0<a_1<\cdots<a_N\}$ of $(\bar a_1,\dots,\bar a_N)$ for which (1.2) continues to be valid (such neighborhood exists simply by continuity). Finally, fix $\delta^\star$ and ε_0 so that Proposition 3.2 holds: such numbers will depend only on $\bar a_j$'s and the (order-one) size of the chosen neighborhood A.

(iii) In the case of only *one* dominant planet (i.e., $\mu_1=\bar\mu_1=O(1)$, $\mu_i=O(\delta)$ for $i\ge 2$), the first two asymptotics in (3.29) do not give any information: in particular we cannot assure that $\bar\Omega_1$ and $\bar\Omega_2$ are different from zero. On the other hand, one could also consider the case of three or more dominant planets and the choice of focusing on *two* dominant planets has been made for simplicity.

4. Existence of -dimensional elliptic invariant tori. In this section we prove the existence of N-dimensional elliptic invariant tori for the $(N+1)$-body problem Hamiltonian $\mathcal{H}$ in (2.14) for any $N\ge 2$.

Let $m_0<\bar\mu_j<4m_0$, let θ, A, $\delta^\star$, and ε_0 be as in Remark 3.2(ii), and fix $0<\delta<\delta^\star$, which henceforth will be kept fixed. In the rest of the paper only ε is

[12]Recall that $m_0<\bar\mu_j<4m_0$; compare with the line before (1.3).

regarded as a free parameter: at the moment, ε is assumed not to exceed ε_0 but later will be required to satisfy stronger smallness conditions. The semimajor axis map

$$\vec{a} : \Lambda = (\Lambda_1, \ldots, \Lambda_N) \mapsto \big(a(\Lambda_1; \mu_1, \varepsilon), \ldots, a(\Lambda_N; \mu_N, \varepsilon)\big) \tag{4.1}$$

is a real-analytic diffeomorphism and we define

$$\mathfrak{I} = \vec{a}^{\,-1}(A),$$

then the Hamiltonian $\mathcal{H}$ is real-analytic (and bounded) on the domain $\mathfrak{I} \times \mathbf{T}^N \times B_R^{2N}$ for a suitable $R > 0$ (here B_r^n denotes the n-ball of radius r and center $0 \in \mathbf{R}^n$).

By Proposition 3.1, the quadratic part $\overline{\mathcal{H}}_{1,2}$ of the averaged Newtonian interaction $\mathcal{H}_1$ has the simple form (3.6), M being the symmetric matrix defined in (3.7). As already pointed out in Remark 3.1, the matrix M can be diagonalized with eigenvalues, which, thanks to our assumptions and to Proposition 3.2, have the form in (3.29) and, therefore, satisfy

$$\inf_{\mathfrak{I}} |\bar{\Omega}_j| > \bar{c}, \qquad \inf_{\mathfrak{I}} \left|\bar{\Omega}_i - \bar{\Omega}_j\right| > \bar{c} \tag{4.2}$$

for any $i \neq j = 1, \ldots, N$ and for a suitable positive constant $\bar{c}$ independent of ε. If $U := U(\Lambda)$ is the symmetric matrix which diagonalizes M, $U^T M U = \operatorname{diag}(\bar{\Omega}_1, \ldots, \bar{\Omega}_N)$, then the map

$$\Xi : (I, \varphi, p, q) \mapsto (\Lambda, \lambda, \eta, \xi), \qquad \text{where} \qquad \begin{cases} p = U^T \eta, \qquad q = U^T \xi, \\ I = \Lambda, \\ \varphi = \lambda + \displaystyle\sum_{h,k,\ell} (\partial_\Lambda U_{k\ell}) U_{h\ell}\, \eta_k\, \xi_\ell, \end{cases} \tag{4.3}$$

is symplectic (and real-analytic) and

$$\overline{\mathcal{H}}_{1,2} \circ \Xi = \frac{1}{2} \sum_{i=1}^N \bar{\Omega}_i(I)(p_i^2 + q_i^2). \tag{4.4}$$

Thus, the $(N+1)$-body problem Hamiltonian $\mathcal{H}$ in (2.14), in the case we are considering, can be written as

$$\mathcal{H} \circ \Xi(I, \varphi, p, q; \varepsilon) = h(I) + f(I, \varphi, p, q; \varepsilon) \tag{4.5}$$

with

$$h := \mathcal{H}_0, \qquad f := \varepsilon f_1(I, p, q; \varepsilon) + \varepsilon f_2(I, \varphi, p, q; \varepsilon),$$

$$f_1 := f_{1,0}(I) + \frac{1}{2} \sum_{i=1}^N \bar{\Omega}_i(I)(p_i^2 + q_i^2) + \tilde{f}_1(I, p, q; \varepsilon),$$

$$f_{1,0} := \overline{\mathcal{H}}_{1,0}, \qquad \tilde{f}_1 := \overline{\mathcal{H}}_{1,*} \circ \Xi, \qquad f_2 := \widetilde{\mathcal{H}}_1 \circ \Xi.$$

Here h is uniformly strictly concave,

$$|\tilde{f}_1| \leq \text{const}\, |(p, q)|^4, \qquad \text{and} \qquad \int_{\mathbf{T}^N} f_2 \, d\varphi = 0.$$

The construction of elliptic invariant tori for the Hamiltonian (4.5) is based on four steps, which we proceed to describe.

4.1. Averaging. Fix $\tau > N-1 \geq 1$ and pick two numbers b_1, b_2 such that

$$0 < b_1 < \frac{1}{2}, \qquad 0 < b_2 < \Big(\frac{1}{2} - b_1\Big)\frac{1}{\tau+1}. \tag{4.6}$$

Since the integrable Hamitlonian h depends only on the action I, the conjugated variable φ is a "fast" angle and, in "first approximation," the $(h+f)$-motions are governed by the averaged Hamiltonian $h+\varepsilon f_1$, which possesses an elliptic equilibrium at $p = q = 0$. As we, now, proceed to describe, one may remove the φ-dependence of the perturbation function f up to high order in ε by using averaging theory; for detailed information on averaging theory in similar situations, see Proposition A.1 of [BCV03] or Proposition 7.1 of [BBV04].

Denote by D^n_R the complex n-ball of center zero and radius $R > 0$ and, for any $V \subset \mathbf{R}^N$, denote by V_R the complex neighborhood of radius $R > 0$ of the set V given by $V_R := \cup_{x\in V} D_R(x)$. Next, define the set $\hat{\mathfrak{I}}$ as the following "Diophantine subset" of $\mathfrak{I}$:

$$\hat{\mathfrak{I}} := \left\{ I \in \mathfrak{I} \,:\, |\partial_I h(I)\cdot k| \geq \frac{\bar\gamma}{|k|^\tau}\ \forall k \in \mathbf{Z}^N \setminus \{0\} \right\} \quad \text{with} \quad \bar\gamma := \text{const}\ \varepsilon^{b_1}. \tag{4.7}$$

Notice that (as it is standard to prove)

$$\text{meas}\left(\mathfrak{I} \setminus \hat{\mathfrak{I}}\right) \leq \text{const}\,\bar\gamma = \text{const}\,\varepsilon^{b_1}. \tag{4.8}$$

The Hamiltonian $h+f$ in (4.5) is real-analytic on the complex domain

$$\mathfrak{D}_{\mathrm{r},s,\rho} := \hat{\mathfrak{I}}_{\mathrm{r}} \times \mathbf{T}^N_s \times D^{2N}_\rho \subset \mathbf{C}^{4N}, \tag{4.9}$$

with

$$\mathrm{r} := \text{const}\,\sqrt{\varepsilon}, \qquad s := \text{const}, \qquad \rho := \text{const}\,\varepsilon^{b_2}. \tag{4.10}$$

The definition of $\hat{\mathfrak{I}}$ is motivated by the necessity to have an estimate on small divisors. In fact, let $I \in \hat{\mathfrak{I}}_{\mathrm{r}}$ (and ε small enough) and let $I_0 \in \hat{\mathfrak{I}}$ be a point at distance less than r from I. Then, for any $k \in \mathbf{Z}^N\backslash\{0\}$ such that $|k| \leq K := \text{const}\,\varepsilon^{-b_2}$, by the second relation in (4.6), by (4.7), and by Cauchy estimates, one finds

$$\begin{aligned}|\partial_I h(I)\cdot k| &\geq |\partial_I h(I_0)\cdot k| - |\partial_I h(I_0) - \partial_I h(I)|\ |k| \\ &\geq \frac{\bar\gamma}{K^\tau} - \max|\partial^2_I h|\ r\ K \\ &\geq \frac{\bar\gamma}{2K^\tau} =: \alpha = \text{ const}\,\varepsilon^{b_1+\tau b_2}, \qquad \Big(0 < |k| \leq K := \text{const}\,\varepsilon^{-b_2}\Big).\end{aligned} \tag{4.11}$$

In order to apply averaging theory (see, e.g., [N77]) so as to remove the ε-dependence up to order $\exp(-\,\text{const}\,K)$, one has to verify the following "smallness condition" (compare condition (A.2), p. 110 in [BCV03])

$$\|f\|_{\mathrm{r},s,\rho} \leq \text{const}\,\frac{\alpha\min\{\mathrm{r}s, \rho^2\}}{K},$$

where the norm $\|\cdot\|_{\mathrm{r},s,\rho}$ is defined as the standard "sup-Fourier norm"

$$\|f\|_{\mathrm{r},s,\rho} := \sum_{k\in\mathbf{Z}^N} \Big(\sup_{(I,p,q)\in\hat{\mathfrak{I}}_{\mathrm{r}}\times D^{2N}_\rho} |f_k(I,p,q)| \Big) e^{|k|s}, \tag{4.12}$$

($f_k(I,p,q)$ denoting Fourier coefficients of the multiperiodic, real-analytic function $\varphi \mapsto f(I,\varphi,p,q)$). Such condition, in view of (4.6), can be achieved by taking ε small enough since, by (4.11) and (4.10), one has

$$\|f\|_{\mathrm{r},s,\rho} = O(\varepsilon) \qquad \text{and} \qquad \frac{\alpha \min\{\mathrm{r}s, \rho^2\}}{K} = O(\varepsilon^{b_1+(\tau+1)b_2+1/2}).$$

Hence, there exists a close-to-identity (real-analytic) symplectic change of variables $(I',\varphi',p',q') \mapsto (I,\varphi,p,q)$ verifying (compare formulae (2.16) and (A.7) of [BCV03])

$$|I'-I| \le \text{ const}\,\varepsilon^{\frac{1}{2}+b_2} \qquad \text{and} \qquad |p'-p|, |q'-q| \le \text{const}\,\sqrt{\varepsilon}, \tag{4.13}$$

and such that the Hamiltonian expressed in the new symplectic variables becomes

$$h(I') + \hat{g}(I',p',q') + \hat{f}(I',\varphi',p',q'), \qquad \hat{g} := \varepsilon f_1(I',p',q') + \varepsilon \hat{f}_1(I',p',q') \tag{4.14}$$

with $\hat{f}_1$ and $\hat{f}$ real-analytic on the complex domain $\mathfrak{D}_{\mathrm{r}/2,s/6,\rho/2}$ and satisfying

$$\begin{aligned} &\|\hat{f}_1\|_{\mathrm{r}/2,s/6,\rho/2} \le \frac{\varepsilon}{\alpha \mathrm{r}} = \text{ const}\,\varepsilon^{b_2+b_3} \quad \text{with} \quad b_3 := \frac{1}{2} - b_1 - (\tau+1)b_2 > 0, \\ &\|\hat{f}\|_{\mathrm{r}/2,s/6,\rho/2} \le \text{ const}\,e^{-\text{ const}\,K} \ll \text{ const}\,\varepsilon^3. \end{aligned} \tag{4.15}$$

4.2. New elliptic equilibrium. Due to the (small) term $\hat{f}_1$ in (4.14), zero is no longer an elliptic equilibrium for the "averaged" (i.e., φ-independent) Hamiltonian $h+\hat{g}$. Using the implicit function theorem, we can find a new elliptic equilibrium for $h+\hat{g}$, which is $\varepsilon^{b_2+b_3}$ close to zero. Hence we construct a real-analytic symplectic transformation

$$(J',\psi',v',u') \mapsto (I',\varphi',p',q') \quad \text{with} \quad I' = J' \quad \text{and } \varepsilon^{b_2+b_3}\text{-close-to-the-identity}, \tag{4.16}$$

such that in the new symplectic variables (J',ψ',v',u') the Hamiltonian takes the form

$$h(J') + \tilde{g}(J',v',u') + \tilde{f}(J',\psi',v',u')$$

with $\tilde{g}$ having $v'=u'=0$ as elliptic equilibrium; the functions $\tilde{g}$ and $\tilde{f}$ are real-analytic on a slightly smaller complex domain, say $\mathfrak{D}_{\mathrm{r}/7,s/7,\rho/7}$, where they satisfy bounds similar to those in (4.15). Furthermore, for $j=1,\dots,N$, the eigenvalues $\tilde{\Omega}_j(J')$ of the symplectic quadratic part of $\tilde{g}$ are purely imaginary and $\varepsilon^{1+b_2+b_3}$-close to $\varepsilon\bar{\Omega}_j(J')$.

4.3. Symplectic diagonalization of the quadratic term. Using a well-known result on the symplectic diagonalization of quadratic Hamiltonians, we can find a real-analytic, symplectic transformation

$$(\tilde{J},\tilde{\psi},\tilde{v},\tilde{u}) \mapsto (J',\psi',v',u') \quad \text{with} \quad J' = \tilde{J} \quad \text{and } \varepsilon^{b_2+b_3}\text{-close-to-the-identity}, \tag{4.17}$$

such that the quadratic part of $\tilde{g}$ becomes, simply, $\sum_{i=1}^N \tilde{\Omega}_i(\tilde{J})\,(\tilde{u}_j^2 + \tilde{v}_j^2)$. Whence, the new Hamiltonian becomes (compare formula (2.22) of [BCV03])

$$\widetilde{\mathcal{H}} := h_0(\tilde{J}) + \sum_{i=1}^N \tilde{\Omega}_i(\tilde{J})\,(\tilde{u}_i^2 + \tilde{v}_i^2) + \tilde{g}_0(\tilde{J},\tilde{v},\tilde{u}) + \tilde{f}_0(\tilde{J},\tilde{\psi},\tilde{v},\tilde{u}), \tag{4.18}$$

where

$$h_0(\tilde{J}) := h(\tilde{J}) + \varepsilon \tilde{g}(\tilde{J}, 0, 0), \tag{4.19}$$

$\tilde{g}_0$, $\tilde{f}_0$, $\tilde{\Omega}_j$ are real-analytic, and

(4.20)

$$|\tilde{g}_0(\tilde{J}, \tilde{v}, \tilde{u})| \leq \text{ const } \varepsilon\, |(\tilde{v}, \tilde{u})|^3, \qquad |\tilde{\Omega}| \leq \text{ const}\, \varepsilon, \qquad \|\tilde{f}\|_{\mathrm{r}/8, s/8, \rho/8} \leq \text{ const}\, \varepsilon^3.$$

Finally, because of (4.2),

$$\inf |\tilde{\Omega}_i| \geq \text{ const } \varepsilon > 0, \qquad \inf |\tilde{\Omega}_2 - \tilde{\Omega}_1| \geq \text{ const } \varepsilon > 0. \tag{4.21}$$

4.4. Applying KAM theory. We rewrite now the Hamiltonian $\widetilde{\mathcal{H}}$ in (4.18) in a form suitable for applying (elliptic) KAM theory. Introducing translated variables $y := \tilde{J} - \mathfrak{p}$ and complex variables $z, \bar{z}$, we define

$$H = \widetilde{\mathcal{H}}\Big(\mathfrak{p} + y, \psi, \frac{z + l\bar{z}}{\sqrt{2}}, \frac{z - \bar{z}}{\mathrm{i}\sqrt{2}}\Big), \tag{4.22}$$

here $\mathfrak{p}$ is regarded as a parameter and the symplectic form is $\sum_{j=1}^N dy_j \wedge d\psi_j + \mathrm{i} \sum_{j=1}^N dz_j \wedge d\bar{z}_j$ with $\mathrm{i} := \sqrt{-1}$. The Hamiltonian H is then seen to have the form

$$H = \mathcal{N} + P$$

with

$$\mathcal{N} = e + \omega \cdot y + \sum_{j=1}^N \Omega_j z_j \bar{z}_j, \qquad e := h_0(\mathfrak{p}), \quad \omega := \partial_{\tilde{J}} h_0(\mathfrak{p}), \quad \Omega := \tilde{\Omega}(\mathfrak{p}), \tag{4.23}$$

and P a perturbation, which can naturally be split into four terms:

$$P = \sum_{1 \leq k \leq 4} P_k$$

with

$$\begin{aligned}
P_1 &= h_0(\mathfrak{p} + y) - h_0(\mathfrak{p}) - \partial_{\tilde{J}} h_0(\mathfrak{p}) \cdot y \sim y^2, \\
P_2 &= \sum_{j=1}^n \Big(\tilde{\Omega}_j(\mathfrak{p} + y) - \tilde{\Omega}_j(\mathfrak{p})\Big)\, z_j \bar{z}_j \sim y|z||\bar{z}|, \\
P_3 &= \tilde{g}_0\Big(\mathfrak{p} + y, \frac{z + \bar{z}}{\sqrt{2}}, \frac{z - \bar{z}}{\mathrm{i}\sqrt{2}}\Big) \sim \varepsilon(|z| + |\bar{z}|)^3, \qquad (\text{by } \ (4.20)), \\
P_4 &= \tilde{f}_0\Big(\mathfrak{p} + y, \psi, \frac{z + \bar{z}}{\sqrt{2}}, \frac{z - \bar{z}}{\mathrm{i}\sqrt{2}}\Big) = O(\varepsilon^3).
\end{aligned} \tag{4.24}$$

The parameter $\mathfrak{p}$ runs over the Diophantine set $\hat{\mathfrak{J}}$ defined in (4.7). Notice that the integrable Hamiltonian $\mathcal{N}$ affords, for any given value of the parameter $\mathfrak{p}$, the N-*dimensional elliptic torus*

$$\{y = 0\} \times \mathbf{T}^N \times \{z = \bar{z} = 0\}, \tag{4.25}$$

which is invariant for the Hamiltonian flow generated by $\mathcal{N}$, the flow being, simply, the Diophantine translation $x \mapsto x + \omega t$, with ω as in (4.23).

Since $\det \partial^2_{\bar{J}} h_0 \neq 0$, we can use the frequencies ω as parameters rather than the actions $\mathfrak{p}$. We, therefore, set

$$\mathcal{O} := \partial_{\bar{J}} h_0(\hat{\mathfrak{I}}) = \left\{ \omega = \partial_{\bar{J}} h_0(\mathfrak{p}) \ : \ \mathfrak{p} \in \hat{\mathfrak{I}} \right\}. \tag{4.26}$$

Notice that, by (4.7), we have

$$\operatorname{meas}\Big(\partial_{\bar{J}} h_0(\mathfrak{I}) \setminus \mathcal{O} \Big) \leq \text{const } \varepsilon^{b_1}. \tag{4.27}$$

Now, if we put $\mathfrak{p} = \mathfrak{p}(\omega) := (\partial_{\bar{J}} h_0)^{-1}(\omega)$ in (4.22), we can rewrite the $(N+1)$-body Hamiltonian in the form

$$H(y, \psi, z, \overline{z}; \omega) := \mathcal{N}(y, z, \bar{z}; \omega) + P(y, \psi, z, \bar{z}; \omega), \tag{4.28}$$

where

(4.29)

$$\mathcal{N}(y, z, \overline{z}; \omega) := e(\omega) + \omega \cdot y + \sum_{j=1}^{N} \Omega_j(\omega) z_j \bar{z}_j, \qquad e(\omega) := h_0(\mathfrak{p}(\omega)), \qquad \Omega(\omega) := \widetilde{\Omega}(\mathfrak{p}(\omega)),$$

and the perturbation $P(y, \psi, z, \bar{z}; \omega)$ is obtained by replacing $\mathfrak{p}$ with $\mathfrak{p}(\omega)$ in (4.24). Recalling (4.10), the Hamiltonian H in (4.28) is real-analytic in

$$(y, \psi, z, \bar{z}; \omega) \in \mathfrak{D}_{\mathfrak{r}^2, \mathfrak{s}, \mathfrak{r}, \mathfrak{d}} := D^N_{\mathfrak{r}^2} \times \mathbf{T}^N_{\mathfrak{s}} \times D^{2N}_{\mathfrak{r}} \times \mathcal{O}^N_{\mathfrak{d}} \tag{4.30}$$

with

$$\mathfrak{r} := \text{const } , \varepsilon^{3/4}, \qquad \mathfrak{s} := \text{const}, \qquad \mathfrak{d} := \text{const} \sqrt{\varepsilon}. \tag{4.31}$$

We recall, now, a well-known KAM result concerning the persistence of lower-dimensional elliptic tori for nearly integrable Hamiltonian systems (see [M65], [E88], [K88]). The version we present here is, essentially, a reformulation of Pöschel's theorem in [P89] (compare, also, with Theorem 5.1 of [BBV04]).

THEOREM 4.1. *Let H have the form in* (4.28), (4.29) *and let it be real-analytic on a domain $\mathfrak{D}_{\mathfrak{r}^2, \mathfrak{s}, \mathfrak{r}, \mathfrak{d}}$ of the form* (4.30) *for some $\mathfrak{r}$, $\mathfrak{s}$, and $\mathfrak{d}$ positive. Assume that*

$$\sup_{\omega \in \mathcal{O}_{\mathfrak{d}}} |\partial_\omega \Omega(\omega)| \leq \frac{1}{4} \tag{4.32}$$

and that the nonresonance (or Melnikov) condition

$$|\Omega(\omega) \cdot k| \geq \gamma_0 \qquad \forall\ 1 \leq |k| \leq 2,\ k \in \mathbf{Z}^N, \forall\ \omega \in \mathcal{O}, \tag{4.33}$$

is satisfied for some $\gamma_0 > 0$. Then, if $\mathfrak{d} \geq \gamma_0$ and P is sufficiently small, i.e.,

$$\|P\|_{\mathfrak{r}, \mathfrak{s}, \mathfrak{d}} := \sup_{\omega \in \mathcal{O}_{\mathfrak{d}}} \|P(\,\cdot; \omega)\|_{\mathfrak{r}^2, \mathfrak{s}, \mathfrak{r}} \leq \text{const}\, \gamma_0\, \mathfrak{r}^2, \tag{4.34}$$

then there exist a normal form $\mathcal{N}_* := e_*(\omega) + \omega \cdot y_* + \Omega_*(\omega) z_* \bar{z}_*$, a Cantor set $\mathcal{O}(\gamma_0) \subset \mathcal{O}$ with

$$\operatorname{meas}\big(\mathcal{O} \setminus \mathcal{O}(\gamma_0)\big) \leq \text{const}\, \gamma_0, \tag{4.35}$$

and a transformation

$$\mathcal{F} :\ D^N_{\mathfrak{r}^2/4} \times \mathbf{T}^N_{\mathfrak{s}/2} \times D^{2N}_{\mathfrak{r}/2} \times \mathcal{O}(\gamma_0) \longrightarrow D^N_{\mathfrak{r}^2} \times \mathbf{T}^N_{\mathfrak{s}} \times D^{2N}_{\mathfrak{r}} \times \mathcal{O}_{\mathfrak{d}}$$
$$(y_*, \psi_*, z_*, \bar{z}_*; \omega) \longmapsto (y, \psi, z, \bar{z}; \omega)$$

real-analytic and symplectic for each ω and Whitney smooth in ω, such that

$$H \circ \mathcal{F} = \mathcal{N}_* + R_* \qquad \text{with} \quad \partial^j_{y_*} \partial^h_{z_*} \partial^k_{\bar{z}_*} R_* = 0 \quad \text{if} \quad 2|j| + |h+k| \le 2. \tag{4.36}$$

In particular, for each $\omega \in \mathcal{O}(\gamma_0)$, the torus $\{y_* = 0\} \times \mathbf{T}^N \times \{z_* = \bar{z}_* = 0\}$ is an N-dimensional, linearly elliptic, invariant torus run by the flow $\psi_* \to \psi_* + \omega t$. Finally

$$|y_* - y|, \quad \mathfrak{r}|z_* - z|, \quad \mathfrak{r}|\bar{z}_* - \bar{z}| \le \text{ const } \frac{\|P\|_{\mathfrak{r},\mathfrak{s},\mathfrak{d}}}{\gamma_0}. \tag{4.37}$$

In this section we have shown that the many-body Hamiltonian (2.14) (under the hypotheses spelled out at the beginning of the section) has indeed the form assumed in the KAM theorem (Theorem 4.1). Furthermore, by (4.21), the elliptic frequencies Ω_i verify the Melnikov conditions (4.33) with

$$\gamma_0 = \text{const } \varepsilon, \tag{4.38}$$

and, by (4.24) and (4.31), the perturbation P verifies, for small ε, the KAM condition (4.34), since

$$\|P\|_{\mathfrak{r},\mathfrak{s},\mathfrak{d}} = O(\mathfrak{r}^4 + \varepsilon\mathfrak{r}^3 + \varepsilon^3) = O(\varepsilon^3) \le \text{const}\, \gamma_0\, \mathfrak{r}^2 = O(\varepsilon^{5/2}). \tag{4.39}$$

Thus, the existence of the desired quasi-periodic orbits follows at once from Theorem 4.1. We may summarize the final result as follows.

THEOREM 4.2. *Let $N \ge 2$ and let $\mathcal{H}$ be the $(N+1)$-body problem Hamiltonian in Poincaré variables defined in* (2.14). *Let $m_0 < \bar{\mu}_j < 4m_0$, let θ, A, $\delta^\star$, and ε_0 be as in Remark* 3.2*(ii). Fix $0 < \delta < \delta^\star$ and let $\mathfrak{I} = \vec{a}^{-1}(A)$ where $\vec{a}$ is the semimajor axis map defined in* (4.1). *Let $\tau > N - 1$ and pick b_1, b_2 as in* (4.6). *Finally, let $0 < \varepsilon^\star < \varepsilon_0$ be such that* (4.39) *holds for any $\varepsilon \le \varepsilon^\star$ and such that all conditions on ε required for constructing the symplectic transformations introduced in sections* 4.1–4.3 *are satisfied for $\varepsilon < \varepsilon^\star$. Then, for any $\varepsilon < \varepsilon^\star$, there exist a Cantor set $\mathfrak{I}_* \subset \mathfrak{I}$, with*

$$\text{meas}\,(\mathfrak{I} \setminus \mathfrak{I}_*) \le \text{const}\, \varepsilon^{b_1}, \tag{4.40}$$

and a Lipschitz continuous family of tori embedding

$$\phi\ :\ (\vartheta, \mathfrak{p}) \in \mathbf{T}^N \times \mathfrak{I}_* \mapsto \Big(\Lambda(\vartheta; \mathfrak{p}),\ \lambda(\vartheta; \mathfrak{p}),\ \eta(\vartheta; \mathfrak{p}),\ \xi(\vartheta; \mathfrak{p}) \Big) \in \mathfrak{I} \times \mathbf{T}^N \times B^{2N}_{\rho_*}$$

with $\rho_* := \text{const}\, \varepsilon^{b_2}$ such that, for any $\mathfrak{p} \in \mathfrak{I}_*$, $\phi(\mathbf{T}^N; \mathfrak{p})$ is a real-analytic elliptic $\mathcal{H}$-invariant torus, on which the $\mathcal{H}$-flow is analytically conjugated to the linear flow $\vartheta \to \vartheta + \omega_* t$, ω_* being (γ, τ)-diophantine with $\gamma = O(\varepsilon^{b_1})$. Furthermore, the following bounds hold uniformly on $\mathbf{T}^N \times \mathfrak{I}_*$:

$$|\Lambda(\vartheta; \mathfrak{p}) - \mathfrak{p}| \le \text{const}\, \varepsilon^{\frac{1}{2} + b_2}, \tag{4.41}$$

$$|\eta(\vartheta; \mathfrak{p})| + |\xi(\vartheta; \mathfrak{p})| \le \text{const}\, \varepsilon^{b_2}. \tag{4.42}$$

Theorem 1.1 follows, now, by taking (recall the definitions of b_k in (4.6))

$$c_1 := b_1, \qquad c_2 := b_2, \qquad c_3 := b_2 + \frac{1}{2}. \tag{4.43}$$

In particular the statements on the density of the set of the osculating major semiaxes, on the bound on the osculating eccentricities and on the variation of the osculating major semiaxes, follows from (4.40), (2.9), (4.42), (2.7), and (4.41).

Appendix A. Poincaré variables for the planar (+ 1)-body problem. We briefly recall in this appendix the classical derivation of the Poincaré variables for the planar N-body problem,[13] showing, in particular, the validity of Lemma 2.1, which is proven in subsections A.1 and A.2; subsections A.3 and A.4 are included for completeness.

A.1. Canonical variables for the two-body problem. Consider two bodies P_0, P_1 of masses $\mathfrak{m}_0$, $\mathfrak{m}_1$ and position $u^{(0)}$, $u^{(1)} \in \mathbf{R}^2$ (with respect to an inertial frame). We assume that P_0 and P_1 interact through gravity, with gravitational constant 1. By Newton's laws, the equations of motion for such two-body problem are

$$\ddot{u}^{(0)} = \mathfrak{m}_1 \frac{(u^{(1)} - u^{(0)})}{|u^{(1)} - u^{(0)}|^3},$$
$$\ddot{u}^{(1)} = \mathfrak{m}_0 \frac{(u^{(0)} - u^{(1)})}{|u^{(0)} - u^{(1)}|^3}.$$

Let

$$\mathfrak{M} := \mathfrak{m}_0 + \mathfrak{m}_1, \quad \mathfrak{m} := \frac{\mathfrak{m}_0 \mathfrak{m}_1}{\mathfrak{M}}, \quad x := u^{(1)} - u^{(0)}, \quad X := \mathfrak{m}\dot{x}. \tag{A.1}$$

Then, the above equations of motion become

$$\ddot{x} = \frac{\mathfrak{M}x}{|x|^3},$$

and the motion of the two bodies is governed by the Hamiltonian

$$\mathcal{K}(X,x) = \frac{1}{2\mathfrak{m}}|X|^2 - \frac{\mathfrak{m}\mathfrak{M}}{|x|}, \tag{A.2}$$

with $(X,x) \in \mathbf{R}^2 \times \mathbf{R}^2$ conjugate variables; i.e., the equations of motion are $\dot{x} = \partial_X \mathcal{K}$, $\dot{X} = -\partial_x \mathcal{K}$.

As well known, such system is integrable and for $\mathcal{K} < 0$ the orbits are ellipses. More precisely, one has the following proposition.

Proposition A.1. *Fix $\Lambda_- > 0 > \mathcal{K}_0$ and let $\Lambda_+ := \left(\frac{\mathfrak{m}^3\mathfrak{M}^2}{-2\mathcal{K}_0}\right)^{\frac{1}{2}} > \Lambda_-$. Then, there exist $\hat{\rho} > 0$ and a real-analytic symplectic transformation*

$$\Psi_{\mathrm{DP}} : \Big((\Lambda,\eta),(\lambda,\xi)\Big) \in \Big([\Lambda_-,\Lambda_+] \times [-\hat{\rho},\hat{\rho}]\Big) \times \Big(\mathbf{T} \times [-\hat{\rho},\hat{\rho}]\Big) \mapsto (X,x) \in \Big\{|x| \geq \frac{\hat{\rho}^2}{\mathfrak{m}^2\mathfrak{M}}\Big\},$$

casting (A.2) *into the integrable Hamiltonian* $(-\mathfrak{m}^3\mathfrak{M}^2)/(2\Lambda^2)$.

[13]For a review of the Poincaré variables in the nonplanar case, see, for instance, [Ch88] and [BCV03].

This classical proposition is a planar version of the classical one, due to Poincaré (see [Poi1905, Chapter III]) and the variables $(\Lambda, \eta, \lambda, \xi)$ are, usually, called (planar) Poincaré variables. The proof of Proposition A.1 is particularly interesting from the physical point of view and rests upon the introduction of three different (famous) changes of variables, which we, now, proceed to describe briefly.

Let ℓ and g denote, respectively, the mean anomaly and the argument of the perihelion.

Step 1. The system is set in *"symplectic" polar variables*; namely, we consider the symplectic map $\Psi_{\rm spc} : ((R,\Phi),(r,\varphi)) \mapsto (X,x)$ (where $r>0$ and $\varphi \in \mathbf{T}$) given by

$$\Psi_{\rm spc} : \begin{cases} x_1 = r\ \cos\varphi, \\ x_2 = r\ \sin\varphi, \end{cases} \qquad X = \begin{pmatrix} \cos\varphi & -\dfrac{\sin\varphi}{r} \\ \sin\varphi & \dfrac{\cos\varphi}{r} \end{pmatrix} \begin{pmatrix} R \\ \Phi \end{pmatrix} \tag{A.3}$$

and consider the new Hamiltonian $\mathcal{K}_{\rm spc} := \mathcal{K} \circ \Psi_{\rm spc}$.

Step 2. There is a symplectic map $\Psi_{\rm D} : ((L,G),(\ell,g)) \mapsto ((R,\Phi),(r,\varphi))$ that integrates the system: $\Psi_{\rm D}$ is obtained via the generating function

$$S(L,G,r,\varphi) = \int \sqrt{-\frac{\mathfrak{m}^4\mathfrak{M}^2}{L^2} + \frac{2\mathfrak{m}^2\mathfrak{M}}{r} - \frac{G^2}{r^2}}\ dr\ +\ G\,\varphi. \tag{A.4}$$

The variables $((L,G),(\ell,g))$ are known as (planar) *Delaunay variables*. In such variables, the new Hamiltonian becomes

$$\mathcal{K}_{\rm D} := \mathcal{K}_{\rm spc} \circ \Psi_{\rm D} = -\frac{\mathfrak{m}^3\mathfrak{M}^2}{2L^2}.$$

Also, if C is the angular momentum of the planet and a is the major semiaxis, by construction, one has that

$$G = |C| \qquad \text{and} \qquad L = \mathfrak{m}\sqrt{\mathfrak{M}a}.$$

Step 3. We need now to remove singularities, which appear for small eccentricity. To this aim, we first introduce (planar) *Poincaré action-angle variables* by means of the linear symplectic transformation

$$\Psi_{\rm P_{aa}} : \Big((\Lambda,H),(\lambda,h)\Big) \mapsto \Big((L,G),(\ell,g)\Big)$$

given by

$$\Psi_{\rm P_{aa}} : \quad \begin{cases} \Lambda = L, \quad H = L - G, \\ \lambda = \ell + g, \quad h = -g. \end{cases} \tag{A.5}$$

Then, we let $\Psi_{\rm P} : ((\Lambda,\eta),(\lambda,\xi)) \mapsto ((\Lambda,H),(\lambda,h))$ be the symplectic map defined by

$$\sqrt{2H}\ \cos h = \eta, \qquad \sqrt{2H}\ \sin h = \xi. \tag{A.6}$$

As Poincaré showed (see [Poi1905], [Ch88], [BCV03]), the symplectic map

$$\Psi_{\rm DP} : \Big((\Lambda,\eta),(\lambda,\xi)\Big) \mapsto (X,x)$$

with

$$\Psi_{\rm DP} := \Psi_{\rm spc} \circ \Psi_{\rm D} \circ \Psi_{\rm P_{aa}} \circ \Psi_{\rm P} \tag{A.7}$$

is real-analytic in a (complex) neighborhood of

$$\Lambda \in [\Lambda_-, \Lambda_+], \qquad |\eta|, |\xi| \le \text{const}\sqrt{\Lambda_-}, \qquad \lambda \in \mathbf{T}.$$

Also, the two-body Hamiltonian, in Poincaré variables, is $\mathcal{K} \circ \Psi = -\frac{\mathfrak{m}^3\mathfrak{M}^2}{2\Lambda^2}$.

Remark A.1. (i) If we denote $(X, x) = \Phi_{\rm DP}((\Lambda, \eta, {\rm p}), (\lambda, \xi, {\rm q}))$, then

$$X = \frac{\mathfrak{m}^4\mathfrak{M}^2}{\Lambda^3}\frac{\partial x}{\partial \lambda}. \tag{A.8}$$

Indeed, from the Hamilton equations one sees that: $\dot\lambda = \partial_\Lambda(-\frac{\mathfrak{m}^3\mathfrak{M}^2}{2\Lambda^2}) = \frac{\mathfrak{m}^3\mathfrak{M}^2}{\Lambda^3}$, and $\dot\Lambda = \dot\xi = \dot\eta = \dot{\rm p} = \dot{\rm q} = 0$. Thus, by the chain rule, $X = \mathfrak{m}\dot x = \mathfrak{m}(\partial_\lambda x)\,\dot\lambda = \frac{\mathfrak{m}^4\mathfrak{M}^2}{\Lambda^3}\frac{\partial x}{\partial\lambda}$, proving (A.8).

(ii) We collect some useful relations among the above-introduced quantities. Let, as usual, e denote the eccentricity of the Keplerian ellipse and let a denote the major semiaxis. Then, by construction, one sees that

$$\Lambda = \mathfrak{m}\sqrt{\mathfrak{M}a}, \qquad \sqrt{\xi^2+\eta^2} = \sqrt{\Lambda}\, e\,(1+O(e^2)). \tag{A.9}$$

Also, if C is the angular momentum of the system, one infers that

$$|C| = \Lambda\sqrt{1-e^2} = \Lambda(1+O(e^2)). \tag{A.10}$$

(iii) A proof of the analyticity of Poincaré variables will also follow by directly inspecting the formulae given in Lemma 2.1, which is proved in the coming section.

A.2. Orbital elements. We now sketch a way to explicitly represent some quantities in terms of Poincaré variables. This will also lead to the proof of Lemma 2.1. Let u and v denote the eccentric anomaly and the true anomaly, respectively. By geometric considerations,

$$u = \ell + e\sin u \tag{A.11}$$

and[14]

$$\cos v = \frac{\cos u - e}{1 - e\cos u}, \tag{A.12}$$

where

$$\ell = \lambda + h. \tag{A.13}$$

Also, by (A.6),

$$H = \frac{\eta^2+\xi^2}{2}. \tag{A.14}$$

An explicit expression taking into account H, the eccentricity, and the major semiaxis is given by

$$H = \Lambda\,(1-\sqrt{1-e^2}) = \Lambda\,\frac{e^2}{2}(1+O(e^2)), \tag{A.15}$$

$$e(H,\Lambda) = \sqrt{\frac{H}{\Lambda}\left(2-\frac{H}{\Lambda}\right)}. \tag{A.16}$$

[14]Such relations are classical and we refer the reader to [Ch88] and [BCV03] for a geometric interpretation of these anomalies.

In light of (A.12),

$$\sin v = \frac{\sqrt{1-e^2}\sin u}{1-e\cos u}. \tag{A.17}$$

By means of (A.11), we have

$$u - \ell = e\,\sin(u-\ell+\ell) = e\,\cos\ell\,\sin(u-\ell) + e\,\sin\ell\,\cos(u-\ell).$$

Thus, in the notation of Lemma 2.1, if G_0 is implicitly defined by

$$G_0(x,y) = x\sin G_0(x,y) + y\cos G_0(x,y),$$

with $G_0(0,0)=0$, we have that G_0 is real-analytic, $G_0(x,y) = y + xy + O_3(x,y)$ and

$$u - \ell = G_0(e\cos\ell, e\sin\ell). \tag{A.18}$$

Therefore, we deduce from (A.18) and (A.13) that

$$u = \lambda + h + G_0(e\cos h\cos\lambda - e\sin h\sin\lambda,\ e\sin h\cos\lambda + e\cos h\sin\lambda). \tag{A.19}$$

Moreover, denoting

$$\hat\eta = \eta/\sqrt{\Lambda}, \qquad \hat\xi = \xi/\sqrt{\Lambda}, \tag{A.20}$$

we deduce from (A.6) and (A.16) that

$$e\sin h = \sqrt{\frac{2H}{\Lambda}}\cdot\sqrt{1-\frac{H}{2\Lambda}}\sin h = \hat\xi\,F_1(\hat\eta^2+\hat\xi^2), \tag{A.21}$$

where $F_1(t) = \sqrt{1-(t/4)}$ is real-analytic for $|t|<4$ (and agrees with the one introduced in Lemma 2.1). Analogously,

$$e\cos h = \hat\eta\,F_1(\hat\eta^2+\hat\xi^2). \tag{A.22}$$

Therefore, substituting (A.21) and (A.22) in (A.19), we can write G_0 as an analytic expression of $(\hat\eta,\hat\xi,\lambda)$: more formally, there exists a real-analytic $(\hat\eta,\hat\xi,\lambda)\mapsto G(\hat\eta,\hat\xi,\lambda)$ (which agrees with the one introduced in (2.6) by (A.21) and (A.22)), so that

$$G_0(e\cos h\cos\lambda - e\sin h\sin\lambda,\ e\sin h\cos\lambda - e\cos h\sin\lambda) = G(\hat\eta,\hat\xi,\lambda).$$

Hence, from (A.19),

$$\begin{aligned} e\cos u &= e\cos h\cos(\lambda+G) - e\sin h\sin(\lambda+G),\\ e\sin u &= e\sin h\cos(\lambda+G) + e\cos h\sin(\lambda+G) \end{aligned} \tag{A.23}$$

with $G = G(\hat\eta,\hat\xi,\lambda)$. Notice also that, from the formulae in (A.16) and (A.14),

$$\frac{1-\sqrt{1-e^2}}{e^2} = F_2(\hat\eta^2+\hat\xi^2)$$

for a suitable real-analytic function F_2 (actually, $F_2(t) = \frac12(1-\frac t4)^{-1}$, which agrees with the notation in Lemma 2.1). Thus, if we set $\varphi = \lambda + v - \ell = v - h$, recalling also

(A.12) and (A.17), we have

$$\begin{aligned}\sin\varphi &= \sin v\cos h - \cos v \sin h\\ &= \frac{1}{1-e\cos u}[\sqrt{1-e^2}\sin u\cos h - \cos u \sin h + e\sin h]\\ &= \frac{1}{1-e\cos u}[\sin(u-h) + e\sin h - F_2\cdot(e\sin u)\cdot(e\cos h)]\\ &= \frac{1}{1-e\cos u}[\sin(\lambda + e\sin u) + e\sin h - F_2\cdot(e\sin u)\cdot(e\cos h)] \qquad \text{(A.24)}\end{aligned}$$

for $F_2 = F_2(\hat\eta^2 + \hat\xi^2)$ and analogously

$$\cos\varphi = \frac{1}{1-e\cos u}[\cos(\lambda + e\sin u) - e\cos h - F_2\cdot(e\sin u)\cdot(e\sin h)]. \qquad \text{(A.25)}$$

Hence, from (A.21), (A.22), (A.23), (A.24), and (A.25), it follows that $\sin\varphi$ and $\cos\varphi$ are real-analytic functions in $\lambda, \hat\eta, \hat\xi$, for $\lambda \in \mathbf{T}$ and small $\hat\xi, \hat\eta$. In particular, if $\mathcal{C}$, $\mathcal{S}$, and $\mathcal{E}_{\mathrm{s}}$ are as defined in Lemma 2.1, we deduce from (A.23), (A.21), and (A.22) that

$$e\sin u = \mathcal{E}_{\mathrm{s}}, \qquad \text{(A.26)}$$

and then from (A.25) and (A.24) that

$$(1-e\cos u)\cos\varphi = \mathcal{C} \qquad \text{and} \qquad (1-e\cos u)\sin\varphi = \mathcal{S}. \qquad \text{(A.27)}$$

Finally, by geometric considerations, we have

$$r = a(1-e\cos u), \qquad \text{(A.28)}$$

where r is the distance between the planet and the sun. Thus, the formulae in Lemma 2.1 follow at once by (A.26), (A.27), (A.3), and (A.8).

A.3. Hamiltonian setting for the planar many-body problem. Consider $(N+1)$ bodies $P_0, \dots, P_N$ of masses $m_0, \dots, m_N$, all lying in the same plane, interacting through gravity (with constant of gravitation 1). Denote by $u^{(i)}$ the position of P_i in a given inertial frame of $\mathbf{R}^2$, with origin in the center of mass of the system. By Newton's laws, we have that

$$\ddot u^{(i)} = \sum_{0\le j\ne i\le N} \frac{m_j(u^{(j)} - u^{(i)})}{|u^{(j)} - u^{(i)}|^3}. \qquad \text{(A.29)}$$

Thus, if $U^{(i)} := m_i \dot u^{(i)}$ denotes the momentum of P_i, we see that the equations of motion (A.29) come from the Hamiltonian

$$\sum_{i=0}^{N} \frac{1}{2m_i}|U^{(i)}|^2 - \sum_{0\le i<j\le N} \frac{m_i m_j}{|u^{(i)} - u^{(j)}|},$$

where $U = (U^{(0)}, \dots, U^{(N)}) \in \mathbf{R}^{2(N+1)}$ and $u = (u^{(0)}, \dots, u^{(N)}) \in \mathbf{R}^{2(N+1)}$ are conjugate symplectic variables.

We now consider P_0 as the "sun" and introduce canonical heliocentric variables via the linear symplectic transformation

$$\begin{aligned}&u^{(0)} = r^{(0)}, \quad u^{(i)} = r^{(0)} + r^{(i)},\\ &U^{(0)} = R^{(0)} - R^{(1)} - \cdots - R^{(N)}, \quad U^{(i)} = R^{(i)}, \qquad \text{for} \quad i = 1, \dots, N. \qquad \text{(A.30)}\end{aligned}$$

Notice that, by our choice of coordinates, $R^{(0)} = 0$. Thus, the planar many-body problem is governed by the $(2N)$-degree-of-freedom Hamiltonian

$$\sum_{i=1}^{N}\left(\frac{m_0+m_i}{2m_0m_i}|R^{(i)}|^2 - \frac{m_0m_i}{|r^{(i)}|}\right) + \sum_{1\le i<j\le N}^{N}\left(\frac{R^{(i)}\cdot R^{(j)}}{m_0} - \frac{m_im_j}{|r^{(i)}-r^{(j)}|}\right).$$

If $m_i = \varepsilon\mu_i$ for $i = 1,\dots,N$, i.e., if the "planets" are very much smaller than the "sun," the momenta $R^{(i)}$ are of order ε. Therefore, it is convenient to introduce the following rescaled symplectic variables:

$$X^{(i)} = \frac{R^{(i)}}{\varepsilon m_0^{5/3}}, \qquad x^{(i)} = \frac{r^{(i)}}{m_0^{2/3}}, \qquad i = 1,\dots,N. \tag{A.31}$$

In such variables, after a time scale of factor $\varepsilon m_0^{7/3}$ we obtain the Hamiltonian in (2.1). Notice that, in that setting, the Hamiltonian $\mathcal{H}_0^{(N)}$ corresponds to the sum of N integrable Hamiltonians of the form (A.2), with $\mathfrak{m}$ and $\mathfrak{M}$ replaced by $\mathrm{m_i}$ and $\mathrm{M_i}$, respectively.

A.4. A parity property. We recall here the well-known fact that the λ-average of $\mathcal{H}_1$ (as in (2.14)) is even in (η,ξ). The proof of this will be accomplished by a 180-degree rotation of the perihelia.

Proposition A.2. *Let*

$$f_1(\Lambda,\eta,\xi) := \frac{1}{\varepsilon}\int_{\mathbf{T}^2}\mathcal{H}_1(\Lambda,\eta,\lambda,\xi)\ d\lambda.$$

Then, $f_1(\Lambda,-\eta,-\xi) = f_1(\Lambda,\eta,\xi)$.

The rescaling by $\frac{1}{\varepsilon}$ is made so that f_1 is a (real-analytic) uniformly bounded (by an order-one constant) function.

Proof. The eccentricity e_i, the semiaxis a_i, and the mean anomaly ℓ_i of the osculating ellipse of P_i are invariant under the map $(\Lambda_i,\lambda_i,H_i,h_i)\mapsto(\Lambda_i,\lambda_i-\pi,H_i,h_i+\pi)$, while the argument of the perihelion g_i changes of π. Let us denote $\vec{\pi} := (\pi,\dots,\pi)\in\mathbf{R}^N$. In light of the consideration above we have that the map

$$(\Lambda,\lambda,H,h)\mapsto(\Lambda,\lambda-\vec{\pi},H,h+\vec{\pi}) \tag{A.32}$$

leaves $|r^{(i)}-r^{(j)}|$ invariant, for any $i,j = 1,\dots,N$. The map in (A.32) corresponds to

$$(\Lambda,\lambda,\xi,\eta)\mapsto(\Lambda,\lambda-\pi,-\xi,-\eta),$$

usually referred to as "space inversion." ∎

Appendix B. Simple eigenvalue perturbations.

Lemma B.1. *Let* $M^\star\in\mathrm{Mat}(m\times m)$, $M_\star\in\mathrm{Mat}(k\times k)$, *and* $M^\sharp\in\mathrm{Mat}(m\times k)$. *Then,*

$$\det\begin{pmatrix} M^\star & M^\sharp \\ \mathbf{0}_{k\times m} & M_\star\end{pmatrix} = \det(M^\star)\det(M_\star).$$

Proof. The proof is obvious. ∎

LEMMA B.2. *Let* $M^\star \in \mathrm{Mat}(m\times m)$, $M_\star \in \mathrm{Mat}(k\times k)$, $M^\sharp \in \mathrm{Mat}(m\times k)$, *and* $M_\sharp \in \mathrm{Mat}(k\times m)$. *Let*

$$M_\epsilon := \begin{pmatrix} M^\star + O(\epsilon) & \epsilon M^\sharp + O(\epsilon^2) \\ \\ \epsilon M_\sharp + O(\epsilon^2) & \epsilon M_\star + O(\epsilon^2) \end{pmatrix}.$$

Then

- *if* $\bar\lambda \neq 0$ *is a simple eigenvalue of* $M^\star$, *then there exists* $\bar\lambda_\epsilon = \bar\lambda + O(\epsilon)$ *which is an eigenvalue of* M_ϵ, *provided* $|\epsilon|$ *is suitably small;*
- *if* $\tilde\lambda$ *is a simple eigenvalue of* $M_\star$ *and* $\det(M^\star)\neq 0$, *then there exists* $\tilde\lambda_\epsilon = \tilde\lambda + O(\epsilon)$ *so that* $\epsilon\tilde\lambda_\epsilon$ *is an eigenvalue of* M_ϵ, *provided* $|\epsilon|$ *is suitably small.*

Proof. For the first claim, apply the implicit function theorem to

$$\mathcal{F}_1(t,\epsilon) := \det(M_\epsilon - t\mathbf{1}_{(m+k)}),$$

noticing that $\mathcal{F}_1(t,0) = (-1)^k t^k \det(M^\star - t\mathbf{1}_m)$. For the second claim, apply the implicit function theorem to

$$\mathcal{F}_2(t,\epsilon) := \det \begin{pmatrix} M^\star - \epsilon t\mathbf{1}_m & M^\sharp \\ \epsilon M_\sharp & M_\star - t\mathbf{1}_k \end{pmatrix},$$

noticing that $\epsilon^k \mathcal{F}_2(t,\epsilon) = \det(M_\epsilon - \epsilon t\mathbf{1}_{(m+k)})$ and that, by Lemma B.1, $\mathcal{F}_2(t,0) = \det(M^\star)\det(M_\star - t\mathbf{1}_k)$. □

Acknowledgment. We thank Carlangelo Liverani for fruitful suggestions.

REFERENCES

[A63] V. I. Arnold, *Small denominators and problems of stability of motion in classical and celestial mechanics*, Uspekhi Mat. Nauk., 18 (1963), pp. 91–192 (in Russian).

[BBV04] M. Berti, L. Biasco, and E. Valdinoci, *Periodic orbits in Hamiltonian systems with elliptic invariant tori and applications to the planetary three-body problem*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (5), 3 (2004), pp. 87–138.

[BCV03] L. Biasco, L. Chierchia, and E. Valdinoci, *Elliptic two–dimensional invariant tori for the planetary three-body problem*, Arch. Ration. Mech. Anal., 170 (2003), pp. 91–135.

[CC03] A. Celletti and L. Chierchia, *KAM Stability and Celestial Mechanics*, Mem. Am. Math. Soc., to appear, http://www.mat.uniroma3.it/users/chierchia/PREPRINTS/SJV_03.pdf.

[Ch88] A. Chenciner, *Intégration du problème de Kepler par la méthode de Hamilton–Jacobi: Coordonnées "action-angles" de Delaunay and coordonnées de Poincaré*, Notes scientifiques et techniques du B.D.L., S026 (1988).

[E88] L. Eliasson, *Perturbations of stable invariant tori for Hamiltonian systems*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 15 (1988), pp. 115–147.

[F02] J. Féjoz, *Quasiperiodic motions in the planar three-body problem*, J. Differential Equations, 183 (2002), pp. 303–341.

[F04] J. Féjoz, *Démonstration du "théorème d'Arnold" sur la stabilité du système planétaire (d'après Michael Herman)*, Ergodic Theory Dyn. Sys., 24 (2004), pp. 1521–1582.

[H95] M. Herman, *Proof of Arnold's Theorem on the Existence of Maximal Tori for Planetary Systems*, private communication, 1995.

[JM66] W. H. Jefferys and J. Moser, *Quasi–periodic solutions for the three-body problem*, Astronom. J., 71 (1966), pp. 568–578.

[K88] S. B. Kuksin, *Perturbation theory of conditionally periodic solutions of infinite–dimensional Hamiltonian systems and its applications to the Korteweg–de Vries equation*, Mat. Sb., 136 (1988), pp. 396–412, 431 (in Russian); Math. USSR-Sb., 64 (1989), pp. 397–413.

[L91] J. Laskar, *Analytical framework in Poincaré variables for the motions of the solar system*, in Predictability, Stability, and Chaos in N-Body Dynamical Systems (Cortina d'Ampezzo, 1990), NATO Adv. Sci. Inst. Ser. B Phys. 272, Plenum, New York, 1991, pp. 93–114.

[L96] J. Laskar, *Large scale chaos and marginal stability in the solar system*, Celestial Mech. Dynam. Astronom., 64 (1996), pp. 115–162.

[LR95] J. Laskar and P. Robutel, *Stability of the planetary three-body problem.* I. *Expansion of the planetary Hamiltonian*, Celestial Mech. Dynam. Astronom., 62 (1995), pp. 193–217.

[M65] V. K. Melnikov, *On certain cases of conservation of almost periodic motions with a small change of the Hamiltonian function*, Dokl. Akad. Nauk SSSR, 165 (1965), pp. 1245–1248 (in Russian).

[N77] N. N. Nekhoroshev, *An exponential estimate of the time of stability of nearly–integrable Hamiltonian systems,* I, Uspekhi. Mat. Nauk., 32 (1977), pp. 5–66 (in Russian); Russian Math. Survey, 32 (1977), pp. 1–65.

[Poi1905] H. Poincaré, *Leçons de Mécanique Céleste*, Tome 1, Gauthier–Villars, Paris, 1905; also available on-line from http://gallica.bnf.fr/scripts/ConsultationTout.exe?E=0&O=N095010.

[P89] J. Pöschel, *On elliptic lower dimensional tori in Hamiltonian system*, Math. Z., 202 (1989), pp. 559–608.

[P96] J. Pöschel, *A KAM–theorem for some nonlinear PDEs*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 23 (1996), pp. 119–148.

[R95] P. Robutel, *Stability of the planetary three-body problem.* II. *KAM theory and existence of quasi–periodic motions*, Celestial Mech. Dynam. Astronom., 62 (1995), pp. 219–261.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1589–1615

QUASIMINIMAL PARTITIONS WITH PRESCRIBED MEASURE*

SÉVERINE RIGOT†

Abstract. We prove that, in a fairly general context, quasiminimal partitions with prescribed measure enjoy quantitative rectifiability properties with universal bounds. Namely, we show that the set of interfaces of a quasiminimal partition is uniformly rectifiable with bounds that depend only on the structural data of the problem.

Key words. partitions, Ahlfors-regular sets, uniform rectifiability

AMS subject classifications. 49Q20, 49N60, 28A75

DOI. 10.1137/S003614100443960X

1. Introduction.

This paper is devoted to the study of quantitative and uniform rectifiability properties for quasiminimal partitioning hypersurfaces between regions in $\mathbb{R}^n$ of prescribed n dimensional measure. Usually one says that a partition is minimal if admissible deformations of it cannot decrease its energy. For quasiminimal partitions one allows the energy to decrease but only in a controlled way. Energies we shall consider here will be surface-like energies of interface type.

There is a wide class of variational problems involving a competition between surface- and volume-like energies where quasiminimality conditions appear naturally. There are also many situations where such a variational problem turns out to be related to a partitioning problem. For instance the various regions delimitated by a quasiminimal partitioning hypersurface seem to provide a model for equilibrium configurations of a system of immiscible and incompressible fluids subject to external forces of volume type (for example the gravity). Roughly speaking the term in the energy corresponding to these external forces is, mostly for homogeneity reasons, negligible compared to the surface-like energy, which reflects the effects of the surface tension between the different fluids. Then, when a configuration has minimal total energy among all other possible configurations, the surface-like energy turns out to enjoy on its own some suitable quasiminimality condition.

A major work in the study of minimal partitions is due to Almgren [1]. He proved existence and regularity of partitioning hypersurfaces between regions of given measure, hypersurfaces which minimize an appropriate surface-like energy with so-called partitioning regular weights assigned to the various interfaces. This work has been the main source for the present paper, the goal being to give (nontrivial) improvements of some of the results in [1] about the regularity theory of minimal partitions (we will not consider here questions of existence).

First we consider more general energies since it seems that relevant physical situations do not fit Almgren's setting (compare in particular (H1), (H2), and (H3) with the hypotheses in [1, section VI.1], see also [8], [11]). We shall however go back to Almgren's setting in section 6. Next we relax the minimality into a quasiminimality condition which reflects more closely the presence of external forces of volume type.

*Received by the editors January 14, 2004; accepted for publication (in revised form) April 15, 2005; published electronically January 27, 2006.
http://www.siam.org/journals/sima/37-5/43960.html

†Laboratoire de Mathématiques, Université Paris-Sud, Bâtiment 425, 91405 Orsay Cedex, France (Severine.Rigot@math.u-psud.fr).

This will actually not be a serious issue here and can be handled without too much pain.

More significantly we prove some uniform and universal versions of properties that appeared more or less explicitly, and in weaker form, in [1]. The properties we have in mind, namely Ahlfors-regularity and condition B (see section 3 for precise definitions), or the weaker forms exploited in [1], constitute some of the main steps in the proof of the regularity results of $C^{1,\alpha}$ type obtained in [1] (this actually turns out to be more generally the case in many other geometric variational problems). Another major motivation for the study of Ahlfors-regularity and condition B comes from the fact that they imply the uniform rectifiability of the set of interfaces. We use here the terminology of David and Semmes who have developed an extensive theory of uniformly rectifiable sets. Uniform rectifiability is a quantitative and uniform version of the classical notion of rectifiability, but it is much stronger because it comes with uniform bounds (see [5] and the references therein). Note that due to the generality adopted here—we assume only some Lipschitz regularity and nondegeneracy of the defining integrands of the energy—one cannot hope to have much more in the way of smoothness than uniform rectifiability, just for reasons of bi-Lipschitz invariance.

Finally we stress that, besides these quantitative rectifiability properties, the main goal of this paper is to prove them with universal bounds. This roughly means that we want to prove that these properties hold for any quasiminimal partition with prescribed measure with quantifiers which depend only on the structural data of the problem (the dimension, the prescribed measure, the way the quasiminimality is formulated, bounds on the defining integrands of the energy) but not on the particular partition under study (see section 3 for precise statements). In particular one of the main steps in our proof is to show that quasiminimal partitions with prescribed measure actually satisfy an apparently stronger condition where there is no volume constraint anymore on the various components of admissible deformations, this new condition being formulated in a way that depends only on the structural data of the problem (see Theorem 3.6). This is one of the main points which improves what was proved in [1] where the same kind of condition without volume constraint was obtained but in a way that depends strongly on the specific geometry of the partition under study (see especially the comment around typical examples of (F, ε, δ) minimal sets on page vii in [1]).

The main issue when studying quasiminimal partitions with prescribed measure is indeed to handle properly the volume constraint and Theorem 3.6 follows actually from the construction of suitable deformations which modify the measure of each component of the partition by a prescribed amount given in advance. As just pointed out an additional difficulty here comes from the fact that we also need to keep a universal control in these constructions just because of the uniform constants we are seeking in the formulation of Ahlfors-regularity and condition B. This will be done through a uniform version of Almgren's original deformations, see Theorem 4.1 (compare with Theorem 4.2 which comes from [1]).

The same kind of results have been previously proved in [10] in a similar context and for the simpler case of quasiminimal crystals, that is, for partitions composed by simply one set and its complement, but with a completely different method.

The paper is organized as follows. In section 2 we recall basic facts about the theory of sets with finite perimeter and define various classes of partitions and of energies. We state the main results of this paper in section 3. Theorem 3.6 which says that any quasiminimal partition with a volume constraint is locally quasiminimal in a suitable universal way is proved in section 4. In particular the key argument of the proof, the above-mentioned construction of suitable deformations, is given there by Theorem 4.1. We prove the Ahlfors-regularity and condition B for locally quasiminimal partitions in section 5. In section 6 we restrict the study to energies with so-called partitioning regular interface coefficients as in [1] and prove refined properties.

2. Preliminaries. Throughout this paper we work in $\mathbb{R}^n$, $n \geq 2$, equipped with its euclidean structure. For $x \in \mathbb{R}^n$ and $t > 0$, we denote by $B(x,t)$ the open ball with center x and radius t. For any set $G \subset \mathbb{R}^n$ we denote by $\mathbb{1}_G$ its characteristic function. If G is measurable, then $|G|$ stands for its Lebesgue measure. Finally we denote by $\mathcal{H}^{n-1}$ the $(n-1)$ dimensional Hausdorff measure.

We first recall some results about the theory of sets with finite perimeter that will be relevant for our purposes. For more details we refer the reader, for instance, to [2], [7], or [12]. If $G \subset \mathbb{R}^n$ is a measurable set and $\Omega \subset \mathbb{R}^n$ is open, the perimeter of G in Ω, denoted by $P(G,\Omega)$, is defined by

$$P(G,\Omega) = \sup\left\{\int_\Omega \mathbb{1}_G \,\operatorname{div}\phi\, dx;\ \phi \in C_0^1(\Omega,\mathbb{R}^n),\ \|\phi\|_\infty \leq 1\right\}.$$

We say that a measurable set G has finite perimeter if $P(G,\mathbb{R}^n) < +\infty$ and we then simply write $P(G) = P(G,\mathbb{R}^n)$.

If G has finite perimeter, then one knows that the set function $\Omega \mapsto P(G,\Omega)$ defined above is the restriction to the open subsets of $\mathbb{R}^n$ of a finite Borel measure, called the perimeter of G and still denoted by $P(G,.)$. Equivalently it follows from Riesz' representation theorem that a measurable set G has finite perimeter if and only if the distributional gradient of its characteristic function $\nabla\mathbb{1}_{\mathrm{G}}$ can be represented by a vector-valued measure. Then it turns out that the total variation $|\nabla\mathbb{1}_{\mathrm{G}}|$ of this measure coincides with $P(G,.)$.

The reduced boundary $\partial_* G$ of a set G with finite perimeter is defined as the set of points $x \in \mathbb{R}^n$ such that

$$\int_{B(x,r)} |\nabla\mathbb{1}_{\mathrm{G}}| > 0 \quad \text{for all } r > 0,$$

the limit

$$\nu_G(x) = \lim_{r\to 0}\left(\int_{B(x,r)} \nabla\mathbb{1}_{\mathrm{G}}\right) / \int_{B(x,r)} |\nabla\mathbb{1}_{\mathrm{G}}|$$

exists, and $\|\nu_G(x)\| = 1$. It follows from the theorem of Besicovitch on differentiation of measures that $\nu_G(x)$ exists and $\|\nu_G(x)\| = 1$ for $|\nabla\mathbb{1}_{\mathrm{G}}|$-a.e. $x \in \mathbb{R}^n$. Moreover, it is well known that the perimeter $P(G,.)$ coincides with $\mathcal{H}^{n-1}_{|\partial_* G}(.)$.

For any measurable set G and $t > 0$, we set

$$G(t) = \left\{x \in \mathbb{R}^n;\ \lim_{r\to 0}\frac{|G \cap B(x,r)|}{|B(x,r)|} = t\right\},$$

and define the essential boundary $\partial^* G$ of G as the set of points where the volume density of G is neither 0 nor 1, $\partial^* G = \mathbb{R}^n \setminus (G(0) \cup G(1))$. It is also well known that if G has finite perimeter, then

$$\partial_* G \subset G(1/2) \subset \partial^* G \quad \text{and} \quad \mathcal{H}^{n-1}(\mathbb{R}^n \setminus (G(0) \cup G(1) \cup \partial_* G)) = 0. \tag{2.1}$$

Note that it follows in particular that $P(G,.)$ coincides with $\mathcal{H}^{n-1}_{|\partial^* G}$. Note also that we always have $\partial^* G \subset \partial G$ where ∂G is the topological boundary of G.

We turn now to the definition of a partition we shall work with. We fix once and for all an integer $N \in \mathbb{N}^*$.

DEFINITION 2.1 (partition). *We say that a family $A = (A_0, \dots, A_N)$ of measurable sets is a partition if*

$$\begin{gathered} A_i \text{ is a set with finite perimeter for all } i \in \{0, \dots, N\}, \\ |A_i \cap A_j| = 0 \text{ when } i \neq j, \\ |\mathbb{R}^n \setminus \cup_{i=0}^N A_i| = 0. \end{gathered}$$

We denote by $\mathcal{P}$ the set of all partitions.

DEFINITION 2.2 (partition with prescribed measure). *Let $a = (a_1, \dots, a_N)$, $a_i > 0$, be fixed. We say that a partition $A \in \mathcal{P}$ has measure a if $|A_i| = a_i$ for all $i \in \{1, \dots, N\}$. We denote by $\mathcal{P}_a$ the set of all partitions with measure a.*

We will be mostly interested here in the regularity properties of the set of interfaces of some partitions. This set is precisely defined as follows.

DEFINITION 2.3 (set of interfaces). *For any $A \in \mathcal{P}$, we define the set S of its interfaces by $S = \cup_{i=0}^N \partial^* A_i$.*

Note that $\partial^* A_i = \cup_{j \neq i} \partial^* A_i \cap \partial^* A_j$ for all $i \in \{0, \dots, N\}$. Note also that the interfaces $S_{ij} = \partial^* A_i \cap \partial^* A_j$, $i \neq j$, are essentially pairwise disjoint in the sense that

$$\mathcal{H}^{n-1}(S_{ij} \cap S_{lk}) = 0$$

whenever $i \neq j$, $l \neq k$, and at least three of the indexes are distinct.

The symmetric difference $A \triangle A'$ between two partitions A and $A' \in \mathcal{P}$ is defined by

$$A \triangle A' = \cup_{i=0}^N A_i \triangle A'_i,$$

where $A_i \triangle A'_i = (A_i \setminus A'_i) \cup (A'_i \setminus A_i)$ denotes the usual symmetric difference between sets. We set

$$|A \triangle A'| = \max_{i \geq 0} |A_i \triangle A'_i|,$$

and $\operatorname{diam}(A \triangle A') = \max_{i \geq 0} \operatorname{diam}(A_i \triangle A'_i)$. We say that A and A' are equivalent if $|A \triangle A'| = 0$. Note that two equivalent partitions have the same set of interfaces.

DEFINITION 2.4 (reduced partition). *We say that $A \in \mathcal{P}$ is a reduced partition if each component A_i of A coincides with the set of its Lebesgue points, that is, $A_i = A_i(1)$ for all $i \in \{0, \dots, N\}$.*

For any given partition A there exists a unique reduced equivalent partition. It is obtained from A replacing each one of its components by the set of their Lebesgue points. Note that if $A \in \mathcal{P}$ is reduced, then $A_i \cap A_j = \emptyset$ for $i \neq j$, $A_i \cap S = \emptyset$ for all $i \in \{0, \dots, N\}$, and $\mathbb{R}^n = (\cup_{i=0}^N A_i) \cup S$.

We now define the energy of a partition. We denote by S^{n-1} the unit sphere in $\mathbb{R}^n$. Given a family of Borel functions $F_{ij} : \mathbb{R}^n \times S^{n-1} \to \mathbb{R}^+$, $i, j \in \{0, \dots, N\}$, $i \neq j$, we define the associated energy E as

$$E(A, U) = \sum_{\substack{i,j \\ i \neq j}} \int_{\partial^* A_i \cap \partial^* A_j} \mathbb{1}_U(x) F_{ij}(x, \nu_{A_i}(x)) \, d\mathcal{H}^{n-1}$$

for any $A \in \mathcal{P}$ and any measurable set $U \subset \mathbb{R}^n$. When $U = \mathbb{R}^n$ we simply write $E(A)$. Note that this definition is consistent because ν_{A_i} is well defined $\mathcal{H}^{n-1}$-a.e. on $\partial^* A_i$ with $\|\nu_{A_i}(x)\| = 1$ for $\mathcal{H}^{n-1}$-a.e. $x \in \partial^* A_i$. Note also that two equivalent partitions have the same energy.

Let $\underline{F} > 0$ and $\overline{F} > 0$ be two fixed positive constants with $\underline{F} \leq \overline{F}$. Given a map $F : \mathbb{R}^n \times S^{n-1} \to \mathbb{R}^+$ we consider the following conditions:

$$\text{(H1)} \qquad \sup_{\mathbb{R}^n \times S^{n-1}} F \leq \overline{F},$$

$$\text{(H2)} \qquad \underline{F} \leq \inf_{\mathbb{R}^n \times S^{n-1}} F,$$

$$\text{(H3)} \qquad F \text{ is } \overline{F}\text{-Lipschitz on } \mathbb{R}^n \times S^{n-1}, \text{ that is,}$$

$$|F(x,u) - F(y,v)| \leq \overline{F} \max(\|x-y\|, \|u-v\|) \quad \text{for all } x, y \in \mathbb{R}^n, \text{ for all } u, v \in S^{n-1}.$$

Then we denote by $\mathcal{E}_0(\underline{F}, \overline{F})$ the class of all those energies that are defined with respect to some family of integrands (F_{ij}) where each F_{ij} satisfies (H1) and (H2), and by $\mathcal{E}(\underline{F}, \overline{F})$ the class of all energies for which each associated defining integrand F_{ij} satisfies (H1), (H2), and (H3). Finally we denote by $\mathcal{E}(\overline{F})$ the class of all energies for which each defining integrand F_{ij} satisfies (H1) and (H3).

When $F_{ij}(x, \nu) = \|\nu\|$ the corresponding energy is simply the sum of the perimeters of the different components. With a slight abuse of terminology and notation we will call it the perimeter of the partition and denote it by P, that is,

$$P(A, U) = \sum_{i=0}^{N} \sum_{\substack{j=0 \\ j \neq i}}^{N} \mathcal{H}^{n-1}(\partial^* A_i \cap \partial^* A_j \cap U) = \sum_{i=0}^{N} P(A_i, U)$$

for all $A \in \mathcal{P}$ and measurable set $U \subset \mathbb{R}^n$. Note that $\mathcal{H}^{n-1}$-a.e. point in S belongs to exactly one interface $\partial^* A_i \cap \partial^* A_j$ for some $i \neq j$ and thus

$$P(A, U) = 2\, \mathcal{H}^{n-1}(S \cap U). \tag{2.2}$$

Finally note that, whenever $E \in \mathcal{E}_0(\underline{F}, \overline{F})$, we have

$$\underline{F}\, P(A, U) \leq E(A, U) \leq \overline{F}\, P(A, U). \tag{2.3}$$

To conclude these preliminaries we fix some more conventions. In general nonindexed letters, typically the letter C, will denote positive constants whose precise value can change at each occurrence (unless otherwise stated). On the other hand, indexed letters (C_0, C_n, $\eta_0, \dots$) should denote constants whose value remains the same throughout a same paragraph (but whose signification and value may change from one paragraph to another, unless otherwise stated). Finally when specifying $C = C(n, N, \dots)$ we mean that the value of C can be chosen depending only on the fixed data inside the parentheses.

3. Quasiminimality conditions and main results. We fix an energy E as defined in section 2. We also fix $\underline{F} > 0$ and $\overline{F} > 0$ with $\underline{F} \leq \overline{F}$ in view of further assumptions we shall make on the energy. We first define quasiminimal partitions with prescribed measure.

DEFINITION 3.1. *Let $g : [0, +\infty) \to [0, +\infty)$ be such that $\lim_{t \to 0} t^{(1-n)/n} g(t) = 0$ and some $a = (a_1, \dots, a_N)$, $a_i > 0$, be fixed. We say that a partition $A \in \mathcal{P}$ is (g, a)-quasiminimal (with respect to E) if $A \in \mathcal{P}_a$ and*

$$E(A) \leq E(A') + g(|A' \triangle A|) \tag{3.1}$$

for all $A' \in \mathcal{P}_a$ such that $A' \triangle A \subset\subset \mathbb{R}^n$.

Note that in this definition admissible perturbations must preserve the measure of the partition. Note also that (3.1) gives significant information only when $|A'\triangle A|$ is small. In that case the way the energy can be decreased is controlled by $g(|A'\triangle A|)$ and hence negligible compared to $|A'\triangle A|^{(n-1)/n}$. Roughly speaking this turns out to be essentially equivalent to requiring the variation of the energy to be negligible compared to the initial energy.

Before stating the main results of this paper we need some more definitions.

Definition 3.2 (Ahlfors-regularity). *We say that a closed subset Σ of $\mathbb{R}^n$ is Ahlfors-regular if there exist some constants $c > 0$ and $C > 0$, called Ahlfors-regularity constants, such that*

$$c\, r^{n-1} \leq \mathcal{H}^{n-1}(\Sigma \cap B(x,r)) \leq C\, r^{n-1}$$

for all $x \in \Sigma$ and $r \leq 1$.

This is a uniform and scale-invariant version of the property of having upper and lower densities with respect to $\mathcal{H}^{n-1}$ that are positive and finite.

Definition 3.3 (condition B). *We say that a subset Σ of $\mathbb{R}^n$ satisfies condition B if there exists a constant $C > 0$, called condition B constant, such that, for each ball B centered on Σ with radius $r \leq 1$, one can find two balls contained in B, with radius $C\, r$, and which are contained in two distinct connected components of $\mathbb{R}^n \setminus \Sigma$.*

This topological condition is a quantitative, uniform, and scale-invariant way of saying that Σ separates well the different connected components of its complement. Sets that are Ahlfors-regular and satisfy condition B have strong rectifiability properties. Namely, they contain "big pieces of Lipschitz graphs" and thus are uniformly rectifiable (see [3] for the original proof or [4] and [6] for simpler proofs). The aim of this paper not being to speak about the theory of uniform rectifiability, we will not enter the details and refer the reader to [5] and the references therein for more information.

We can now state the main result of this paper about quasiminimal partitions with prescribed measure.

Theorem 3.4. *Assume that $E \in \mathcal{E}(\underline{F}, \overline{F})$. Let $g : [0,+\infty) \to [0,+\infty)$ be non-decreasing and such that $\lim_{t\to 0} t^{(1-n)/n} g(t) = 0$ and let $a = (a_1, \ldots, a_N)$, $a_i > 0$, be fixed. Let A be a (g,a)-quasiminimal partition with respect to E and let S denote its set of interfaces. Then S is a closed Ahlfors-regular set which satisfies condition B. Moreover, the Ahlfors-regularity and condition B constants can be taken depending only on n, N, $\underline{F}$, $\overline{F}$, g, and a.*

As already stressed in the introduction, besides the Ahlfors-regularity and condition B on their own, the main point in this result is that one can find universal Ahlfors-regularity and condition B constants, which do not depend on the specific quasiminimal partition A.

We now wish to explain the main lines of the proof. Let us first introduce another notion of quasiminimality where admissible perturbations are not required to be volume preserving anymore, but a little bit localized instead (this constraint will actually not be a serious issue here).

Definition 3.5 (locally quasiminimal partitions). *Let $g : [0,+\infty) \to [0,+\infty)$ be such that $\lim_{t\to 0} t^{(1-n)/n} g(t) = 0$ and let $\delta > 0$ be fixed. We say that a partition $A \in \mathcal{P}$ is (g,δ)-quasiminimal (with respect to E) if A satisfies (3.1) for all $A' \in \mathcal{P}$ such that $\operatorname{diam}(A'\triangle A) < \delta$.*

In a first step we want to get rid of the volume constraint, proving that quasiminimal partitions with prescribed measure are locally quasiminimal in an appropriate and universal way.

THEOREM 3.6. *Assume that $E \in \mathcal{E}(\underline{F}, \overline{F})$. Let $g : [0, +\infty) \to [0, +\infty)$ be nondecreasing and such that $\lim_{t\to 0} t^{(1-n)/n} g(t) = 0$ and let $a = (a_1, \ldots, a_N)$, $a_i > 0$, be fixed. Then there exist $\tilde{g} : [0, +\infty) \to [0, +\infty)$ such that $\lim_{t\to 0} t^{(1-n)/n} \tilde{g}(t) = 0$ and $\delta > 0$ such that any (g, a)-quasiminimal partition with respect to E is $(\tilde{g}, \delta)$-quasiminimal with respect to E. In particular $\tilde{g}$ and δ can be chosen depending only on n, N, $\underline{F}$, $\overline{F}$, g, and a.*

We shall prove this theorem in section 4. Let us mention that the assumption on g to be nondecreasing is here mostly for technical convenience. As already stressed, Theorem 3.6 is, at least from a technical point of view, the most delicate point here, essentially because, in view of the last part of Theorem 3.4, we need to find $\tilde{g}$ and δ universal. The proof of Theorem 3.6 relies on the construction of suitable deformations of a partition which modify the measure of each component of a prescribed amount given in advance. This will be done in Theorem 4.1 which is actually the crucial step in the proof.

Next, with Theorem 3.6 in hand, we only need to worry about locally quasiminimal partitions and we shall prove in section 5 the Ahlfors-regularity and condition B in that context (see Theorem 5.1). Arguments are based on comparison arguments that one can now easily implement because there is no volume constraint anymore. Let us also mention that we will actually get more information. First note that if A and A' are two equivalent partitions and if A is quasiminimal, then A' is quasiminimal as well and both partitions have the same set of interfaces. In particular one can always assume with no loss of generality that the quasiminimal partition is reduced (remember the remarks around Definition 2.4). Then it is possible to prove that each component is open and that one can replace essential boundaries with topological boundaries in the definition of the set of interfaces.

THEOREM 3.7. *Assume that $E \in \mathcal{E}_0(\underline{F}, \overline{F})$. Let $g : [0, +\infty) \to [0, +\infty)$ be such that $\lim_{t\to 0} t^{(1-n)/n} g(t) = 0$ and let $\delta > 0$ be fixed. Let A be a reduced (g, δ)-quasiminimal partition with respect to E and let S denote its set of interfaces. Then each component A_i, $i \in \{0, \ldots, N\}$, of A is open and $S = \cup_{i=0}^{N} \partial A_i$. Moreover each component with finite Lebesgue measure is bounded.*

Note that we do not say in Theorem 3.7 that each boundary ∂A_i satisfies the regularity properties, the Ahlfors-regularity and condition B, on its own and neither that $\partial^* A_i$ and ∂A_i coincide. This is actually not true in general. However it turns out, as we shall show in section 6, that this is the case when one imposes more conditions on the defining integrands F_{ij} of the energy. These conditions correspond to those taken in [1]. We refer the reader to section 6 for more details and precise statements; see, in particular, Theorem 6.1.

4. Quasiminimal partitions with a volume constraint are locally quasiminimal. This section is devoted to the proof of Theorem 3.6. We first construct in section 4.1 suitable families of deformations of a partition that allow to modify the measure of each component by some prescribed amount given in advance. The proof of Theorem 3.6 itself will be done in section 4.2.

4.1. Measure-prescribed deformations. We first define what we mean by suitable families of deformations. Let $\overline{F} > 0$ be fixed. Let $A \in \mathcal{P}$ be a partition, $\Omega \subset \mathbb{R}^n$ be open, $\varepsilon > 0$ and $C > 0$ be fixed. We say that a map $\psi : (-\varepsilon, \varepsilon)^N \times \mathbb{R}^n \to \mathbb{R}^n$

belongs to $\mathcal{D}_\Omega(A,\varepsilon,C)$ if for all $v=(v_1,\dots,v_N)\in(-\varepsilon,\varepsilon)^N$ we have

$$\psi(v,.) \text{ is a } C^1 \text{ diffeomorphism}, \quad \psi(0,.)=\mathrm{Id}, \tag{4.1}$$

$$\mathcal{O}(\psi(v,.)) \subset\subset \Omega, \tag{4.2}$$

where $\mathcal{O}(f)$ is defined for $f:\mathbb{R}^n\to\mathbb{R}^n$ by $\mathcal{O}(f)=\{x\in\mathbb{R}^n;\ f(x)\neq x\}$,

$$|\psi(v,A_i)|-|A_i|=v_i, \tag{4.3}$$

$$|\psi(v,A_i)\triangle A_i| \le C\,P(A_i,\Omega)\,|v|, \tag{4.4}$$

for all $i\in\{1,\dots,N\}$, with $|v|=\max|v_i|$, and finally

$$E(\psi(v,A),\psi(v,U)) \le E(A,U)+C\,P(A,U)\,|v| \tag{4.5}$$

for all energy $E\in\mathcal{E}(\overline{F})$ and all measurable set $U\subset\mathbb{R}^n$.

In other words one can view an element ψ in $\mathcal{D}_\Omega(A,\varepsilon,C)$ as a family $\psi(v,.)$ of deformations of the partition A. The main point is that for any given $v\in(-\varepsilon,\varepsilon)^N$, one can use $\psi(v,.)$ to add, or remove, depending on the sign of v_i, exactly the prescribed amount of mass v_i to the component A_i (see (4.3)). At the same time (4.4) and (4.5) give a control of order $|v|$ on the measure of the symmetric difference between $\psi(v,A)$ and A and on the induced variation of the energy. This will fit exactly our needs for the estimation of each term in the quasiminimality condition.

For technical reasons we will actually need to use two such families of deformations, each one acting far away from the other. Let $\delta>0$ be fixed. We say that (ψ_1,ψ_2) belongs to $\mathcal{CD}_\Omega(A,\varepsilon,\delta,C)$ if $\psi_j\in\mathcal{D}_\Omega(A,\varepsilon,C)$ for $j=1,2$ and

$$\mathrm{dist}(\mathcal{O}(\psi_1(v,.)),\mathcal{O}(\psi_2(v,.)))\ge\delta$$

for all $v\in(-\varepsilon,\varepsilon)^N$. When $\Omega=\mathbb{R}^n$ we just write $\mathcal{D}(A,\varepsilon,C)$ and $\mathcal{CD}(A,\varepsilon,\delta,C)$.

We are now ready to state the main result about the existence of such families of deformations. Let $\underline{a}>0$, $\overline{a}>0$, $\overline{P}>0$ be fixed with $\underline{a}\le\overline{a}$. We denote by $\mathcal{P}_{\underline{a},\overline{a},\overline{P}}$ the set of all partitions $A=(A_0,\dots,A_N)\in\mathcal{P}$ such that

$$\underline{a}\le\min_{i\ge1}|A_i|\le\max_{i\ge1}|A_i|\le\overline{a} \qquad \text{and} \qquad P(A)\le\overline{P}.$$

Theorem 4.1. *There exist three constants $\varepsilon_0>0$, $\delta_0>0$, and $C_0>0$, depending only on n, N, $\overline{F}$, $\underline{a}$, $\overline{a}$, and $\overline{P}$, such that, for any $A\in\mathcal{P}_{\underline{a},\overline{a},\overline{P}}$, we have $\mathcal{CD}(A,\varepsilon_0,\delta_0,C_0)\neq\emptyset$.*

It has been proved in [1] that for a fixed $A\in\mathcal{P}$ we have $\mathcal{CD}(A,\varepsilon,\delta,C)\neq\emptyset$ for suitable parameters $\varepsilon=\varepsilon(A)$, $\delta=\delta(A)$, and $C=C(A)$ which depend strongly on the specific geometry of A. As already explained in the introduction this does not fit our needs. However it also follows from [1] that $\mathcal{CD}(A',\varepsilon,\delta,C)\neq\emptyset$ with the same parameters ε, δ, and C as for A, as soon as A' is close enough to A in measure. This will be used in the proof of Theorem 4.1. More precisely, one has the following result.

Theorem 4.2 (see [1]). *Let $A\in\mathcal{P}$ and $\Omega\subset\mathbb{R}^n$ be open. Assume that $|A_i\cap\Omega|>0$ for all $i\in\{0,\dots,N\}$. Then there exist $\varepsilon>0$, $\delta>0$, $C>0$, and $\eta>0$ such that, if $A'\in\mathcal{P}$ is such that $|A\triangle A'|\le\eta$, then $\mathcal{CD}_\Omega(A',\varepsilon,\delta,C)\neq\emptyset$. Moreover if $\Omega=\cup_{l=1}^s\Omega_l$ where Ω_l is open and $\Omega_l\cap\Omega_r=\emptyset$ for $l\neq r$, one can construct $(\psi_1,\psi_2)\in\mathcal{CD}_\Omega(A',\varepsilon,\delta,C)$ in such a way that*

$$\{x\in\Omega_l\,;\ \psi_j(v,x)\neq x\}\cup\psi_j(v,\{x\in\Omega_l\,;\ \psi_j(v,x)\neq x\})\subset\subset\Omega_l$$

for all $l\in\{1,\dots,s\}$, $j\in\{1,2\}$, and $v\in(-\varepsilon,\varepsilon)^N$.

This theorem is essentially given by Proposition VI.12 in [1]. We add here the requirement (4.4) about the measure of the symmetric difference between the initial partition and the deformed one. This condition never appears in [1] because Almgren deals with minimal partitions and does not need to handle the extra term given by the function g in the quasiminimality condition. Next we want to control the variation of the energy for all energies in the class $\mathcal{E}(\overline{F})$ (see (4.5)) which is a priori quite larger than the class of energies considered in [1]. Finally we also add, mainly for technical reasons, the localization in Ω and each Ω_l. It turns actually out that one can easily check that the deformations constructed in [1] satisfy these properties (or at least only minor technical modifications of them). We now turn to the proof of Theorem 4.1.

Proof of Theorem 4.1. We argue by contradiction and assume that one can find a sequence A_ν of partitions in $\mathcal{P}_{\underline{a},\overline{a},\overline{P}}$ such that $\mathcal{CD}(A_\nu,\nu^{-1},\nu^{-1},\nu)=\emptyset$ for all $\nu\in\mathbb{N}^*$. We would like to find some limit, say A, to the sequence A_ν and then apply Theorem 4.2 to get a contradiction. For notational convenience, let us denote in this proof the different components of the partition A_ν by $A_\nu(i)$, $i\in\{0,\dots,N\}$, (and similarly for any other partition we shall introduce). In other words we need to know that each sequence $A_\nu(i)$ converges to some set $A(i)$ in L^1. This is not true in general because the sets $A_\nu(i)$ are not a priori uniformly bounded as subsets of $\mathbb{R}^n$ and classical embedding theorems only ensure convergence in L^1_{loc}. For that reason we first construct from A_ν a uniformly bounded sequence A'_ν for which we will have the required convergence. We must also do the construction in such a way to be able to go back to the original sequence A_ν when applying Theorem 4.2. Some of the arguments in what follows are similar in spirit to arguments used in [1], but for different purposes though.

Step 1. One can find $C_1=C_1(n,\underline{a},\overline{P})>0$ and for all $\nu\in\mathbb{N}^*$ and $i\in\{1,\dots,N\}$ a point $p_\nu(i)\in\mathbb{R}^n$ such that

$$|A_\nu(i)\cap B(p_\nu(i),1)|\ge C_1,$$

see for instance [9, Lemma 4.1]. This essentially comes from the fact that a set with positive and finite Lebesgue measure and finite perimeter cannot spread out too much and must meet in a substantial way (which depends only on a lower bound for its Lebesgue measure and an upper bound for its perimeter) at least one ball with unit radius. Next we fix some $R=R(n,N,\underline{a},\overline{a},\overline{P})\ge 4$ large enough so that

$$|B(0,R/2)|-N\overline{a}\ge C_1.$$

Step 2. Up to a subsequence one can always assume that

$$\lim_{\nu\to+\infty}\|p_\nu(i)-p_\nu(j)\|$$

exists in $\mathbb{R}^+\cup\{+\infty\}$ for all $i,j\in\{1,\dots,N\}$. Let $\Lambda_1,\dots,\Lambda_s$ be a partitioning of $\{1,\dots,N\}$ so that $k,l\in\Lambda_r$ if and only if there exist $k_1=k,\dots,k_t=l$ in $\{1,\dots,N\}$ such that

$$\lim_{\nu\to+\infty}\|p_\nu(k_j)-p_\nu(k_{j+1})\|\le 2\,R$$

for $j=1,\dots,t-1$. We also assume, passing to a subsequence if necessary, that for all $r\in\{1,\dots,s\}$ and $k,l\in\Lambda_r$,

$$\lim_{\nu\to+\infty}p_\nu(k)-p_\nu(l)$$

exists. Finally for each $r\in\{1,\dots,s\}$, we fix an index $k_r\in\Lambda_r$.

Step 3. For each $\nu \in \mathbb{N}^*$ and $r \in \{1, \dots, s\}$, we set

$$B_{\nu,r} = \cup_{k\in\Lambda_r} B(p_\nu(k), R).$$

Note that if $l, r \in \{1, \dots, s\}$ with $l \neq r$ we have

$$\lim_{\nu\to+\infty} \|p_\nu(i) - p_\nu(j)\| > 2\,R$$

for all $i \in \Lambda_l$ and $j \in \Lambda_r$, hence $B_{\nu,l} \cap B_{\nu,r} = \emptyset$ if ν is large enough. For each $r \in \{1, \dots, s\}$ and $i \in \{1, \dots, N\}$, we set

$$A_{\nu,r}(i) = A_\nu(i) \cap B_{\nu,r}.$$

Step 4. Let $C_2 = C_2(N) \geq 2$ be a constant chosen large enough so that, for all ν large enough, we have

$$\|p_\nu(j) - p_\nu(k_r)\| \leq C_2 R/2$$

for all $r \in \{1, \dots, s\}$ and $j \in \Lambda_r$. Then we have

$$B_{\nu,r} \subset B(p_\nu(k_r), C_2 R).$$

For each $\nu \in \mathbb{N}^*$, $r \in \{1, \dots, s\}$, and $i \in \{1, \dots, N\}$, we set

$$A'_{\nu,r}(i) = \tau_{\nu,r}(A_{\nu,r}(i)),$$

where $\tau_{\nu,r}$ is the translation defined by $\tau_{\nu,r}(x) = x - p_\nu(k_r) + (0, \dots, 0, 2\,C_2 R\, r)$. We have for all ν large enough,

$$A'_{\nu,r}(i) \subset \tau_{\nu,r}(B_{\nu,r}) \subset B(r),$$

where $B(r) = B((0, \dots, 0, 2\,C_2 R\, r), C_2 R)$. Note that for $l, r \in \{1, \dots, s\}$, $l \neq r$, we have $B(l) \cap B(r) = \emptyset$. Finally for $i \in \{1, \dots, N\}$, we set

$$A'_\nu(i) = \cup_{r=1}^{s} A'_{\nu,r}(i)$$

and

$$\begin{aligned} A'_\nu(0) &= \mathbb{R}^n \setminus \cup_{i=1}^{N} A'_\nu(i), \\ A'_\nu &= (A'_\nu(0), \dots, A'_\nu(N)). \end{aligned}$$

Step 5. By construction we have, for some constant $C_3 = C_3(n, N, \underline{a}, \overline{a}, \overline{P}) > 0$ and all ν large enough,

$$A'_\nu(i) \subset B(0, C_3)$$

for each $i \in \{1, \dots, N\}$. Moreover, recalling that $A_\nu \in \mathcal{A}$, we have for ν large enough,

$$\begin{aligned} P(A'_\nu(i)) &\leq \sum_{r=1}^{s} P(A'_{\nu,r}(i)) = \sum_{r=1}^{s} P(A_{\nu,r}(i)) \\ &\leq \sum_{r=1}^{s} P(A_\nu(i), B_{\nu,r}) + P(B_{\nu,r}) \\ &\leq P(A_\nu(i)) + \sum_{j=1}^{N} P(B(p_\nu(j), R)) \leq C \end{aligned}$$

for some suitable $C = C(n, N, \underline{a}, \overline{a}, \overline{P}) > 0$. Hence we have uniformly bounded sequences of sets with uniformly bounded perimeter and thus, passing to a subsequence, one can assume that for each $i \in \{1, \ldots, N\}$ the set $A'_\nu(i)$ converges in L^1 to some set $A(i)$ with finite perimeter (in the sense of convergence of the corresponding characteristic functions, see for instance [7, Theorem 1.19]). Note that it follows that the sequence $A'_\nu(0)$ converges to $A(0)$ in L^1 where

$$A(0) = \mathbb{R}^n \setminus \cup_{i=1}^N A(i).$$

Let us set

$$A = (A(0), \ldots, A(N)) \in \mathcal{P}.$$

Step 6. For each $r \in \{1, \ldots, s\}$ and $j \in \Lambda_r$, we set

$$p_j = \lim_{\nu \to +\infty} \tau_{\nu,r}(p_\nu(j)),$$

(this limit is well defined because of Step 2) and

$$\begin{aligned} \Omega_r &= \cup_{j \in \Lambda_r} B(p_j, R/2), \\ \Omega &= \cup_{r=1}^s \Omega_r. \end{aligned}$$

By construction we have $\Omega_l \cap \Omega_r = \emptyset$ whenever $l \neq r$ and we also have that the partitions A'_ν and $\tau_{\nu,r}(A_\nu)$ coincide on the open set Ω_r provided ν is large enough, that is,

$$A'_\nu(i) \cap \Omega_r = \tau_{\nu,r}(A_\nu(i)) \cap \Omega_r \tag{4.6}$$

for all $r \in \{1, \ldots, s\}$ and all $i \in \{1, \ldots, N\}$. Indeed for each $r \in \{1, \ldots, s\}$ and $j \in \Lambda_r$ set $q_\nu(j) = \tau_{\nu,r}^{-1}(p_j)$. We have $q_\nu(j) - p_\nu(j) = p_j - \tau_{\nu,r}(p_\nu(j))$ and it follows that, if ν is large enough,

$$B(p_\nu(j), R/4) \subset B(q_\nu(j), R/2) \subset B(p_\nu(j), R).$$

(The left inclusion will be used later in Step 7.) In particular we get for all ν large enough and all $r \in \{1, \ldots, s\}$,

$$\begin{aligned} \Omega_r &= \tau_{\nu,r}(\cup_{j \in \Lambda_r} B(q_\nu(j), R/2)) \\ &\subset \tau_{\nu,r}(\cup_{j \in \Lambda_r} B(p_\nu(j), R)) \\ &= \tau_{\nu,r}(B_{\nu,r}) \subset B(r). \end{aligned}$$

Since $B(l) \cap B(r) = \emptyset$ whenever $l \neq r$, it follows that the sets Ω_r are pairwise disjoint. Next (4.6) follows automatically from the previous inclusions and the construction of $A'_\nu(i)$ in Step 4 and that of $A_{\nu,r}(i)$ in Step 3.

Step 7. We have

$$|A(i) \cap \Omega| > 0$$

for all $i \in \{0, \ldots, N\}$. Indeed if $i \in \{1, \ldots, N\}$ and $r \in \{1, \ldots, s\}$ are such that $i \in \Lambda_r$, then

$$\begin{aligned} |A'_\nu(i) \cap \Omega| &\geq |A'_{\nu,r}(i) \cap B(p_i, R/2)| \\ &= |A_{\nu,r}(i) \cap B(q_\nu(i), R/2)| \\ &\geq |A_\nu(i) \cap B(p_\nu(i), R/4)| \geq C_1 > 0 \end{aligned}$$

for all ν large enough, by choice of C_1 in Step 1 and remembering that $R \geq 4$. Passing to the limit when ν goes to infinity, we get the conclusion for $i \in \{1,\dots,N\}$. Finally, if ν is large enough, we have by construction $|A'_\nu(i) \cap \Omega| \leq |A_\nu(i)| \leq \overline{a}$ for all $i \in \{1,\dots,N\}$ hence $|A(i) \cap \Omega| \leq \overline{a}$ if $i \neq 0$. Then, for any $j \in \{1,\dots,N\}$, we get

$$\begin{aligned}|A(0) \cap \Omega| &\geq |A(0) \cap B(p_j, R/2)|\\ &\geq |B(p_j, R/2)| - |\cup_{i=1}^N A(i)|\\ &\geq |B(p_j, R/2)| - N\overline{a} \geq C_1 > 0\end{aligned}$$

by choice of R in Step 1.

Step 8. It follows from Step 7 that one can apply Theorem 4.2 to the partition A and the open set Ω and find corresponding ε, δ, C, and $\eta > 0$. Then, if ν is large enough, we have $|A'_\nu \triangle A| \leq \eta$ so that one can find $(\psi^1_\nu, \psi^2_\nu) \in \mathcal{CD}_\Omega(A'_\nu, \varepsilon, \delta, C)$ with

$$\{x \in \Omega_r;\ \psi^j_\nu(v,x) \neq x\} \cup \psi^j_\nu(v, \{x \in \Omega_r;\ \psi^j_\nu(v,x) \neq x\}) \subset\subset \Omega_r$$

for $j \in \{1,2\}$, all $r \in \{1,\dots,s\}$, and $v \in (-\varepsilon,\varepsilon)^N$. We set

$$\varphi^j_\nu(v,x) = (\tau^{-1}_{\nu,r} \circ \psi^j_\nu(v,.) \circ \tau_{\nu,r})(x)$$

if $x \in \tau^{-1}_{\nu,r}(\Omega_r)$ and

$$\varphi^j_\nu(v,x) = x$$

otherwise. Remember that, if ν is large enough, $\tau^{-1}_{\nu,r}(\Omega_r) \subset B_{\nu,r}$ and $B_{\nu,r} \cap B_{\nu,l} = \emptyset$ for $l \neq r$ (see Step 3). Hence $\tau^{-1}_{\nu,r}(\Omega_r) \cap \tau^{-1}_{\nu,l}(\Omega_l) = \emptyset$ and $\varphi^j_\nu(v,.)$ is well defined.

Next we clearly have, at least if ν is large enough,

$$\{x \in \tau^{-1}_{\nu,r}(\Omega_r);\ \varphi^j_\nu(v,x) \neq x\} \cup \varphi^j_\nu(v, \{x \in \tau^{-1}_{\nu,r}(\Omega_r);\ \varphi^j_\nu(v,x) \neq x\}) \subset\subset \tau^{-1}_{\nu,r}(\Omega_r)$$

and

$$\mathcal{O}(\varphi^j_\nu(v,.)) \subset\subset \cup_{r=1}^s \tau^{-1}_{\nu,r}(\Omega_r).$$

Since each $\psi^j_\nu(v,.)$ is a C^1 diffeomorphism, it follows in particular that $\varphi^j_\nu(v,.)$ is a C^1 diffeomorphism for $j \in \{1,2\}$ and each $v \in (-\varepsilon,\varepsilon)^N$.

To compute the variation of the measure and of the energy of A_ν under φ^j_ν, we clearly need to worry only about what happens inside the pairwise disjoint open sets $\tau^{-1}_{\nu,r}(\Omega_r)$. Recalling that the partitions $\tau_{\nu,r}(A_\nu)$ and A'_ν coincide on Ω_r (see (4.6)), we have

$$\begin{aligned}|\varphi^j_\nu(v, A_\nu(i))| - |A_\nu(i)| &= \sum_{r=1}^s |\varphi^j_\nu(v, A_\nu(i) \cap \tau^{-1}_{\nu,r}(\Omega_r))| - |A_\nu(i) \cap \tau^{-1}_{\nu,r}(\Omega_r)|\\ &= \sum_{r=1}^s |\psi^j_\nu(v, A'_\nu(i) \cap \Omega_r)| - |A'_\nu(i) \cap \Omega_r|\\ &= |\psi^j_\nu(v, A'_\nu(i))| - |A'_\nu(i)| = v_i\end{aligned}$$

for all ν large enough and $i \in \{1,\dots,N\}$ according to the property (4.3) of the map $\psi^j_\nu \in \mathcal{D}_\Omega(A'_\nu, \varepsilon, C)$. Similarly, using (4.4) for ψ^j_ν, we have for all ν large enough and $i \in \{1,\dots,N\}$,

$$\begin{aligned}|\varphi^j_\nu(v, A_\nu(i)) \triangle A_\nu(i)| &= |\psi^j_\nu(v, A'_\nu(i)) \triangle A'_\nu(i)|\\ &\leq CP(A'_\nu(i), \Omega)|v| \leq CP(A_\nu(i))|v|.\end{aligned}$$

The last inequality follows from the invariance of the perimeter under translations together with (4.6). Indeed we have for all ν large enough,

$$\begin{aligned} P(A'_\nu(i),\Omega) &= \sum_{r=1}^{s} P(A'_\nu(i),\Omega_r) = \sum_{r=1}^{s} P(\tau_{\nu,r}(A_\nu(i)),\Omega_r) \\ &= \sum_{r=1}^{s} P(A_\nu(i),\tau_{\nu,r}^{-1}(\Omega_r)) \le P(A_\nu(i)). \end{aligned}$$

Next we consider an energy $E \in \mathcal{E}(\overline{F})$ defined with respect to some family (F_{ij}). Note that

$$E(\tau(\tilde{A}),\tau(U)) = E_\tau(\tilde{A},U)$$

for any partition $\tilde{A} \in \mathcal{P}$, any measurable set $U \subset \mathbb{R}^n$, and any translation $\tau : \mathbb{R}^n \to \mathbb{R}^n$ where $E_\tau \in \mathcal{E}(\overline{F})$ is defined with respect to the family $((x,y) \mapsto F_{ij}(\tau(x),y))$. It follows that if U is a measurable set, then, for ν large enough,

$$\begin{aligned} &E(\varphi^j_\nu(v,A_\nu),\varphi^j_\nu(v,U)) - E(A_\nu,U) \\ &\quad = \sum_{r=1}^{s} E(\tau_{\nu,r}^{-1} \circ \psi^j_\nu(v,A'_\nu),\tau_{\nu,r}^{-1}(\Omega_r) \cap \varphi^j_\nu(v,U)) - E(\tau_{\nu,r}^{-1}(A'_\nu),\tau_{\nu,r}^{-1}(\Omega_r) \cap U) \\ &\quad = \sum_{r=1}^{s} E_{\tau_{\nu,r}^{-1}}(\psi^j_\nu(v,A'_\nu),\Omega_r \cap \psi^j_\nu(v,\tau_{\nu,r}(U))) - E_{\tau_{\nu,r}^{-1}}(A'_\nu,\Omega_r \cap \tau_{\nu,r}(U)). \end{aligned}$$

Since $E_{\tau_{\nu,r}^{-1}} \in \mathcal{E}(\overline{F})$ and $\psi^j_\nu(v,\Omega_r \cap \tau_{\nu,r}(U)) = \Omega_r \cap \psi^j_\nu(v,\tau_{\nu,r}(U))$ for all ν large enough and all $r \in \{1,\dots,s\}$, it follows from the property (4.5) of ψ^j_ν that

$$\begin{aligned} &E_{\tau_{\nu,r}^{-1}}(\psi^j_\nu(v,A'_\nu),\Omega_r \cap \psi^j_\nu(v,\tau_{\nu,r}(U))) - E_{\tau_{\nu,r}^{-1}}(A'_\nu,\Omega_r \cap \tau_{\nu,r}(U)) \\ &\qquad \le C\,P(A'_\nu,\Omega_r \cap \tau_{\nu,r}(U))\,|v| = C\,P(A_\nu,\tau_{\nu,r}^{-1}(\Omega_r) \cap U)\,|v| \end{aligned}$$

and thus

$$E(\varphi^j_\nu(v,A_\nu),\varphi^j_\nu(v,U)) - E(A_\nu,U) \le C\,P(A_\nu,U)\,|v|$$

for all ν large enough.

Finally by construction it is clear that one can find $\delta' > 0$, may be smaller than δ but still strictly positive and not depending on ν, such that

$$\operatorname{dist}(\mathcal{O}(\varphi^1_\nu(v,.)),\mathcal{O}(\varphi^2_\nu(v,.))) \ge \delta'.$$

Hence taking ν large enough and so that $\nu \ge \max(\varepsilon^{-1},\delta'^{-1},C)$ we have just proved that

$$(\varphi^1_\nu,\varphi^2_\nu) \in \mathcal{CD}(A_\nu,\nu^{-1},\nu^{-1},\nu)$$

which gives the required contradiction and concludes the proof. □

Although we will not need this refinement here one can note that, with essentially the same proof modulo only minor technical modifications, one can get a localized version of Theorem 4.1 similar to that given in Theorem 4.2. More precisely, assume that $\Omega \subset \mathbb{R}^n$ is open and, for simplicity, bounded and that $\Omega = \cup_{l=1}^{s}\Omega_l$ with the sets

Ω_l open and pairwise disjoint. Let $\mathcal{P}_{\underline{a},\overline{P},\Omega}$ denote the class of all partitions $A \in \mathcal{P}$ such that

$$\underline{a} \leq \min_{i=0,\ldots,N} |A_i \cap \Omega| \quad \text{and} \quad P(A,\Omega) \leq \overline{P}$$

for some $\underline{a} > 0$ and $\overline{P} > 0$. Then one can find $\varepsilon_0 > 0$, $\delta_0 > 0$, and $C_0 > 0$, depending only on n, N, $\overline{F}$, $\underline{a}$, and $\overline{P}$, such that for all $A \in \mathcal{P}_{\underline{a},\overline{P},\Omega}$ one can find $(\psi_1,\psi_2) \in \mathcal{CD}_\Omega(A,\varepsilon_0,\delta_0,C_0)$ such that

$$\{x \in \Omega_l\,;\ \psi_j(v,x) \neq x\} \cup \psi_j(v,\{x \in \Omega_l\,;\ \psi_j(v,x) \neq x\}) \subset\subset \Omega_l$$

for all $l \in \{1,\ldots,s\}$, $j \in \{1,2\}$, and $v \in (-\varepsilon_0,\varepsilon_0)^N$.

4.2. Proof of Theorem 3.6. We now prove Theorem 3.6. We let $0 < \underline{F} \leq \overline{F} < +\infty$, $E \in \mathcal{E}(\underline{F},\overline{F})$, $g : [0,+\infty) \to [0,+\infty)$, $a = (a_1,\ldots,a_N)$ be as in the statement and $A \in \mathcal{P}_a$ be a (g,a)-quasiminimal partition with respect to E. Recall that we want to find some suitable $\tilde{g} : [0,+\infty) \to [0,+\infty)$ such that $\lim_{t\to 0} t^{(1-n)/n}\tilde{g}(t) = 0$ and $\delta > 0$ both universal and such that A is a $(\tilde{g},\delta)$-quasiminimal partition with respect to E. Throughout this subsection, when saying that some object is universal, we mean more precisely that it can be chosen depending only on n, N, $\underline{F}$, $\overline{F}$, g, and a.

Set $\underline{a} = \min_{i\geq 1} a_i$ and $\overline{a} = \max_{i\geq 1} a_i$. Next we claim that one can find $\overline{P} > 0$ universal such that $P(A) \leq \overline{P}$. To see this, setting $B_t = B(0,t)$ for all $t > 0$, we choose $R > 0$ large enough so that

$$P(A_i,\mathbb{R}^n \setminus B_R) \leq 1$$

for all $i \in \{0,\ldots,N\}$ and so that one can find N open balls $B_1,\ldots,B_N$ with $B_i \subset\subset B_R$, $|B_i| = a_i$ and $B_i \cap B_j = \emptyset$ for $i \neq j$. By the coarea formula and Tchebytchev's inequality, we have

$$|\{t \in (R,R+1);\ \mathcal{H}^{n-1}(A_i \cap \partial B_t) > C\,\overline{a}\}|_1 \leq \frac{|A_i \cap (B_{R+1} \setminus B_R)|}{C\,\overline{a}} \leq C^{-1},$$

where $|.|_1$ denotes the one dimensional Lebesgue measure. Then, choosing C large enough, depending only on N, and recalling that

$$P(A_i \setminus B_t) = P(A_i,\mathbb{R}^n \setminus \overline{B}_t) + \mathcal{H}^{n-1}(A_i \cap \partial B_t)$$

for a.e. $t > 0$ (see for instance [7, Remark 2.14]), one can find $t \in (R,R+1)$ such that this last equality holds and

$$\mathcal{H}^{n-1}(A_i \cap \partial B_t) \leq C\,\overline{a}$$

for all $i \in \{1,\ldots,N\}$. We set

$$A_i' = (A_i \setminus B_t) \cup B_i'$$

for $i \in \{1,\ldots,N\}$, where $B_i' \subset B_i$ is a ball with $|B_i'| = |A_i \cap B_t|$, and

$$A_0' = \mathbb{R}^n \setminus \cup_{i=1}^N A_i'.$$

We have $A' = (A_0',\ldots,A_N') \in \mathcal{P}_a$ and $A' \triangle A \subset\subset \mathbb{R}^n$. Next, for all $i \in \{1,\ldots,N\}$, we have

$$\begin{aligned} P(A_i') &= P(B_i') + P(A_i \setminus B_t) \\ &\leq C\,\overline{a}^{\frac{n-1}{n}} + P(A_i,\mathbb{R}^n \setminus \overline{B}_t) + \mathcal{H}^{n-1}(A_i \cap \partial B_t) \\ &\leq C\,\overline{a}^{\frac{n-1}{n}} + P(A_i,\mathbb{R}^n \setminus B_R) + C\,\overline{a} \leq C \end{aligned}$$

for some universal constants $C > 0$. When $i = 0$, we have

$$P(A'_0) = P(\mathbb{R}^n \setminus \cup_{i=1}^N A'_i) \leq \sum_{i=1}^n P(A'_i).$$

Then it follows from (2.3) that

$$E(A') \leq \overline{F}\, P(A') \leq C$$

for some suitable universal constant $C > 0$. On the other hand, for $i \in \{1, \ldots, N\}$, we have $|A'_i \triangle A_i| \leq 2\,\overline{a}$ and $A'_0 \triangle A_0 \subset \cup_{i=1}^N A'_i \triangle A_i$, hence $g(|A' \triangle A|) \leq g(2\,N\,\overline{a})$ since g is nondecreasing. Then it follows from (2.3) together with the (g, a)-quasiminimality (3.1) of A that

$$P(A) \leq \underline{F}^{-1}\, E(A) \leq \underline{F}^{-1}\, (E(A') + g(|A' \triangle A|)) \leq \overline{P}$$

for some suitable universal constant $\overline{P} > 0$ as claimed.

Thus we have $A \in \mathcal{P}_{\underline{a},\overline{a},\overline{P}}$ as in section 4.1 and Theorem 4.1 gives some $\varepsilon_0 > 0$, $\delta_0 > 0$, and $C_0 > 0$ universal such that $\mathcal{CD}(A, \varepsilon_0, \delta_0, C_0) \neq \emptyset$. Let $(\psi_1, \psi_2) \in \mathcal{CD}(A, \varepsilon_0, \delta_0, C_0)$ and let us fix $\delta > 0$ universal and small enough so that

$$|B_\delta| < \varepsilon_0 \quad \text{and} \quad 4\delta < \delta_0.$$

Now let $A' \in \mathcal{P}$ be such that $\operatorname{diam}(A' \triangle A) < \delta$ (A' has here nothing to do with the partition constructed a few lines above). To prove the theorem we need to compare the energy of A' with that of A. We first use the suitable deformations given by ψ_1 (or ψ_2) to construct from A' a partition $A'' \in \mathcal{P}_a$. Then we will use the (g, a)-quasiminimality of A to get a comparison between the energies.

For $i \in \{1, \ldots, N\}$ we set $v_i = |A_i| - |A'_i|$. Since $\operatorname{diam}(A' \triangle A) < \delta$ then $A' \triangle A \subset\subset B$ for some ball B with radius δ and thus

$$|v_i| = ||A_i \setminus A'_i| - |A'_i \setminus A_i|| \leq |A'_i \triangle A_i| \leq |B|,$$

that is, $v = (v_1, \ldots, v_N) \in (-\varepsilon_0, \varepsilon_0)^N$ by choice of δ. Next let us assume that

$$2B \cap \mathcal{O}(\psi_1(v, .)) = \emptyset,$$

where $2B$ denotes the ball concentric to B with radius 2δ. Otherwise, since

$$\operatorname{dist}(\mathcal{O}(\psi_1(v, .)), \mathcal{O}(\psi_2(v, .))) \geq \delta_0$$

and $4\delta < \delta_0$, we would have $2B \cap \mathcal{O}(\psi_2(v, .)) = \emptyset$ and would simply use ψ_2 instead of ψ_1 in what follows. Then we set

$$A''_i = (A'_i \cap B) \cup (\psi_1(v, A_i) \cap \mathbb{R}^n \setminus B)$$

and $A'' = (A''_0, \ldots, A''_N)$. Since A'_i coincides with A_i on a neighborhood of $\mathbb{R}^n \setminus B$ and $\psi_1(v, A_i)$ with A_i on B (note that since $\psi_1(v, .)$ is a C^1 diffeomorphism, we have $\psi_1(v, \mathcal{O}(\psi_1(v, .))) = \mathcal{O}(\psi_1(v, .))$) it follows from the property (4.3) of $\psi_1 \in \mathcal{D}(A, \varepsilon_0, C_0)$ that, for all $i \in \{1, \ldots, N\}$,

$$\begin{aligned} |A''_i| &= |A'_i \cap B| + |\psi_1(v, A_i) \cap \mathbb{R}^n \setminus B| \\ &= |A'_i| + |\psi_1(v, A_i)| - |A_i| \\ &= |A'_i| + v_i = |A_i| \end{aligned}$$

hence $A'' \in \mathcal{P}_a$. For the same reasons we have

$$E(A', \mathbb{R}^n \setminus B) = E(A, \mathbb{R}^n \setminus B),$$
$$E(\psi_1(v, A), B) = E(A, B).$$

Next A'' coincides with A' on B and with $\psi_1(v, A)$ on a neighborhood of $\mathbb{R}^n \setminus B$, hence

$$\begin{aligned} E(A'') &= E(A', B) + E(\psi_1(v, A), \mathbb{R}^n \setminus B) \\ &= E(A') + E(\psi_1(v, A)) - E(A) \\ &\leq E(A') + C_0\, P(A)\, |v| \\ &\leq E(A') + C_0\, \overline{P}\, |A' \triangle A| \end{aligned}$$

according to (4.5). Finally we have

$$A'' \triangle A = (A' \triangle A) \cup (\psi_1(v, A) \triangle A)$$

and thus, since, thanks to (4.4),

$$|\psi_1(v, A_i) \triangle A_i| \leq C_0\, P(A_i)\, |v| \leq C_0\, P(A_i)\, |A' \triangle A|$$

for all $i \in \{1, \ldots, N\}$ and $|\psi_1(v, A_0) \triangle A_0| \leq |\cup_{i=1}^N \psi_1(v, A_i) \triangle A_i|$, we get

$$|A'' \triangle A| \leq (2\, C_0\, \overline{P} + 1)\, |A' \triangle A|.$$

To conclude we use the (g, a)-quasiminimality of A to compare $E(A)$ and $E(A'')$,

$$\begin{aligned} E(A) &\leq E(A'') + g(|A'' \triangle A|) \\ &\leq E(A') + C_0\, \overline{P}\, |A' \triangle A| + g((2\, C_0\, \overline{P} + 1)\, |A' \triangle A|) \end{aligned}$$

because g is nondecreasing. Then setting $\tilde{g}(t) = C_0\, \overline{P}\, t + g((2\, C_0\, \overline{P} + 1)\, t)$ we have just proved that A is $(\tilde{g}, \delta)$-quasiminimal and this concludes the proof.

5. Ahlfors-regularity and condition B. This section is devoted to the study of locally quasiminimal partitions. Namely, we fix some $0 < \underline{F} \leq \overline{F} < +\infty$, $E \in \mathcal{E}_0(\underline{F}, \overline{F})$, $g : [0, +\infty) \to [0, +\infty)$ such that $\lim_{t \to 0} t^{(1-n)/n} g(t) = 0$ and $\delta > 0$ and we prove the following theorem, Theorem 5.1, together with Theorem 3.7.

THEOREM 5.1. *Let A be a (g, δ)-quasiminimal partition with respect to E and let S denote its set of interfaces. Then S is a closed Ahlfors-regular set with Ahlfors-regularity constants that can be chosen depending only on n, $\underline{F}$, $\overline{F}$, g, and δ, and S satisfies condition B with a condition B constant that can be chosen depending only on n, $\underline{F}$, $\overline{F}$, g, δ, and N.*

We shall actually get slightly refined properties, see especially Lemma 5.7 (see also section 6 for a particular case). For simplicity, and with no loss of generality, we assume that $\delta < 1$. We also assume that A is a reduced (g, δ)-quasiminimal partition (remember that the assumption on A to be reduced is not restrictive, see the comments before Theorem 3.7). We begin with the upper bound in the Ahlfors-regularity.

LEMMA 5.2. *There exist a constant $C_0 = C_0(n, \underline{F}, \overline{F}) > 0$ and a radius $r_0 = r_0(n, g, \delta) \leq 1$ such that, for all $x \in \mathbb{R}^n$ and $r \leq r_0$, we have*

$$\mathcal{H}^{n-1}(S \cap B(x, r)) \leq C_0\, r^{n-1}.$$

Proof. The proof is based on a simple comparison argument. We take $x \in \mathbb{R}^n$, $r > 0$ and set

$$\begin{aligned} A'_0 &= A_0 \cup B(x,r), \\ A'_i &= A_i \backslash B(x,r) \end{aligned}$$

for $i \in \{1, \ldots, N\}$. Then $A' = (A'_0, \ldots, A'_N) \in \mathcal{P}$ and, since $A' \triangle A \subset B(x,r)$, we have $\operatorname{diam}(A' \triangle A) < \delta$ provided $r < \delta/2$. Moreover $|A' \triangle A| \leq |B(x,r)|$ hence, if r is small enough, depending only on n and g, we get from the behavior of g near 0 that

$$g(|A' \triangle A|) \leq r^{n-1}.$$

On the other hand, A' and A coincide on the open set $\mathbb{R}^n \setminus \overline{B}(x,r)$ hence they have the same energy there and it follows from the (g,δ)-quasiminimality of A that

$$E(A, \overline{B}(x,r)) \leq E(A', \overline{B}(x,r)) + g(|A' \triangle A|) \leq E(A', \overline{B}(x,r)) + r^{n-1}.$$

We have

$$\begin{aligned} E(A, \overline{B}(x,r)) &\geq \sum_{i=0}^{N} \sum_{\substack{j=0 \\ j \neq i}}^{N} \int_{\partial^* A_i \cap \partial^* A_j} \mathbb{1}_{B(x,r)}(y) F_{ij}(y, \nu_{A_i}(y)) \; d\mathcal{H}^{n-1} \\ &\geq \underline{F} \sum_{i=0}^{N} \sum_{\substack{j=0 \\ j \neq i}}^{N} \mathcal{H}^{n-1}(\partial^* A_i \cap \partial^* A_j \cap B(x,r)) \\ &= 2\, \underline{F}\, \mathcal{H}^{n-1}(S \cap B(x,r)), \end{aligned}$$

where the last equality follows from (2.2). On the other hand, we clearly have $E(A', B(x,r)) = 0$. Next, any point y in $\partial B(x,r)$ is a point of lower density at least $1/2$ for A'_0 because $B(x,r) \subset A'_0$. Then, since $\partial_* A'_i \subset A'_i(1/2)$, such points $y \in \partial B(x,r)$ can only belong to at most a single set $\partial_* A'_i$ with $i \neq 0$. It follows that

$$\mathcal{H}^{n-1}(\partial^* A'_i \cap \partial^* A'_j \cap \partial B(x,r)) = 0$$

whenever i, $j \in \{1, \ldots, N\}$, $i \neq j$, and hence

$$\begin{aligned} &E(A', \partial B(x,r)) \\ &\quad = \sum_{i=1}^{N} \int_{\partial^* A'_0 \cap \partial^* A'_i} \mathbb{1}_{\partial B(x,r)}(y) (F_{0i}(y, \nu_{A'_0}(y)) + F_{i0}(y, \nu_{A'_i}(y)))\, d\mathcal{H}^{n-1} \\ &\quad \leq 2\, \overline{F} \sum_{i=1}^{N} \mathcal{H}^{n-1}(\partial^* A'_0 \cap \partial^* A'_i \cap \partial B(x,r)) \\ &\quad = 2\, \overline{F}\, \mathcal{H}^{n-1}(\partial^* A'_0 \cap \partial B(x,r)) \leq C\, r^{n-1} \end{aligned}$$

(remember the comments after Definition 2.3 for the equality on the last line). Then we get

$$\begin{aligned} \mathcal{H}^{n-1}(S \cap B(x,r)) &\leq (2\, \underline{F})^{-1}\, E(A, \overline{B}(x,r)) \\ &\leq (2\, \underline{F})^{-1}\, (E(A', \partial B(x,r)) + r^{n-1}) \leq C_0\, r^{n-1} \end{aligned}$$

for some $C_0 = C_0(n, \underline{F}, \overline{F}) > 0$ as wanted. $\square$

Analogues of the next lemma are quite standard in the regularity theory of (quasi-)minimal sets and is technically one of our main steps. It says that if the proportion of some component inside a ball is large enough, then this component contains one half the ball.

LEMMA 5.3. *There exists* $\eta_0 = \eta_0(n, \underline{F}, \overline{F}, g) > 0$ *such that, for all* $i \in \{0, \dots, N\}$, $x \in \mathbb{R}^n$, *and* $r \le \delta/2$, *if* $|B(x,r) \backslash A_i| \le \eta_0\, r^n$, *then* $B(x, r/2) \subset A_i$.

Proof. Since A is a reduced partition, and hence $A_i = A_i(1)$, it is sufficient to show that $|B(x,r/2)\backslash A_i| = 0$ when $i \in \{0, \dots, N\}$, $x \in \mathbb{R}^n$, and $r \le \delta/2$ are fixed as in the statement. For $t > 0$, we set $m(t) = |B(x,t)\backslash A_i|$ and assume that $m(r) \le \eta_0\, r^n$ for some $\eta_0 > 0$ to be fixed small later. Arguing by contradiction, we also assume that $m(r/2) > 0$. Using a similar comparison argument as in the proof of Lemma 5.2, we set for $t \in (r/2, r)$,

$$A'_i = A_i \cup B(x,t),$$

and, for $j \neq i$,

$$A'_j = A_j \backslash B(x,t).$$

Then $A' = (A'_0, \dots, A'_N) \in \mathcal{P}$ and, since $A' \triangle A \subset B(x,t)$, we have $\operatorname{diam}(A' \triangle A) < \delta$. Since A' and A coincide on the open set $\mathbb{R}^n \setminus \overline{B}(x,t)$, they have the same energy there and we get from the (g,δ)-quasiminimality of A that

$$E(A, \overline{B}(x,t)) \le E(A', \overline{B}(x,t)) + g(|A' \triangle A|).$$

We have

$$\begin{aligned} E(A, \overline{B}(x,t)) &\ge \sum_{\substack{j=0 \\ j\neq i}}^{N} \int_{\partial^* A_i \cap \partial^* A_j} \mathbb{1}_{B(x,t)}(y) F_{ij}(y, \nu_{A_i}(y))\, d\mathcal{H}^{n-1} \\ &\ge \underline{F} \sum_{\substack{j=0 \\ j\neq i}}^{N} \mathcal{H}^{n-1}(\partial^* A_i \cap \partial^* A_j \cap B(x,t)) \\ &= \underline{F}\, \mathcal{H}^{n-1}(\partial^* A_i \cap B(x,t)) \\ &\ge C\ \min(|B(x,t) \setminus A_i|, |A_i \cap B(x,t)|)^{\frac{n-1}{n}} = C\, m(t)^{\frac{n-1}{n}}, \end{aligned}$$

where the equality follows from the comments after Definition 2.3 and the last inequality from the relative isoperimetric inequality for balls, provided η_0 is small enough, how small depending only on n. On the other hand, $E(A', B(x,t)) = 0$ and, arguing as in Lemma 5.2 (see the estimation of $E(A', \partial B(x,r))$ there), we have

$$E(A', \partial B(x,t)) \le 2\, \overline{F}\, \mathcal{H}^{n-1}(\partial^* A'_i \cap \partial B(x,t)) = 2\, \overline{F}\, P(A_i \cup B(x,t), \partial B(x,t)).$$

Combining these inequalities we get

$$C\, m(t)^{\frac{n-1}{n}} \le P(A_i \cup B(x,t), \partial B(x,t)) + g(|A' \triangle A|) \tag{5.1}$$

for some suitable constant $C = C(n, \underline{F}, \overline{F}) > 0$. We have

$$|A' \triangle A| = m(t) \le \eta_0\, r^n \le \eta_0$$

thus, if η_0 is small enough,

$$g(|A'\triangle A|) \leq \frac{C\, m(t)^{\frac{n-1}{n}}}{2},$$

where C is the constant that shows up in (5.1). Then, recalling that

$$P(A_i \cup B(x,t), \partial B(x,t)) = m'(t)$$

for a.e. $t > 0$ (this follows for instance from [7, Remark 2.14] and the coarea formula), it follows that

$$m(t)^{\frac{n-1}{n}} \leq C\, P(A_i \cup B(x,t), \partial B(x,t)) \leq C\, m'(t)$$

for a.e. $t \in (r/2, r)$ and for some suitable constant $C = C(n, \underline{F}, \overline{F}) > 0$. Finally, since by assumption $m(t) > 0$ for all $t \geq r/2$, we can rewrite this inequality in the following way:

$$C \leq m(t)^{\frac{1-n}{n}}\, m'(t),$$

and, integrating over $(r/2, r)$, we get

$$r \leq C\,(m(r)^{\frac{1}{n}} - m(r/2)^{\frac{1}{n}}) \leq C\, m(r)^{\frac{1}{n}} \leq C\, \eta_0^{\frac{1}{n}}\, r$$

for some $C = C(n, \underline{F}, \overline{F}) > 0$. This is impossible if η_0 is small enough and gives the contradiction. □

We draw in the following lemma some easy consequences of Lemma 5.3 which help to clarify the picture and give the first part of Theorem 3.7.

Lemma 5.4. *For all $i \in \{0, \dots, N\}$, A_i is open and $S = \cup_{i=0}^N \partial A_i$.*

Proof. Let $i \in \{0, \dots, N\}$ be fixed and $x \in A_i$. Since A is reduced, x is point of density 1 for A_i hence $|B(x,r)\backslash A_i| \leq \eta_0\, r^n$ for all r small enough where η_0 is given by Lemma 5.3. Then according to that lemma we have $B(x, r/2) \subset A_i$ hence A_i is open. Next we always have $S \subset \cup_{i=0}^N \partial A_i$. On the other hand, if $x \in \partial A_i$ for some $i \in \{0, \dots, N\}$, then $x \notin \cup_{j=0}^N A_j$ because these sets are open and pairwise disjoint, and hence $x \in S$ as required (remember what follows Definition 2.4). □

We now prove the lower bound in the Ahlfors-regularity.

Lemma 5.5. *There exists $c_0 = c_0(n, \underline{F}, \overline{F}, g) > 0$ such that, for all $x \in S$ and $r \leq \delta/2$, we have*

$$\mathcal{H}^{n-1}(S \cap B(x,r)) \geq c_0\, r^{n-1}.$$

Proof. Let $c_0 > 0$ be a constant that will be fixed small later, depending only on n, $\underline{F}$, $\overline{F}$, and g, and assume by contradiction that one can find $x \in S$ and $r \leq \delta/2$ such that

$$\mathcal{H}^{n-1}(S \cap B(x,r)) < c_0\, r^{n-1}.$$

First one can find $i \in \{0, \dots, N\}$ such that

$$|A_i \cap B(x,r)| > |B(x,r)|/2.$$

Otherwise the relative isoperimetric inequality for balls would give

$$\begin{aligned} |A_i \cap B(x,r)|^{\frac{n-1}{n}} &= \min(|A_i \cap B(x,r)|, |B(x,r) \setminus A_i|)^{\frac{n-1}{n}} \\ &\leq C\, \mathcal{H}^{n-1}(\partial^* A_i \cap B(x,r)) \end{aligned}$$

for all $i \in \{0, \ldots, N\}$. Using (2.2) we would get

$$\begin{aligned}\sum_{i=0}^{N} |A_i \cap B(x,r)|^{\frac{n-1}{n}} &\leq C \sum_{i=0}^{N} \mathcal{H}^{n-1}(\partial^* A_i \cap B(x,r)) \\ &\leq C\, \mathcal{H}^{n-1}(S \cap B(x,r)) \leq C\, c_0\, r^{n-1}.\end{aligned}$$

On the other hand, since $|A_i \cap B(x,r)| \leq |B(x,r)|/2$ for all $i \in \{0, \ldots, N\}$, we have

$$\sum_{i=0}^{N} |A_i \cap B(x,r)|^{\frac{n-1}{n}} \geq 2^{\frac{1}{n}}\, |B(x,r)|^{-\frac{1}{n}} \sum_{i=0}^{N} |A_i \cap B(x,r)| = 2^{\frac{1}{n}}\, |B(x,r)|^{\frac{n-1}{n}}.$$

Hence we would have

$$|B(x,r)|^{\frac{n-1}{n}} \leq C\, c_0\, r^{n-1}$$

for some dimensional constant $C > 0$ but this is impossible if c_0 is small enough. Thus let us fix $i \in \{0, \ldots, N\}$ such that $|A_i \cap B(x,r)| > |B(x,r)|/2$. Then once again by the relative isoperimetric inequality for balls we have

$$\begin{aligned}|B(x,r) \backslash A_i| &\leq C\, \mathcal{H}^{n-1}(\partial^* A_i \cap B(x,r))^{\frac{n}{n-1}} \\ &\leq C\, \mathcal{H}^{n-1}(S \cap B(x,r))^{\frac{n}{n-1}} \\ &\leq C\, c_0^{\frac{n}{n-1}}\, r^n,\end{aligned}$$

and Lemma 5.3 implies that $B(x, r/2) \subset A_i$ provided $c_0 = c_0(n, \underline{F}, \overline{F}, g)$ is small enough. In particular $x \in A_i$. But A is a reduced partition and thus $S \cap A_i = \emptyset$ which gives the required contradiction. $\square$

Corollary 5.6. *The set of interfaces S is a closed Ahlfors-regular set with regularity constants that can be chosen depending only on n, $\underline{F}$, $\overline{F}$, g, and δ.*

Proof. First, according to Lemma 5.4, S is the finite union of the topological boundaries ∂A_i and hence is closed. Next Lemmas 5.2 and 5.5 give the required inequalities

$$c_0\, r^{n-1} \leq \mathcal{H}^{n-1}(S \cap B(x,r)) \leq C_0\, r^{n-1}$$

for all $x \in S$ and $r \leq r_1 = \min(\delta/2, r_0)$. If $r \in [r_1, 1]$ we obviously have

$$(c_0\, r_1^{n-1})\, r^{n-1} \leq c_0\, r_1^{n-1} \leq \mathcal{H}^{n-1}(S \cap B(x, r_1)) \leq \mathcal{H}^{n-1}(S \cap B(x,r)).$$

Finally the upper inequality in the Ahlfors-regularity for $r \in [r_1, 1]$ follows from a standard covering argument that we include here for the sake of completeness. Let us consider a maximal family $\mathcal{A}$ of points in $S \cap B(x,r)$ at mutual distance greater than or equal to r_1. The balls $B(y, r_1/2)$, $y \in \mathcal{A}$, are pairwise disjoint and contained in $B(x, 2r)$, hence,

$$\begin{aligned}\text{card } \mathcal{A} &\leq C \sum_{y \in \mathcal{A}} |B(y, r_1/2)| = C\, |\cup_{y \in \mathcal{A}} B(y, r_1/2)| \\ &\leq C\, |B(x, 2r)| \leq C\, r^n \leq C\, r^{n-1}\end{aligned}$$

for some suitable constants $C > 0$. Since $S \cap B(x,r) \subset \cup_{y \in \mathcal{A}} B(y, r_1)$, it follows that

$$\begin{aligned}\mathcal{H}^{n-1}(S \cap B(x,r)) &\leq \sum_{y \in \mathcal{A}} \mathcal{H}^{n-1}(S \cap B(y, r_1)) \\ &\leq C\, (\text{card } \mathcal{A})\, r_1^{n-1} \leq C\, r^{n-1}\end{aligned}$$

for some suitable constant $C = C(n, \underline{F}, \overline{F}, g, \delta) > 0$. $\square$

We prove in the next lemma a slightly refined version of condition B.

LEMMA 5.7. *There exists a constant* $C_1 = C_1(n, N, \underline{F}, \overline{F}, g, \delta) > 0$ *such that, for all* $x \in S$ *and* $r \leq 1$*, there exist* i*,* $j \in \{0, \dots, N\}$*,* $i \neq j$*, and two balls* B_i *and* B_j *with radius* $C_1\, r$ *such that* $B_i \subset A_i \cap B(x, r)$ *and* $B_j \subset A_j \cap B(x, r)$*.*

Proof. Let $x \in S$ and $r \leq 1$ be fixed. The set S is Ahlfors-regular according to Corollary 5.6 hence Lemma 5.8, to be proved below, applied with $B = B(x, r/2)$ gives us a point $y \in B(x, r/2)$ such that $B(y, \gamma\, r) \cap S = \emptyset$ for some $\gamma = \gamma(n, \underline{F}, \overline{F}, g, \delta) \leq 1/4$. Recalling that $\mathbb{R}^n = (\cup_{i=0}^{N} A_i) \cup S$ because A is reduced and that S is, according to Lemma 5.4, the union of the topological boundaries of the open and disjoint sets A_i, it follows automatically that one can find $i \in \{0, \dots, N\}$ such that $B(y, \gamma\, r) \subset A_i$. Note that $B(y, \gamma\, r) \subset B(x, r)$.

Next, since $x \in S$ and $A_i \cap S = \emptyset$, we have $|B(x, r/2) \setminus A_i| \geq C\, r^n$ for some constant $C = C(n, \underline{F}, \overline{F}, g, \delta) > 0$, because otherwise Lemma 5.3 would give $x \in A_i$. Let us choose $j \in \{0, \dots, N\} \setminus \{i\}$ such that

$$|B(x, r/2) \cap A_j| = \max_{l \in \{0, \dots, N\} \setminus \{i\}} |B(x, r/2) \cap A_l|.$$

Then we have

$$\begin{aligned} |B(x, r/2) \cap A_j| &\geq N^{-1} \sum_{\substack{l=0 \\ l \neq i}}^{N} |B(x, r/2) \cap A_l| \\ &= N^{-1}\, |B(x, r/2) \setminus A_i| \geq \beta\, r^n \end{aligned}$$

for some $\beta = \beta(n, N, \underline{F}, \overline{F}, g, \delta) > 0$. Then, using once again Lemma 5.8 with now $B = B(x, r/2) \cap A_j$, one can find $\gamma' = \gamma'(n, N, \underline{F}, \overline{F}, g, \delta) \leq 1/4$ and $y' \in B(x, r/2) \cap A_j$ such that $B(y', \gamma'\, r) \subset B(x, r)$ does not meet S. In particular $B(y', \gamma'\, r)$ does not meet ∂A_j hence $B(y', \gamma'\, r) \subset A_j \cap B(x, r)$. Finally $B_i = B(y, C_1\, r)$ and $B_j = B(y', C_1\, r)$ with $C_1 = \min(\gamma, \gamma')$ give the conclusion. ∎

LEMMA 5.8. *Let* $\Sigma \subset \mathbb{R}^n$ *be an Ahlfors-regular set, let* $x \in \Sigma$*,* $r \leq 1$ *be fixed, and let* $B \subset B(x, r/2)$ *be such that* $|B| \geq \beta\, r^n$ *for some* $\beta > 0$*. Then there exist* $\gamma \leq 1/4$*, depending only on* n*,* β *and the Ahlfors-regularity constants for* Σ*, and* $y \in B$ *such that* $B(y, \gamma\, r) \cap \Sigma = \emptyset$*.*

Proof. We argue by contradiction and assume that for any $y \in B$ one can find $z_y \in B(y, \gamma\, r) \cap \Sigma$ with $\gamma \leq 1/4$ to be fixed small later. Then let us consider a maximal family $\mathcal{A}$ of points in B at mutual distance greater than or equal to $4\, \gamma\, r$. We have $B \subset \cup_{y \in \mathcal{A}} B(y, 4\, \gamma\, r)$, hence,

$$\beta\, r^n \leq |B| \leq \sum_{y \in \mathcal{A}} |B(y, 4\, \gamma\, r)| \leq C\, (\text{card } \mathcal{A})\, \gamma^n\, r^n,$$

and thus card $\mathcal{A} \geq C\, \gamma^{-n}$ for some constant $C = C(n, \beta) > 0$. Since the balls $B(y, 2\, \gamma\, r)$, $y \in \mathcal{A}$, are pairwise disjoint and $B(z_y, \gamma\, r) \subset B(y, 2\, \gamma\, r) \subset B(x, r)$, we get from the Ahlfors-regularity of Σ that

$$\begin{aligned} C\, r^{n-1} &\geq \mathcal{H}^{n-1}(\Sigma \cap B(x, r)) \\ &\geq \sum_{y \in \mathcal{A}} \mathcal{H}^{n-1}(\Sigma \cap B(y, 2\, \gamma\, r)) \\ &\geq \sum_{y \in \mathcal{A}} \mathcal{H}^{n-1}(\Sigma \cap B(z_y, \gamma\, r)) \\ &\geq C'\, (\text{card } \mathcal{A})\, \gamma^{n-1}\, r^{n-1} \geq C'\, \gamma^{-1}\, r^{n-1} \end{aligned}$$

which is impossible if γ is small enough, depending only on n, β, and the Ahlfors-regularity constants for Σ. $\square$

Let us now conclude. The Ahlfors-regularity of S is given by Corollary 5.6. Next it is not hard to see that Lemma 5.7 implies condition B stated in Definition 3.3. Indeed the fact that the balls B_i and B_j given by Lemma 5.7 are contained in two distinct connected components of $\mathbb{R}^n \setminus S$ follows from the fact that A_i and A_j are open and disjoint and, for instance, that $\partial A_i \subset S$. Note that we do not know in general whether the components of the partition are connected or not. Hence Lemma 5.7 is slightly stronger than condition B because it says not only that the two balls B_i and B_j are contained in two distinct connected components of $\mathbb{R}^n \setminus S$, but moreover in two distinct components of the partition.

The first part of Theorem 3.7 is given by Lemma 5.4 and it remains to prove that any component with finite Lebesgue measure is bounded. This follows from the Ahlfors-regularity and a covering argument. Let A_i be such that $|A_i| < +\infty$ and let $\mathcal{A}$ be a maximal family of points in ∂A_i at mutual distance greater than or equal to 1. The balls $B(x, 1/2)$, $x \in \mathcal{A}$, are pairwise disjoint and since $\partial A_i \subset S$ and S is Ahlfors-regular, we have

$$\begin{aligned}\operatorname{card} \mathcal{A} &\leq C \sum_{x\in\mathcal{A}} \mathcal{H}^{n-1}(S \cap B(x,1/2)) \\ &\leq C\, \mathcal{H}^{n-1}(S \cap \cup_{x\in\mathcal{A}} B(x,1/2)) \leq C\, \mathcal{H}^{n-1}(S).\end{aligned}$$

According to (2.2) we have $2\,\mathcal{H}^{n-1}(S) = P(A) < +\infty$ hence $\operatorname{card} \mathcal{A} < +\infty$. Since $\partial A_i \subset \cup_{x\in\mathcal{A}} B(x,1)$ and A_i has finite Lebesgue measure, it follows that $\operatorname{diam}(A_i) = \operatorname{diam}(\partial A_i) < +\infty$. Hence A_i is bounded as claimed.

6. Partitioning regular interface coefficients. We consider in this section a particular class of energies similar to that considered in [1] and prove refined properties for quasiminimal partitions. Namely, we show that in that case each topological boundary ∂A_i of some component of a reduced quasiminimal partition A is Ahlfors-regular and satisfies condition B (and even a slightly stronger condition, see Lemma 6.5) on its own. This is quite stronger than the general Theorem 5.1 which concerns only the set of interfaces.

First we need some definitions. Following [1, Chapter VI] we say that the coefficients σ_{ij}, $i,j \in \{0,\dots,N\}$, $i \neq j$, are partitioning regular if $\sigma_{ij} > 0$ and $\sigma_{ij} = \sigma_{ji}$ for all $i,j \in \{0,\dots,N\}$, $i \neq j$, and if, for all $i \in \{0,\dots,N\}$ and $b = (b_0,\dots,b_N) \in (\mathbb{R}^+)^{N+1}$ such that $b_l > 0$ for some $l \neq i$, one can find $j \in \{0,\dots,N\}$, $j \neq i$, such that

$$b_j\, \sigma_{ij} > \sum_{k\neq i,j} b_k\, (\sigma_{jk} - \sigma_{ik}).$$

Then we set

$$\sigma = \inf_i \left(\inf \left(\sup_{j,\, j\neq i} \left(b_j\, \sigma_{ij} - \sum_{k\neq i,j} b_k\, (\sigma_{jk} - \sigma_{ik}) \right) ;\, b \in (\mathbb{R}^+)^{N+1},\, b_i = 0,\, |b| = 1 \right) \right).$$

We have $\sigma > 0$. Next we say that a Borel map $F : \mathbb{R}^n \times S^{n-1} \to \mathbb{R}^+$ is admissible if F is even in its second variable, $F(x,\nu) = F(x,-\nu)$ for all $x \in \mathbb{R}^n$, $\nu \in S^{n-1}$, and if

$$F^- \leq F(x,\nu) \leq F^+$$

for all $x \in \mathbb{R}^n$, $\nu \in S^{n-1}$, and for some constants F^- and F^+ with $0 < F^- \leq F^+ < +\infty$. Then we say that E is an energy with partitioning regular interface coefficients if E is defined with respect to a family (F_{ij}) with $F_{ij} = \sigma_{ij}\, F$ for some partitioning regular coefficients σ_{ij} and some admissible function F. Heuristically the main point here is that the partitioning regularity of the coefficients σ_{ij} guarantees that it can always be advantageous in terms of minimizing the energy $E(A)$ to remove some component A_i and add it to some other component A_j for some judiciously chosen j (see, e.g., the argument in Lemma 6.2). Note that $E \in \mathcal{E}_0(\underline{\sigma} F^-, \overline{\sigma} F^+)$ with $\underline{\sigma} = \min_{i,j} \sigma_{ij}$ and $\overline{\sigma} = \max_{i,j} \sigma_{ij}$. Note also that for all $A \in \mathcal{P}$ and $U \subset \mathbb{R}^n$ measurable, we have

$$E(A, U) = \sum_{\substack{i,j \\ i \neq j}} \sigma_{ij} E_{ij}(A, U),$$

where

$$E_{ij}(A, U) = \int_{\partial^* A_i \cap \partial^* A_j} \mathbb{1}_U(x)\, F(x, \nu_{A_i}(x))\, d\mathcal{H}^{n-1}.$$

Since $\nu_{A_i}(x) = -\nu_{A_j}(x)$ for $\mathcal{H}^{n-1}$-a.e. $x \in \partial^* A_i \cap \partial^* A_j$ and all $i \neq j$, we also have $E_{ij}(A, U) = E_{ji}(A, U)$.

The main result reads now as follows.

THEOREM 6.1. *Let E be an energy with partitioning regular interface coefficients, let $g : [0, +\infty) \to [0, +\infty)$ be such that $\lim_{t \to 0} t^{(1-n)/n} g(t) = 0$, and let $\delta > 0$ be fixed. Let A be a reduced (g, δ)-quasiminimal partition with respect to E. Then, for all $i \in \{0, \dots, N\}$, we have $\partial^* A_i = \partial A_i$ and ∂A_i is an Ahlfors-regular set which satisfies condition B. Moreover, the Ahlfors-regularity and condition B constants can be chosen depending only on n, F^-, F^+, $\underline{\sigma}$, $\overline{\sigma}$, σ, g, δ, and N.*

Of course, assuming moreover that F satisfies (H3) and combining Theorem 6.1 with Theorem 3.6, one gets similar results for quasiminimal partitions with prescribed measure.

We fix for the rest of this section an energy E with partitioning regular interface coefficients, a function $g : [0, +\infty) \to [0, +\infty)$ such that $\lim_{t \to 0} t^{(1-n)/n} g(t) = 0$, some $\delta < 1$, and a reduced (g, δ)-quasiminimal partition A with respect to E. The main tool in the proof of Theorem 6.1 is the following analogue of Lemma 5.3 when the proportion of some component A_i inside a ball is assumed to be small. Then we prove that one half the ball is entirely contained in the interior of the complement of A_i.

LEMMA 6.2. *There exists $\eta_1 = \eta_1(n, F^-, F^+, \sigma, \overline{\sigma}, g, N) > 0$ such that for all $i \in \{0, \dots, N\}$, $x \in \mathbb{R}^n$, and $r \leq \delta/2$, if $|B(x, r) \cap A_i| \leq \eta_1 r^n$, then $B(x, r/2) \subset \mathbb{R}^n \setminus \overline{A}_i$.*

Proof. Since A is a reduced quasiminimal partition, we know from Lemma 5.4 that A_i is open hence it will be sufficient to prove that $|B(x, r/2) \cap A_i| = 0$. Thus let $i \in \{0, \dots, N\}$, $x \in \mathbb{R}^n$, and $r \leq \delta/2$ be such that $|B(x, r) \cap A_i| \leq \eta_1 r^n$ for some $\eta_1 > 0$ to be fixed small later and let us set $m(t) = |B(x, t) \cap A_i|$ for all $t > 0$. Arguing by contradiction we assume that $m(r/2) > 0$. For $t > 0$, we set $b_k = E_{ik}(A, B(x, t))$ when $k \neq i$ and $b_i = 0$, and, using the partitioning regularity of the coefficients, we choose $j \in \{0, \dots, N\}$, $j \neq i$, such that

$$b_j\, \sigma_{ij} - \sum_{k \neq i, j} b_k\, (\sigma_{jk} - \sigma_{ik}) \geq \sigma\, |b|.$$

Then we define $A' \in \mathcal{P}$ setting $A'_i = A_i \setminus B(x, t)$, $A'_j = A_j \cup (A_i \cap B(x, t))$, and $A'_k = A_k$ for $k \neq i, j$. The partitions A and A' coincide on the open set $\mathbb{R}^n \setminus \overline{B}(x, t)$

and hence have the same energy there. Inside $B(x,t)$ we have $\partial^* A'_i \cap B(x,t) = \emptyset$. On the other hand, we have $\partial^* A'_j \cap B(x,t) \subset (\partial^* A_j \cup \partial^* A_i) \cap B(x,t)$ with $\nu_{A'_j} = \nu_{A_j}$ $\mathcal{H}^{n-1}$-a.e. on $\partial^* A'_j \cap \partial^* A_j \cap B(x,t)$, $\nu_{A'_j} = \nu_{A_i}$ $\mathcal{H}^{n-1}$-a.e. on $\partial^* A'_j \cap \partial^* A_i \cap B(x,t)$. Obviously $\partial^* A'_k = \partial^* A_k$ for all $k \neq i,j$. It follows that for $k \neq i,j$,

$$\begin{aligned}
E_{jk}(A', B(x,t)) &= \int_{\partial^* A'_j \cap \partial^* A_k} \mathbb{1}_{B(x,t)}(y)\, F(y, \nu_{A'_j}(y))\, d\mathcal{H}^{n-1} \\
&\leq \int_{\partial^* A_j \cap \partial^* A_k} \mathbb{1}_{B(x,t)}(y)\, F(y, \nu_{A_j}(y))\, d\mathcal{H}^{n-1} \\
&\quad + \int_{\partial^* A_i \cap \partial^* A_k} \mathbb{1}_{B(x,t)}(y)\, F(y, \nu_{A_i}(y))\, d\mathcal{H}^{n-1} \\
&= E_{jk}(A, B(x,t)) + E_{ik}(A, B(x,t)),
\end{aligned}$$

and hence

$$\begin{aligned}
E(A', B(x,t)) \leq & \sum_{\substack{k,l \neq i,j \\ k \neq l}} \sigma_{kl}\, E_{kl}(A, B(x,t)) \\
& + 2 \sum_{k \neq i,j} \sigma_{jk}\, E_{jk}(A, B(x,t)) + 2 \sum_{k \neq i,j} \sigma_{jk}\, E_{ik}(A, B(x,t)) \\
\leq & \sum_{\substack{k,l \neq i,j \\ k \neq l}} \sigma_{kl}\, E_{kl}(A, B(x,t)) \\
& + 2 \sum_{k \neq i,j} \sigma_{jk}\, E_{jk}(A, B(x,t)) + 2 \sum_{k \neq i,j} \sigma_{ik}\, E_{ik}(A, B(x,t)) \\
& + 2\, \sigma_{ij}\, E_{ij}(A, B(x,t)) + 2 \sum_{k \neq i,j} b_k\, (\sigma_{jk} - \sigma_{ik}) - 2\, b_j\, \sigma_{ij} \\
\leq & E(A, B(x,t)) - 2\, \sigma\, |b|.
\end{aligned}$$

Next let us estimate $E(A', \partial B(x,t))$. We have $\partial^* A'_j \subset \partial^* A_j \cup \partial^* (A_i \cap B(x,t))$ with $\nu_{A'_j} = \nu_{A_j}$ $\mathcal{H}^{n-1}$-a.e. on $\partial^* A'_j \cap \partial^* A_j$, hence,

$$\begin{aligned}
\sum_{k \neq i,j} \sigma_{jk} E_{jk}(A', \partial B(x,t)) \leq & \sum_{k \neq i,j} \sigma_{jk} E_{jk}(A, \partial B(x,t)) \\
& + F^+\, \overline{\sigma} \sum_{k \neq i,j} \mathcal{H}^{n-1}(\partial^* A_k \cap \partial^* (A_i \cap B(x,t)) \cap \partial B(x,t)),
\end{aligned}$$

and

$$\begin{aligned}
\sigma_{ij} E_{ij}(A', \partial B(x,t)) \leq & F^+\, \overline{\sigma}\, (\mathcal{H}^{n-1}(\partial^* A_j \cap \partial^* (A_i \setminus B(x,t)) \cap \partial B(x,t)) \\
& + \mathcal{H}^{n-1}(\partial^* (A_i \cap B(x,t)) \cap \partial^* (A_i \setminus B(x,t)) \cap \partial B(x,t))).
\end{aligned}$$

Similarly we have

$$\sum_{k \neq i,j} \sigma_{ik} E_{ik}(A', \partial B(x,t)) \leq F^+\, \overline{\sigma} \sum_{k \neq i,j} \mathcal{H}^{n-1}(\partial^* A_k \cap \partial^* (A_i \setminus B(x,t)) \cap \partial B(x,t))).$$

It follows that

$$\begin{aligned} E(A', \partial B(x,t)) \leq & \sum_{\substack{k,l \neq i,j \\ k \neq l}} \sigma_{kl}\, E_{kl}(A, \partial B(x,t)) + 2 \sum_{k \neq i,j} \sigma_{jk}\, E_{jk}(A, \partial B(x,t)) \\ & + 2\, F^+\, \overline{\sigma} \sum_{k \neq i,j} \mathcal{H}^{n-1}(\partial^* A_k \cap \partial^*(A_i \cap B(x,t)) \cap \partial B(x,t)) \\ & + 2\, F^+\, \overline{\sigma} \sum_{k \neq i} \mathcal{H}^{n-1}(\partial^* A_k \cap \partial^*(A_i \setminus B(x,t)) \cap \partial B(x,t)) \\ & + 2\, F^+\, \overline{\sigma}\, \mathcal{H}^{n-1}(\partial^*(A_i \cap B(x,t)) \cap \partial^*(A_i \setminus B(x,t)) \cap \partial B(x,t)). \end{aligned}$$

Since $\mathcal{H}^{n-1}$-a.e. point belongs to at most two of the sets $\partial^* A_k$ (see what follows Definition 2.3), we get

$$\begin{aligned} E(A', \partial B(x,t)) \leq & \; E(A, \partial B(x,t)) \\ & + F^+\, \overline{\sigma}\, (4\, P(A_i \cap B(x,t), \partial B(x,t)) + 6\, P(A_i \setminus B(x,t), \partial B(x,t))) \end{aligned}$$

and thus

$$E(A', \partial B(x,t)) \leq E(A, \partial B(x,t)) + 10\, F^+\, \overline{\sigma}\, m'(t)$$

for a.e. $t > 0$. Then we get from the (g, δ)-quasiminimality of A that

$$2\, \sigma\, |b| \leq 10\, F^+\, \overline{\sigma}\, m'(t) + g(|A' \triangle A|).$$

On other hand, we have

$$\begin{aligned} |b| = \max_k |b_k| & \geq N^{-1} \sum_{k \neq i} E_{ik}(A, B(x,t)) \\ & \geq N^{-1}\, F^- \sum_{k \neq i} \mathcal{H}^{n-1}(\partial^* A_k \cap \partial^* A_i \cap B(x,t)) \\ & = N^{-1}\, F^-\, P(A_i, B(x,t)) \geq C\, m(t)^{\frac{n-1}{n}} \end{aligned}$$

for some constant $C = C(n, F^-, N) > 0$ provided η_1 is small enough. We also have $|A' \triangle A| = m(t) \leq \eta_1\, r^n \leq \eta_1$ hence, if η_1 is small enough,

$$g(|A' \triangle A|) \leq C\, \sigma\, m(t)^{\frac{n-1}{n}},$$

where C is the same constant as above. It follows that

$$m(t)^{\frac{n-1}{n}} \leq C\, m'(t)$$

for a.e. $t \in (r/2, r)$ and for some suitable constant $C = C(n, F^-, F^+, \sigma, \overline{\sigma}, N) > 0$. Then, dividing both sides by $m(t)^{(n-1)/n}$ which does not vanish for $t \geq r/2$ and integrating on $t \in (r/2, r)$, we get

$$r \leq C\, (m(r)^{\frac{1}{n}} - m(r/2)^{\frac{1}{n}}) \leq C\, m(r)^{\frac{1}{n}} \leq C\, \eta_1^{\frac{1}{n}}\, r.$$

This is impossible if η_1 is small enough and concludes the proof. $\square$

Let $\eta_2 = \min(\eta_0, \eta_1)$ where η_0 and η_1 are given by Lemmas 5.3 and 6.2, respectively. For all $i \in \{0, \ldots, N\}$, we set

$$\partial_i = \{x \in \mathbb{R}^n \, ; \, \min(|A_i \cap B(x,r)|, |B(x,r) \setminus A_i|) > \eta_2 \, r^n \text{ for all } r \leq \delta/2\}. \tag{6.1}$$

The first part of Theorem 6.1 is given by the following lemma.

LEMMA 6.3. *For all $i \in \{0, \ldots, N\}$, we have $\partial_i = \partial^* A_i = \partial A_i$.*

Proof. First since A_i and $\mathbb{R}^n \setminus \overline{A}_i$ are open (see Lemma 5.4) we have $\partial_i \cap (A_i \cup \mathbb{R}^n \setminus \overline{A}_i) = \emptyset$. Then it follows from Lemmas 5.3 and 6.2 that $\mathbb{R}^n$ is the disjoint union of the three sets A_i, $\mathbb{R}^n \setminus \overline{A}_i$, and ∂_i and we get in particular that $\partial_i = \overline{A}_i \setminus A_i = \partial A_i$. On the other hand, we have $A_i = A_i(1)$ by definition of a reduced partition hence, to prove that $\partial_i = \partial^* A_i$, we only need to check that $\mathbb{R}^n \setminus \overline{A}_i = A_i(0)$. One always has $\mathbb{R}^n \setminus \overline{A}_i \subset A_i(0)$. Conversely if $x \in A_i(0)$, then $\min(|A_i \cap B(x,r)|, |B(x,r) \setminus A_i|) = |B(x,r) \cap A_i| \leq \eta_2 \, r^n$ for all r small enough, hence $x \in \mathbb{R}^n \setminus \overline{A}_i$ according to Lemma 6.2. □

Next we prove the Ahlfors-regularity of ∂A_i. As a convention for the rest of this section, we say that a constant is universal if its value can be chosen depending only on n, F^-, F^+, $\underline{\sigma}$, $\overline{\sigma}$, σ, g, δ, and N.

LEMMA 6.4. *For all $i \in \{0, \ldots, N\}$, ∂A_i is an Ahlfors-regular set with Ahlfors-regularity constants that can be chosen universal.*

Proof. Let $x \in \partial A_i$ and $r \leq 1$ be fixed. Since $\partial A_i \subset S$ by Lemma 5.4 and S is Ahlfors-regular with universal regularity constants (see Corollary 5.6), we have

$$\mathcal{H}^{n-1}(\partial A_i \cap B(x,r)) \leq \mathcal{H}^{n-1}(S \cap B(x,r)) \leq C \, r^{n-1}$$

for some universal constant $C > 0$. On the other hand, the relative isoperimetric inequality for balls together with Lemma 6.3 and (6.1) gives

$$\eta_2^{\frac{n-1}{n}} \, r^{n-1} < \min(|A_i \cap B(x,r)|, |B(x,r) \setminus A_i|)^{\frac{n-1}{n}} \leq C \, \mathcal{H}^{n-1}(\partial A_i \cap B(x,r))$$

provided $r \leq \delta/2$. Finally the lower inequality in the Ahlfors-regularity for all radii $r \leq 1$ follows easily, with a slightly different constant (depending on δ). □

LEMMA 6.5. *There exists a universal constant $C_2 > 0$ such that for all $i \in \{0, \ldots, N\}$, $x \in \partial A_i$, and $r \leq 1$, one can find a ball B_i with radius $C_2 \, r$ such that $B_i \subset A_i \cap B(x,r)$.*

Proof. Let $i \in \{0, \ldots, N\}$, $x \in \partial A_i$, and $r \leq 1$ be fixed and set $B = A_i \cap B(x, r/2)$. It follows from Lemma 6.3 and (6.1) that $|B| \geq \beta \, r^n$ for some universal $\beta > 0$ and from Lemma 6.4 that ∂A_i is Ahlfors-regular with universal regularity constants. Hence one can apply Lemma 5.8 to find a universal constant $C_2 \leq 1/4$ and a point $y \in A_i \cap B(x, r/2)$ such that $B(y, C_2 \, r) \cap \partial A_i = \emptyset$. This implies automatically that $B(y, C_2 \, r) \subset A_i$ and since $B(y, C_2 \, r) \subset B(x,r)$ this gives the required conclusion. □

This implies easily that ∂A_i satisfies condition B. Indeed combining Lemma 6.5 with Lemma 5.7 we get the existence of a universal constant $C_3 > 0$ such that for all $i \in \{0, \ldots, N\}$, $x \in \partial A_i$, and $r \leq 1$, one can find two balls B and B' with radius $C_3 \, r$ and such that $B \subset A_i \cap B(x,r)$ and $B' \subset A_j \cap B(x,r)$ for some $j \neq i$. In particular we have $B' \subset \mathbb{R}^n \setminus \overline{A}_i$ hence B and B' are contained in two distinct connected components of $\mathbb{R}^n \setminus \partial A_i$. This concludes the proof of Theorem 6.1.

REFERENCES

[1] F. J. Almgren, *Existence and regularity almost everywhere of solutions to elliptic variational problems with constraints*, Mem. Amer. Math. Soc., 4 (1976), pp. viii + 199.

[2] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press, Oxford, UK, 2000.

[3] G. David, *Morceaux de graphes lipschitziens et intégrales singulières sur une surface*, Rev. Mat. Iberoamericana, 4 (1988), pp. 73–114.

[4] G. David and D. Jerison, *Lipschitz approximation to hypersurfaces, harmonic measure, and singular integrals*, Indiana Univ. Math. J., 39 (1990), pp. 831–845.

[5] G. David and S. Semmes, *Analysis of and on Uniformly Rectifiable Sets*, AMS, Providence, RI, 1993.

[6] G. David and S. Semmes, *Quantitative rectifiability and Lipschitz mappings*, Trans. Amer. Math. Soc., 337 (1993), pp. 855–889.

[7] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser Verlag, Basel, 1984.

[8] G. P. Leonardi, *Infiltrations in immiscible fluids systems*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 425–436.

[9] G. P. Leonardi and S. Rigot, *Isoperimetric sets on Carnot groups*, Houston J. Math., 29 (2003), pp. 609–637.

[10] S. Rigot, *Quasiminimal crystals with a volume constraint and uniform rectifiability*, J. Math. Pures Appl. (9), 82 (2003), pp. 1651–1695.

[11] B. White, *Existence of least-energy configurations of immiscible fluids*, J. Geom. Anal., 6 (1996), pp. 151–161.

[12] W. P. Ziemer, *Weakly Differentiable Functions. Sobolev Spaces and Functions of Bounded Variation*, Springer-Verlag, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1616–1648

LANDAU–DE GENNES MODEL OF LIQUID CRYSTALS WITH SMALL GINZBURG–LANDAU PARAMETER*

XING-BIN PAN†

Abstract. Liquid crystals may exhibit two types of behaviors, in comparison with two types of superconductors. We wish to explore type-I behavior of liquid crystals by using the Landau–de Gennes model with small Ginzburg–Landau parameter κ. In this paper, for small κ, we give an estimate on the critical wave number Q_{c_3}, at which a liquid crystal undergoes a phase transition from a nematic state to a smectic state. We show that, if κ is small, a smectic phase can nucleate only from a special class of nematic states, which depends on the geometry of the domain occupied by the liquid crystal. If the wave number is further lowered, the liquid crystal is in a uniform smectic state, which resembles the perfect superconducting state of type-I superconductors.

Key words. liquid crystal, phase transition, Landau–de Gennes model, critical wave number

AMS subject classifications. 82D30, 35J55, 35Q55

DOI. 10.1137/050631318

1. Introduction.

As predicted by P. G. de Gennes in [dG], smectic liquid crystals can exhibit either type-I or type-II behavior, in comparison with two types of superconductivity. Both behavior types have been reported; see [dGP, p. 512]. Type I superconductivity has been successfully described by the minimizers of the Ginzburg–Landau functional with small Ginzburg–Landau parameter [GL], and it is natural to expect that the minimizers of the Landau–de Gennes model of liquid crystals with small Ginzburg–Landau parameter exhibit type-I behavior. In this paper, continuing our study in [P1],[1] we shall examine this expectation.

According to de Gennes's theory [dG, dGP], the state of a liquid crystal can be described (at least for a temperature close to the transition point) by a complex-valued function Ψ called *order parameter*, a real vector field of unit length $\mathbf{n}$ called *director field*, and a real number q called *wave number*, which depends on the material and temperature. $\Psi = 0$ for a nematic phase, and $\Psi \neq 0$ for a smectic phase. $(\Psi, \mathbf{n})$ is a minimizer of the Landau–de Gennes energy (see [dGP, C, BCLP]):

$$\mathcal{L}[\Psi, \mathbf{n}] = \int_\Omega \big\{ c|\nabla_{q\mathbf{n}}\Psi|^2 + \mathcal{F}_A(|\Psi|) + \mathcal{F}_N(\mathbf{n}, \nabla\mathbf{n}) \big\} dx,$$

where Ω is the region occupied by the liquid crystal, c is a real constant, $\nabla_{q\mathbf{n}}\Psi = \nabla\Psi - iq\mathbf{n}\Psi$ with $i = \sqrt{-1}$, $\mathcal{F}_A(|\Psi|)$ and $\mathcal{F}_N(\mathbf{n}, \nabla\mathbf{n})$ denote the smectic energy density and the nematic Oseen–Frank energy density, respectively;

$$\mathcal{F}_A(|\Psi|) = r|\Psi|^2 + \frac{u}{2}|\Psi|^4,$$

*Received by the editors April 25, 2004; accepted for publication (in revised form) May 16, 2005; published electronically January 27, 2006. This work was partially supported by the National Natural Science Foundation of China, the Science Foundation of the Ministry of Education of China, and NUS grant R-146-000-033-112.

http://www.siam.org/journals/sima/37-5/63131.html

†Department of Mathematics, East China Normal University, Shanghai 200062, People's Republic of China (xbpan@math.ecnu.edu.cn).

[1]Our investigation in [P1] was motivated by Lin [L] and Aviles and Giga [AG] on the phase transitions of liquid crystals, and by Calderer [C] and Bauman et al. [BCLP] on the Landau–de Gennes model.

where $u > 0$, $r < 0$, and $r \sim \alpha(T - T_0)$ as temperature T approaches a critical temperature T_0; see [dGP, p. 508]. In the case without external electromagnetic fields, the Oseen–Frank energy density is given by

$$\begin{aligned}\mathcal{F}_N(\mathbf{n}, \nabla\mathbf{n}) = \mathcal{K}_1|\mathrm{div}\ \mathbf{n}|^2 + \mathcal{K}_2|\mathbf{n}\cdot \mathrm{curl}\ \mathbf{n} + \tau|^2 + \mathcal{K}_3|\mathbf{n}\wedge \mathrm{curl}\ \mathbf{n}|^2 \\ + (\mathcal{K}_2 + \mathcal{K}_4)[\mathrm{tr}(\nabla\mathbf{n})^2 - (\mathrm{div}\ \mathbf{n})^2],\end{aligned}$$

where $\mathcal{K}_j$, $j = 1, 2, 3, 4$, called elastic coefficients, are material constants, among them $\mathcal{K}_1$, $\mathcal{K}_2$, and $\mathcal{K}_3$ are positive; τ is a real number referring to the chiral pitch in some liquid crystal materials. As explained in [P1], we shall drop the last term in the Oseen–Frank energy density, and assume $\mathcal{K}_3 = \mathcal{K}_2$. After making the rescaling

$$\Psi = \sqrt{\frac{|r|}{u}}\psi, \quad \kappa = \sqrt{\frac{|r|}{c}}, \quad K_j = \frac{u}{c|r|}\mathcal{K}_j, \quad j = 1, 2,$$

and replacing $\mathcal{L}$ by $\frac{u}{c|r|}\mathcal{L}$, we are led to the following functional:[2]

$$\mathcal{G}[\psi, \mathbf{n}] = \int_\Omega \Big\{|\nabla_{q\mathbf{n}}\psi|^2 + \frac{\kappa^2}{2}(1 - |\psi|^2)^2 + K_1|\mathrm{div}\ \mathbf{n}|^2 + K_2|\mathrm{curl}\ \mathbf{n} + \tau\mathbf{n}|^2\Big\}dx. \tag{1.1}$$

As mentioned in [P1], we believe that the simplified energy functional (1.1) catches the main feature and most difficulties of the Landau–de Gennes model with the full Oseen–Frank energy. Note that in [dGP] the quantity $\xi = \sqrt{\frac{c}{|r|}}$ is called the *order parameter coherence length*. Hence

$$\kappa = \frac{1}{\xi}.$$

We prefer keeping the notation κ in our model instead of ξ, as κ corresponds to the Ginzburg–Landau parameter in the Ginzburg–Landau model of superconductivity (when taking the penetration length as unit of length). For convenience, let us call κ in (1.1) the *Ginzburg–Landau parameter* for liquid crystals.[3] From the analogy between liquid crystals and superconductivity, we expect that liquid crystals exhibit type-I behavior when the Ginzburg–Landau parameter κ is small.

As explained in [P1], the natural space for variational problems of (1.1) is

$$W^{1,2}(\Omega, \mathbb{C}) \times V(\Omega, \mathbb{S}^2),$$

where[4]

$$\begin{aligned}V(\Omega, \mathbb{R}^3) &= \big\{\mathbf{u} \in L^2(\Omega, \mathbb{R}^3) : \ \ \mathrm{div}\ \mathbf{u} \in L^2(\Omega), \ \ \mathrm{curl}\ \mathbf{u} \in L^2(\Omega, \mathbb{R}^3)\big\}, \\ V(\Omega, \mathbb{S}^2) &= \big\{\mathbf{n} \in V(\Omega, \mathbb{R}^3) : \ \ |\mathbf{n}(x)| = 1 \quad \text{a.e. in } \Omega\big\}.\end{aligned}$$

$V(\Omega, \mathbb{R}^3)$ is a Hilbert space with the inner product and norm defined by

$$\begin{aligned}(\mathbf{u}, \mathbf{v})_V &= \int_\Omega \{\mathrm{div}\ \mathbf{u}\ \mathrm{div}\ \mathbf{v} + \mathrm{curl}\ \mathbf{u}\cdot \mathrm{curl}\ \mathbf{v} + \mathbf{u}\cdot\mathbf{v}\}dx, \\ \|\mathbf{u}\|_V &= \Big\{\|\mathrm{div}\ \mathbf{u}\|^2_{L^2(\Omega)} + \|\mathrm{curl}\ \mathbf{u}\|^2_{L^2(\Omega)} + \|\mathbf{u}\|^2_{L^2(\Omega)}\Big\}^{1/2}.\end{aligned}$$

[2]In [P1] we used the functional $\mathcal{E}[\psi, \mathbf{n}] = \mathcal{G}[\psi, \mathbf{n}] - \frac{\kappa^2}{2}|\Omega|$, which is zero on the trivial critical points.

[3]In physical literature on liquid crystals, the letter κ has been used also for other ratios. For instance, Renn and Lubensky [RL] defined the twist Ginzburg parameter $\kappa_2 = \frac{\lambda_2}{\xi}$.

[4]The space $V(\Omega, \mathbb{R}^3)$ was denoted by $H(\mathrm{curl}, \mathrm{div}, \Omega)$ in Dautray and Lions [DL].

Here and throughout, unless specified otherwise, $\|\cdot\|_{L^2(\Omega)}$ denotes both the usual L^2 norm for scalar functions and the L^2 norm for vector fields. Throughout this paper we assume that

$$\Omega \text{ is a bounded, simply connected domain in } \mathbb{R}^3 \text{ with smooth boundary.} \tag{1.2}$$

We concern the global minimizers of $\mathcal{G}$ in $W^{1,2}(\Omega,\mathbb{C})\times V(\Omega,\mathbb{S}^2)$, without prescribing boundary data of the director fields. Let

$$C(K_1,K_2,\kappa,\tau,q)=\inf_{(\psi,\mathbf{n})\in W^{1,2}(\Omega,\mathbb{C})\times V(\Omega,\mathbb{S}^2)}\mathcal{G}[\psi,\mathbf{n}]. \tag{1.3}$$

Note that if $(\psi,\mathbf{n})$ is a minimizer of $\mathcal{G}$ with wave number q, then $(\psi,-\mathbf{n})$, $(-\psi,\mathbf{n})$, and $(\bar\psi,\mathbf{n})$ are minimizers of $\mathcal{G}$ with wave number $-q$. Therefore in the following we shall always assume $q\ge 0$. The functional $\mathcal{G}$ has many trivial minimizers $(0,\mathbf{n})$ with $\mathbf{n}\in\mathcal{C}(\tau)$, where

$$\mathcal{C}(\tau)=\{\mathbf{n}\in V(\Omega,\mathbb{S}^2):\ \text{div } \mathbf{n}=0,\ \text{curl } \mathbf{n}+\tau\mathbf{n}=\mathbf{0} \text{ in } \Omega\}. \tag{1.4}$$

The trivial minimizers correspond to the nematic state. The classification of vector fields in $\mathcal{C}(\tau)$ has been proved in [BCLP].

It is useful to find the regime of parameters that the global minimizers are nontrivial. In [P1], motivated by the analogy between liquid crystals and superconductivity (see [dG, dGP]), we introduced the *critical wave number* Q_{c_3}, which is a critical number for q such that if $|q|<Q_{c_3}$, then there exist nontrivial minimizers:

$$Q_{c_3}(K_1,K_2,\kappa,\tau)=\inf\{q>0:\ \mathcal{G} \text{ has only trivial minimizers}\}. \tag{1.5}$$

In [P1] we also gave estimates of Q_{c_3} and investigated the behavior of nontrivial minimizers for large K_1, K_2. As we are now concerned with the type-I behavior of liquid crystals, we shall estimate the value of Q_{c_3} when the Ginzburg–Landau parameter κ is small, with K_1, K_2 fixed; and we shall examine the behavior of the minimizers when the wave number q is less than Q_{c_3}.

To state the main results of this paper, let us define, for a vector field $\mathbf{u}\in V(\Omega,\mathbb{R}^3)$,[5]

$$\omega(\mathbf{u})\equiv\omega(\mathbf{u},\Omega)=\inf_{\phi\in W^{1,2}(\Omega)}\fint_\Omega|\nabla\phi-\mathbf{u}|^2dx, \tag{1.6}$$

where $\fint_\Omega\phi dx=\frac{1}{|\Omega|}\int_\Omega\phi dx$, and $|\Omega|$ is the volume of Ω. It is easy to show that $\omega(\mathbf{u})$ is achieved by the unique (real-valued) solution $\zeta_{\mathbf{u}}$ of the equation

$$\begin{cases}\Delta\zeta_{\mathbf{u}}=\text{div } \mathbf{u} & \text{in } \Omega,\\ \dfrac{\partial\zeta_{\mathbf{u}}}{\partial\nu}=\gamma_\nu\mathbf{u} & \text{on } \partial\Omega,\\ \displaystyle\int_\Omega\zeta_{\mathbf{u}}\,dx=0,\end{cases} \tag{1.7}$$

[5]In the case $\mathbf{u}=\mathbf{F_h}$, where $\mathbf{h}$ is a unit vector, and $\mathbf{F_h}$ is a vector field satisfying curl $\mathbf{F_h}=\mathbf{h}$ and div $\mathbf{F_h}=0$ in Ω, the number $\omega(\mathbf{F_h})$ was denoted by $\omega(\mathbf{h})$ in [P2], and was denoted by $\lambda(\mathbf{h})$ in [P1].

where $\gamma_\nu \mathbf{u}$ is the trace of $\mathbf{u}$ on $\partial\Omega$, and $\gamma_\nu \mathbf{u}$ is the restriction of $\mathbf{u}\cdot\nu$ to $\partial\Omega$ if $\mathbf{u}$ is smooth. Up to an additive constant, the minimizer is unique. We shall call $\zeta_{\mathbf{u}}$ the minimizer of $\omega(\mathbf{u})$. Let

$$
\begin{aligned}
&\omega_*(\tau) = \inf_{\mathbf{n}\in\mathcal{C}(\tau)} \omega(\mathbf{n}),\\
&\mathcal{C}_*(\tau) = \{\mathbf{n}\in\mathcal{C}(\tau):\ \omega(\mathbf{n}) = \omega_*(\tau)\}.
\end{aligned}
\tag{1.8}
$$

We can show that $0<\omega_*(\tau)<1$ (see (2.2) in section 2). Note that $\omega(-\mathbf{n}) = \omega(\mathbf{n})$. So if $\mathbf{n}\in\mathcal{C}_*(\tau)$, then $-\mathbf{n}\in\mathcal{C}_*(\tau)$.

THEOREM 1.1. *Assume condition* (1.2)*, and fix the positive constants K_1, K_2, and τ. For small $\kappa>0$ we have*

$$
Q_{c_3}(K_1,K_2,\kappa,\tau) = \frac{\kappa}{\sqrt{\omega_*(\tau)}} + o(\kappa). \tag{1.9}
$$

Let $q = a_\kappa\kappa$ with $0<a_\kappa\kappa<Q_{c_3}(K_1,K_2,\kappa,\tau)$ and $\lim_{\kappa\to 0} a_\kappa = a_0$, where $0<a_0\le \frac{1}{\sqrt{\omega_*(\tau)}}$. Then

$$
C(K_1,K_2,\kappa,\tau,a_\kappa\kappa) = a_0^2\omega_*(\tau)\Big[1-\frac{1}{2}a_0^2\omega_*(\tau)\Big]|\Omega|\kappa^2 + o(\kappa^2); \tag{1.10}
$$

and for any sequence $\kappa\to 0$, there exist a subsequence $\kappa_j\to 0$ and $\mathbf{n}_0\in\mathcal{C}_*(\tau)$, such that the minimizers $(\psi_{\kappa_j},\mathbf{n}_{\kappa_j})$ of the functional $\mathcal{G}$ have expansions

$$
\begin{aligned}
\psi_{\kappa_j} &= c_{\kappa_j}\big[1 + i a_{\kappa_j}\kappa_j(\zeta_{\mathbf{n}_0} + \phi_{\kappa_j})\big],\\
\mathbf{n}_{\kappa_j} &= \mathbf{n}_0 + \kappa_j \mathbf{v}_{\kappa_j},
\end{aligned}
\tag{1.11}
$$

where $\zeta_{\mathbf{n}_0}$ is the minimizer of $\omega(\mathbf{n}_0)$, c_{κ_j} is a complex number, $\phi_{\kappa_j}\in W^{1,2}(\Omega,\mathbb{C})$, $\mathbf{v}_{\kappa_j}\in V(\Omega,\mathbb{R}^3)$, and as $\kappa_j\to 0$,

$$
\begin{aligned}
&|c_{\kappa_j}| \to [1-a_0^2\omega_*(\tau)]^{1/2},\\
&\|\phi_{\kappa_j}\|_{W^{1,2}(\Omega,\mathbb{C})}\to 0,\\
&\|\mathrm{div}\mathbf{v}_{\kappa_j}\|_{L^2(\Omega)}\to 0,\\
&\|\mathrm{curl}\mathbf{v}_{\kappa_j} + \tau\mathbf{v}_{\kappa_j}\|_{L^2(\Omega)}\to 0.
\end{aligned}
\tag{1.12}
$$

Remark. (1) When $0<a_0<\frac{1}{\sqrt{\omega_*(\tau)}}$, the expansion (1.11) shows that the minimizers have the approximation

$$
|\psi_{\kappa_j}| \sim [1-a_0^2\omega_*(\tau)]^{1/2}, \qquad \mathbf{n}_{\kappa_j}\sim\mathbf{n}_0 \quad \text{as } \kappa_j\to 0.
$$

So we may say that the liquid crystal is in a *uniform smectic state* when κ is small, which resembles the perfect superconducting state of type-I superconductors under an applied magnetic field below the critical field H_c.

(2) From (1.11) we also see that, when the wave number q decreases from Q_{c_3}, a smectic phase nucleates from a nematic state $(0,\mathbf{n}_0)$, where $\mathbf{n}_0\in\mathcal{C}_*(\tau)$. Let

$$
\begin{aligned}
\mathcal{N}(\tau) &= \{(0,\mathbf{n}):\ \mathbf{n}\in\mathcal{C}(\tau)\},\\
\mathcal{N}_*(\tau) &= \{(0,\mathbf{n}):\ \mathbf{n}\in\mathcal{C}_*(\tau)\}.
\end{aligned}
$$

Every element in $\mathcal{N}(\tau)$ represents a nematic state; however, only the nematic states in $\mathcal{N}_*(\tau)$ allow a smectic state to nucleate. Note that the structure of $\mathcal{N}_*(\tau)$ reflects the geometry of the domain Ω.

Theorem 1.1 indicates some similarity between smectic liquid crystals with small Ginzburg–Landau parameter κ and type-I superconductors. In fact, the estimate (1.9) for Q_{c_3} is similar to the estimate of the critical field H_c for type I superconductors obtained in [P2, Theorem 1]; and the expansion (1.11) resembles the expansions of minimizers for type-I superconductivity; see [P2, Theorem 2]. However, there is an important difference between liquid crystals and superconductors. For liquid crystals we require the director field to satisfy the pointwise unit length constraint

$$|\mathbf{n}(x)| = 1 \quad \text{on } \Omega. \tag{1.13}$$

The constraint (1.13) is one of the sources of complications, especially for analysis of high order terms in (1.11). Another source of complications is that the functional $\mathcal{I}$ is not coercive in the orthogonal complement of the kernel, where $\mathcal{I}$ is defined by the last two terms in $\mathcal{G}$:

$$\mathcal{I}[\mathbf{v}] = \int_\Omega \{K_1|\text{div } \mathbf{v}|^2 + K_2|\text{curl } \mathbf{v} + \tau\mathbf{v}|\}dx. \tag{1.14}$$

The kernel $X(\tau)$ of $\mathcal{I}$ consists of all solutions of the following equations in $V(\Omega, \mathbb{R}^3)$:

$$\text{div } \mathbf{u} = 0, \qquad \text{curl } \mathbf{u} + \tau\mathbf{u} = \mathbf{0} \quad \text{ in } \Omega. \tag{1.15}$$

Obviously $X(\tau)$ is a closed subspace of $V(\Omega, \mathbb{R}^3)$, and $\mathcal{C}(\tau) = X(\tau) \cap V(\Omega, \mathbb{S}^2)$. With respect to the inner product $(\cdot, \cdot)_V$, the space $V(\Omega, \mathbb{R}^3)$ has orthogonal decomposition

$$V(\Omega, \mathbb{R}^3) = X(\tau) \oplus X_V^\perp(\tau).$$

Unfortunately, $\mathcal{I}$ is not coercive on $X_V^\perp(\tau)$, and a sequence satisfying $\mathcal{I}[\mathbf{u}_j] \to 0$ may exhibit boundary concentrations. Fortunately we are able to classify vector fields that satisfy (1.15) coupled with a pointwise constraint (Lemma 3.1), and obtain an orthogonality property (Proposition 3.3). These results will be useful in sections 4 and 5.

The outline of this paper is as follows. Section 2 contains some preliminary results that will be needed in later sections. In section 3 we classify the solutions of (1.15) coupled with a pointwise constraint. In section 4 we prove the estimate of Q_{c_3} for small κ, and provide an energy upper bound that is better than (1.10). In section 5 we investigate the behavior of nontrivial minimizers as $\kappa \to 0$.

We would like to mention that the mathematical theory of phase transitions of liquid crystals has been studied extensively by many mathematicians, among others we mention Hardt, Kinderlehrer, and Lin [HKL], Hardt and Kinderlehrer [HK], Brezis [B], Lin [L], and Aviles and Giga [AG]. Following the work by P. G. de Gennes [dG], the investigation on liquid crystals based on the Landau–de Gennes theory has been successfully conducted by many physicists, and the similarity between superconductivity and liquid crystals has been explored; see, for instance, Chen and Lubensky [ChL], Lubensky and Renn [LR], and Renn and Lubensky [RL]. The mathematical theory of the Landau–de Gennes model has been investigated by Calderer [C] and Bauman et al. [BCLP].

2. Preliminaries.

Given a vector field $\mathbf{u} \in V(\Omega, \mathbb{R}^3)$, let $\omega(\mathbf{u})$ and $\omega_*(\tau)$ be the numbers defined in (1.6) and (1.8), respectively, and let $\zeta_{\mathbf{u}}$ be the minimizer of $\omega(\mathbf{u})$.

LEMMA 2.1. *For any $\tau > 0$ and $\mathbf{n} \in \mathcal{C}(\tau)$ we have*

$$0 < \omega(\mathbf{n}) < 1. \tag{2.1}$$

In particular,

$$0 < \omega_*(\tau) < 1. \tag{2.2}$$

Proof. Using (1.7) we have, for any smooth function ϕ,

$$\int_\Omega (\nabla\zeta_{\mathbf{u}} - \mathbf{u}) \cdot \nabla\phi dx = 0. \tag{2.3}$$

Note that $\mathbf{u}$ has canonical decomposition

$$\mathbf{u} = \mathbf{v}_{\mathbf{u}} + \nabla\zeta_{\mathbf{u}}, \tag{2.4}$$

where $\mathbf{v}_{\mathbf{u}}$ satisfies

$$\begin{cases} \operatorname{div} \mathbf{v}_{\mathbf{u}} = 0, \quad \operatorname{curl} \mathbf{v}_{\mathbf{u}} = \operatorname{curl} \mathbf{u} & \text{in } \Omega, \\ \gamma_\nu \mathbf{v}_{\mathbf{u}} = 0 & \text{on } \partial\Omega. \end{cases}$$

Using (2.3) we find

$$\int_\Omega |\mathbf{u}|^2 dx = \int_\Omega |\mathbf{v}_{\mathbf{u}}|^2 dx + \int_\Omega |\nabla\zeta_{\mathbf{u}}|^2 dx.$$

Hence

$$\omega(\mathbf{u}) = -\!\!\!\!\!\!\int_\Omega |\mathbf{v}_{\mathbf{u}}|^2 dx = -\!\!\!\!\!\!\int_\Omega (|\mathbf{u}|^2 - |\nabla\zeta_{\mathbf{u}}|^2) dx.$$

In particular, for a unit vector field $\mathbf{n} \in V(\Omega, \mathbb{S}^2)$ we have $0 \le \omega(\mathbf{n}) \le 1$. If $\mathbf{n}$ has canonical decomposition $\mathbf{n} = \mathbf{v}_{\mathbf{n}} + \nabla\zeta_{\mathbf{n}}$, then

$$-\!\!\!\!\!\!\int_\Omega |\nabla\zeta_{\mathbf{n}}|^2 dx = -\!\!\!\!\!\!\int_\Omega (|\mathbf{n}|^2 - |\mathbf{v}_{\mathbf{n}}|^2) dx = 1 - \omega(\mathbf{n}). \tag{2.5}$$

Moreover, $\omega(\mathbf{n}) = 0$ if and only if $\mathbf{n} = \nabla\zeta_{\mathbf{n}}$; and $\omega(\mathbf{n}) = 1$ if and only if $\zeta_{\mathbf{n}} = 0$.

Now we assume $\mathbf{n} \in \mathcal{C}(\tau)$. Since $\operatorname{curl} \mathbf{n} = -\tau\mathbf{n} \ne \mathbf{0}$, we see that $\mathbf{n} \not\equiv \nabla\zeta_{\mathbf{n}}$. So $\omega(\mathbf{n}) > 0$. It has been proved in [BCLP] that every $\mathbf{n} \in \mathcal{C}(\tau)$ has the form

$$\begin{gathered} \mathbf{n} = Q\mathbf{N}_\tau(Q^t x), \\ \text{where} \quad \mathbf{N}_\tau = (\cos(\tau x_3), \sin(\tau x_3), 0)^\top, \quad Q \in SO(3). \end{gathered} \tag{2.6}$$

Here and thereafter, for a row vector $\mathbf{w}$, $\mathbf{w}^\top$ denotes the corresponding column vector. Note that $\zeta_{\mathbf{n}}$ depends on Q through $\mathbf{n}$. Let $\zeta_{\tau,Q}$ be the solution of

$$\begin{cases} \Delta\zeta_{\tau,Q} = 0 & \text{in } Q^t\Omega, \\ \dfrac{\partial\zeta_{\tau,Q}}{\partial\nu} = \mathbf{N}_\tau \cdot \nu & \text{on } \partial(Q^t\Omega), \\ \displaystyle\int_{Q^t\Omega} \zeta_{\tau,Q}\, dx = 0. \end{cases}$$

Then

$$\zeta_{\mathbf{n}}(x) = \zeta_{\tau,Q}(Q^t x).$$

From (2.6) we see that, for any bounded domain Ω with smooth boundary,

$$\mathbf{N}_\tau \cdot \nu \not\equiv 0 \quad \text{on } \partial(Q^t\Omega).$$

Hence $\zeta_{\tau,Q} \not\equiv 0$. Thus, $\omega(\mathbf{n}) < 1$. So (2.1) is true. Note that $\omega_*(\tau)$ is achieved by some $\mathbf{n}_0$ in $\mathcal{C}_*(\tau)$. So (2.2) is true. □

From the above discussion we see that

$$\omega(\mathbf{n}, \Omega) = \fint_\Omega |\nabla\zeta_{\mathbf{n}} - \mathbf{n}|^2 dx = \fint_{Q^t\Omega} |\nabla\zeta_{\tau,Q} - \mathbf{N}_\tau|^2 dx = \omega(\mathbf{N}_\tau, Q^t\Omega).$$

Therefore

$$\begin{aligned}\omega_*(\tau) &= \inf_{Q\in SO(3)} \fint_{Q^t\Omega} |\nabla\zeta_{\tau,Q} - \mathbf{N}_\tau|^2 dx \\ &= \inf_{Q\in SO(3)} \inf_{\phi\in W^{1,2}(Q^t\Omega)} \fint_{Q^t\Omega} |\nabla\phi - \mathbf{N}_\tau|^2 dx.\end{aligned}$$

Next we consider an eigenvalue problem. Given a vector field $\mathbf{u} \in V(\Omega, \mathbb{R}^3)$, let $\mu(\varepsilon\mathbf{u})$ be the lowest eigenvalue of the equation

$$\begin{cases} -\nabla^2_{\varepsilon\mathbf{u}}\phi = \mu\phi & \text{in } \Omega, \\ (\nabla_{\varepsilon\mathbf{u}}\phi)\cdot\nu = 0 & \text{on } \partial\Omega. \end{cases} \tag{2.7}$$

Lemma 2.2. *For any $\mathbf{n} \in V(\Omega, \mathbb{S}^2)$ we have, as $\varepsilon \to 0$,*

$$\varepsilon^2\omega(\mathbf{n}) + O(\varepsilon^3) \le \mu(\varepsilon\mathbf{n}) \le \omega(\mathbf{n})\varepsilon^2 + \frac{1-\omega(\mathbf{n})}{|\Omega|}\|\zeta_{\mathbf{n}}\|^2_{L^2(\Omega)}\varepsilon^4 + O(\varepsilon^6). \tag{2.8}$$

Proof. The lower bound can be proved as in [P2, Lemma 2.1]. To prove the upper bound, let $\phi = 1 + i\varepsilon\zeta_{\mathbf{n}}$. Then

$$\int_\Omega |\nabla_{\varepsilon\mathbf{n}}\phi|^2 dx = \varepsilon^2 \int_\Omega \{|\nabla\zeta_{\mathbf{n}} - \mathbf{n}|^2 + \varepsilon^2|\zeta_{\mathbf{n}}\mathbf{n}|^2\} dx = \varepsilon^2|\Omega|\omega(\mathbf{n}) + \varepsilon^4\|\zeta_{\mathbf{n}}\|^2_{L^2(\Omega)}.$$

Here we have used the fact $|\mathbf{n}(x)| \equiv 1$. Then we use the inequality

$$\mu(\varepsilon\mathbf{n}) \le \frac{\int_\Omega |\nabla_{\varepsilon\mathbf{n}}\phi|^2 dx}{\int_\Omega |\phi|^2 dx}$$

to get the upper bound in (2.8). □

It has been proved in [P1] (see the proof of Lemma 3.2 there) that, for a simply connected domain, a sequence bounded in $V(\Omega, \mathbb{R}^3)$ is bounded in $W^{1,2}_{\mathrm{loc}}(\Omega, \mathbb{R}^3)$. The proof given in [P1] is based on the canonical decomposition of vector fields. This conclusion is true for general bounded domains, and we include it here for our convenience.

Lemma 2.3. *Let Ω be a bounded and smooth domain in $\mathbb{R}^3$. For any subdomain $D_1 \Subset D_2 \subseteq \Omega$, there exists a constant $C(D_1, D_2)$ such that for any $\mathbf{u} \in V(\Omega, \mathbb{R}^3)$ we have*

$$\|\mathbf{u}\|_{W^{1,2}(D_1,\mathbb{R}^3)} \le C(D_1, D_2)\left\{\|\mathrm{div}\mathbf{u}\|_{L^2(D_2)} + \|\mathrm{curl}\mathbf{u}\|_{L^2(D_2)} + \|\mathbf{u}\|_{L^2(D_2)}\right\}.$$

In particular, every bounded sequence in $V(\Omega, \mathbb{R}^3)$ is bounded in $W^{1,2}_{loc}(\Omega, \mathbb{R}^3)$.

The proof of Lemma 2.3 will be given in Appendix A.

3. Solutions of the equation curl $\mathbf{u}$ + $\quad\mathbf{u} = \mathbf{0}$ under a pointwise constraint. Assume condition (1.2) and let $\tau > 0$. For a given $\mathbf{n}_0 \in \mathcal{C}(\tau)$, we shall classify all the solutions in $V(\Omega, \mathbb{R}^3)$ to (1.15) coupled with a pointwise orthogonality constraint $\mathbf{n}_0 \cdot \mathbf{u} = 0$, that is,

$$\mathbf{u} \in V(\Omega, \mathbb{R}^3), \quad \text{div}\, \mathbf{u} = 0, \quad \text{curl}\, \mathbf{u} + \tau \mathbf{u} = \mathbf{0}, \quad \mathbf{n}_0 \cdot \mathbf{u} = 0 \quad \text{in } \Omega. \tag{3.1}$$

LEMMA 3.1. *Let $\mathbf{n}_0 = Q\mathbf{N}_\tau(Q^t x)$ with $Q \in SO(3)$. The solutions of* (3.1) *are given by*

$$\mathbf{u}(x) = a\mathbf{U}_1 + b\mathbf{U}_2 + c\mathbf{U}_3,$$

where a, b, c are arbitrary real numbers,

$$\mathbf{U}_j(x) = Q\mathbf{V}_j(Q^t x), \tag{3.2}$$

and

$$\begin{aligned} \mathbf{V}_1 &= \big(\tau y_1 \sin(\tau y_3), -\tau y_1 \cos(\tau y_3), \cos(\tau y_3)\big)^\top, \\ \mathbf{V}_2 &= \big(\tau y_2 \sin(\tau y_3), -\tau y_2 \cos(\tau y_3), \sin(\tau y_3)\big)^\top, \\ \mathbf{V}_3 &= \big(\tau \sin(\tau y_3), -\tau \cos(\tau y_3), 0\big)^\top. \end{aligned} \tag{3.3}$$

Proof. We prove only the conclusion in the case where $\tau = 1$ and

$$\mathbf{n}_0 = \mathbf{N}_1 = (\cos x_3, \sin x_3, 0)^\top.$$

The conclusion for the general case is obtained by rotation and rescaling. From (3.1) we see that $\mathbf{u}$ is a solution of

$$\Delta \mathbf{u} + \mathbf{u} = \mathbf{0}, \tag{3.4}$$

and hence $\mathbf{u}$ is smooth in the interior of Ω. Let us take a small cube located in the interior of Ω and restrict ourselves to the cube. From the pointwise orthogonality constraint $\mathbf{u} \cdot \mathbf{N}_1 = 0$ we have

$$u_1 \cos x_3 + u_2 \sin x_3 = 0. \tag{3.5}$$

Hence $u_2 = -u_1 \cot x_3$ and

$$\partial_{x_3} u_2 = -\cot x_3 \partial_{x_3} u_1 + \frac{u_1}{\sin^2 x_3}. \tag{3.6}$$

Applying the Laplacian operator to (3.5) and using (3.4) we find

$$\begin{aligned} 0 &= \Delta(u_1 \cos x_3 + u_2 \sin x_3) \\ &= \cos x_3 \Delta u_1 + \sin x_3 \Delta u_2 - 2 \sin x_3 \partial_{x_3} u_1 + 2 \cos x_3 \partial_{x_3} u_2 - u_1 \cos x_3 - u_2 \sin x_3 \\ &= -2 \sin x_3 \partial_{x_3} u_1 + 2 \cos x_3 \partial_{x_3} u_2. \end{aligned}$$

So

$$\partial_{x_3} u_2 = \tan x_3 \partial_{x_3} u_1.$$

Combining this with (3.6) we find

$$\frac{\partial_{x_3} u_1}{u_1} = \cot x_3.$$

So there exists a function $f(x_1, x_2)$ such that

$$u_1 = f(x_1, x_2) \sin x_3, \qquad u_2 = -f(x_1, x_2) \cos x_3.$$

Now we use the condition div $\mathbf{u} = 0$ to find

$$\partial_{x_1} f(x_1, x_2) \sin x_3 - \partial_{x_2} f(x_1, x_2) \cos x_3 + \partial_{x_3} u_3 = 0.$$

Hence we can write

$$u_3 = \partial_{x_1} f(x_1, x_2) \cos x_3 + \partial_{x_2} f(x_1, x_2) \sin x_3 + g(x_1, x_2).$$

In order to determine f and g, we check each component of the equation curl $\mathbf{u}+\mathbf{u} = \mathbf{0}$ to find

$$\begin{aligned}
&f_{x_1x_2}(x_1, x_2) \cos x_3 + f_{x_2x_2}(x_1, x_2) \sin x_3 + g_{x_2}(x_1, x_2) = 0,\\
&f_{x_1x_1}(x_1, x_2) \cos x_3 + f_{x_1x_2}(x_1, x_2) \sin x_3 + g_{x_1}(x_1, x_2) = 0,\\
&g(x_1, x_2) = 0.
\end{aligned}$$

So ∇f is a constant vector, and $f(x_1, x_2) = ax_1 + bx_2 + c$. Hence on the cube we have

$$\begin{aligned}
u_1 &= (ax_1 + bx_2 + c) \sin x_3,\\
u_2 &= -(ax_1 + bx_2 + c) \cos x_3,\\
u_3 &= a \cos x_3 + b \sin x_3.
\end{aligned} \tag{3.7}$$

Since Ω is connected, (3.7) must be true everywhere on Ω.

On the other hand, computation shows that for any constants a, b, c, the vector field $\mathbf{u}$ defined by (3.7) satisfies the conditions

$$\text{div } \mathbf{u} = 0, \qquad \text{curl } \mathbf{u} + \mathbf{u} = \mathbf{0}, \qquad \mathbf{u} \cdot \mathbf{N}_1 = 0. \qquad \square$$

Notation. For $\mathbf{n}_0 \in \mathcal{C}(\tau)$ and $\tau > 0$, $\mathcal{U}(\tau, \mathbf{n}_0)$ denotes the set of all solutions of (3.1).

It follows from Lemma 3.1 that $\mathcal{U}(\tau, \mathbf{n}_0)$ is a linear space of dimension 3. Note that $\mathcal{U}(\tau, \mathbf{n}_0) = \mathcal{U}(\tau, -\mathbf{n}_0)$.

In addition to (3.7), if we further request that $\mathbf{u}$ satisfies the unit length condition $|\mathbf{u}(x)| = 1$ on Ω, then

$$\mathbf{u} = \pm(\sin x_3, -\cos x_3, 0)^\top.$$

Thus we get the following corollary.

COROLLARY 3.2. *Given $\mathbf{n}_0 \in \mathcal{C}(\tau)$, the set $\mathcal{U}(\tau, \mathbf{n}_0) \cap V(\Omega, \mathbb{S}^2)$ contains exactly two elements.*

The following result is useful in sections 4 and 5.

PROPOSITION 3.3. *Let $\mathbf{n}_0 \in \mathcal{C}_*(\tau)$ and $\mathbf{u}_0 \in \mathcal{U}(\tau, \mathbf{n}_0)$. We have*

$$\int_\Omega \mathbf{u}_0 \cdot \nabla \zeta_{\mathbf{n}_0} dx = 0, \tag{3.8}$$

where $\zeta_{\mathbf{n}_0}$ *is the solution of* (1.7) *associated with* $\mathbf{n}_0$.

Proof. We give the proof for $\tau = 1$.

Step 1. Let us introduce some notation. For a 3×3 nonsingular matrix $Q = (q_{ij})$ we assign a vector $\mathbf{q} = (q_{11}, q_{12}, \dots, q_{33})^\top \in \mathbb{R}^9$. Let

$$\mathcal{A} = \big\{\mathbf{q} = (q_{11}, q_{12}, \dots, q_{33})^\top \in \mathbb{R}^9 :\ \det(q_{ij}) > 0\big\},$$
$$\mathcal{Q} = \big\{\mathbf{q} = (q_{11}, q_{12}, \dots, q_{33})^\top \in \mathbb{R}^9 :\ (q_{ij}) \in SO(3)\big\}.$$

Then $\mathcal{A}$ is an open set in $\mathbb{R}^9$. Let

$$\mathbf{n}(x,\mathbf{q}) = Q\mathbf{N}_1(Q^t x), \qquad \zeta(x,\mathbf{q}) = \zeta_{\mathbf{n}(x,\mathbf{q})}.$$

Thus

$$n(x,\mathbf{q})_k = q_{k1}\cos y_3 + q_{k2}\sin y_3, \text{where } y_j = \sum_{i=1}^{3} q_{ij}x_i, \quad j = 1,2,3;$$

and $\zeta(x,\mathbf{q})$ is the unique solution of the equation

$$\begin{cases} \Delta\zeta = 0 & \text{in } \Omega,\\ \dfrac{\partial\zeta}{\partial\nu} = Q\mathbf{N}_1(Q^t x)\cdot\nu & \text{on } \partial\Omega,\\ \displaystyle\int_\Omega \zeta\, dx = 0. \end{cases}$$

If Ω is a bounded domain with $C^{3+\alpha}$ boundary, then the solution of the equation is $C^{3+\alpha}$ in x; see [GT, Theorem 6.3] (also see [ADN, LM]). We can show that ζ is C^2 in the parameter q_{ij} for $\mathbf{q} = (q_{11}, \dots, q_{33})^\top \in \mathcal{A}$, and $\frac{\partial^2\zeta}{\partial q_{ij}\partial q_{lm}}$ is C^2 in x. Of course if $q \in \mathcal{Q}$, then $\mathbf{n} \equiv \mathbf{n}(x,\mathbf{q}) \in \mathcal{C}(\tau)$ and $\zeta(x,\mathbf{q}) = \zeta_{\mathbf{n}}$. Let us define a function on $\mathcal{A}$ by

$$F(\mathbf{q}) = \frac{1}{2}\int_\Omega |\nabla\zeta(x,\mathbf{q}) - \mathbf{n}(x,\mathbf{q})|^2 dx.$$

Then $F(\mathbf{q})$ is a C^2 function in $\mathcal{A}$.

Step 2. Assume $\mathbf{n}_0 = (n_1^0, n_2^0, n_3^0)^\top \in \mathcal{C}_*(1)$, and $\mathbf{n}_0 = Q_0\mathbf{N}_1(Q_0^t x)$, where $Q_0 = (q_{ij}^0)$. Then $\mathbf{q}_0 = (q_{11}^0, \dots, q_{33}^0)$ is a minimizer of the following problem:

$$\text{Minimize } F(\mathbf{q}) \text{ subjected to the conditions } g_k(\mathbf{q}) = 0,\ k = 1, \dots, 6,$$

where

$$g_1(\mathbf{q}) = \sum_{j=1}^{3} q_{j1}^2 - 1, \qquad g_2(\mathbf{q}) = \sum_{j=1}^{3} q_{j2}^2 - 1,$$
$$g_3(\mathbf{q}) = \sum_{j=1}^{3} q_{j3}^2 - 1, \qquad g_4(\mathbf{q}) = \sum_{j=1}^{3} q_{j1}q_{j2},$$
$$g_5(\mathbf{q}) = \sum_{j=1}^{3} q_{j1}q_{j3}, \qquad g_6(\mathbf{q}) = \sum_{j=1}^{3} q_{j2}q_{j3}.$$

Thus $\mathbf{q}_0$ is a solution of the Euler–Lagrange equations

$$\frac{\partial F}{\partial q_{ij}}(\mathbf{q}_0)=\sum_{k=1}^{6}\lambda_k\frac{\partial g_k}{\partial q_{ij}}(\mathbf{q}_0),\qquad i,j=1,2,3,$$

where λ_k's are constants that may depend on $\mathbf{q}_0$. We compute

$$\frac{\partial F}{\partial q_{ij}}(\mathbf{q}_0)=\int_\Omega(\nabla\zeta_{\mathbf{n}_0}-\mathbf{n}_0)\cdot\Big(\nabla\frac{\partial\zeta}{\partial q_{ij}}(x,\mathbf{q}_0)-\frac{\partial\mathbf{n}}{\partial q_{ij}}(x,\mathbf{q}_0)\Big)dx.$$

Applying (2.3) for

$$\mathbf{u}=\mathbf{n}_0,\qquad \phi=\frac{\partial\zeta}{\partial q_{ij}}(x,\mathbf{q}_0),$$

we find

$$\frac{\partial F}{\partial q_{ij}}(\mathbf{q}_0)=-\int_\Omega(\nabla\zeta_{\mathbf{n}_0}-\mathbf{n}_0)\cdot\frac{\partial\mathbf{n}}{\partial q_{ij}}(x,\mathbf{q}_0)dx.$$

We compute

$$\begin{aligned}&\frac{\partial n_k}{\partial q_{i1}}(x,\mathbf{q}_0)=\delta_{ik}\cos y_3,\\&\frac{\partial n_k}{\partial q_{i2}}(x,\mathbf{q}_0)=\delta_{ik}\sin y_3,\\&\frac{\partial n_k}{\partial q_{i3}}(x,\mathbf{q}_0)=-x_i(q^0_{k1}\sin y_3-q^0_{k2}\cos y_3),\end{aligned}$$

where $y_3=q^0_{13}x_1+q^0_{23}x_2+q^0_{33}x_3$. Thus the Euler–Lagrange equations can be written as

$$\begin{aligned}&-\int_\Omega\left(\frac{\partial\zeta_{\mathbf{n}_0}}{\partial x_i}-n^0_i\right)\cos y_3dx=2\lambda_1q^0_{i1}+\lambda_4q^0_{i2}+\lambda_5q^0_{i3},\\&-\int_\Omega\left(\frac{\partial\zeta_{\mathbf{n}_0}}{\partial x_i}-n^0_i\right)\sin y_3dx=2\lambda_2q^0_{i2}+\lambda_4q^0_{i1}+\lambda_6q^0_{i3},\\&\sum_{k=1}^{3}\int_\Omega\left(\frac{\partial\zeta_{\mathbf{n}_0}}{\partial x_k}-n^0_k\right)\left(q^0_{k1}\sin y_3-q^0_{k2}\cos y_3\right)x_idx=2\lambda_3q^0_{i3}+\lambda_5q^0_{i1}+\lambda_6q^0_{i2}.\end{aligned}\tag{3.9}$$

If we multiply the first equation of (3.9) by q^0_{i3} and take sum over i, we get

$$-\sum_{i=1}^{3}q^0_{i3}\int_\Omega\left(\frac{\partial\zeta_{\mathbf{n}_0}}{\partial x_i}-n^0_i\right)\cos y_3dx=\lambda_5.$$

The left side of the above equality is equal to

$$-\int_\Omega(\nabla\zeta_{\mathbf{n}_0}-\mathbf{n}_0)\nabla\sin y_3dx,$$

which is zero by (2.3) (letting $\mathbf{u}=\mathbf{n}_0$). Thus $\lambda_5=0$. Similarly, we multiply the second equality in (3.9) by q^0_{i3} and take sum over i to get

$$\lambda_6=-\sum_{i=1}^{3}q^0_{i3}\int_\Omega\left(\frac{\partial\zeta_{\mathbf{n}_0}}{\partial x_i}-n^0_i\right)\sin y_3dx=\int_\Omega(\nabla\zeta_{\mathbf{n}_0}-\mathbf{n}_0)\nabla\cos y_3dx=0.$$

So we can write (3.9) as

$$
\begin{aligned}
&-\int_\Omega \left(\frac{\partial \zeta_{\mathbf{n}_0}}{\partial x_i} - n_i^0\right)\cos y_3 dx = 2\lambda_1 q_{i1}^0 + \lambda_4 q_{i2}^0,\\
&-\int_\Omega \left(\frac{\partial \zeta_{\mathbf{n}_0}}{\partial x_i} - \mathbf{n}_i^0\right)\sin y_3 dx = 2\lambda_2 q_{i2}^0 + \lambda_4 q_{i1}^0,\\
&\sum_{k=1}^3 \int_\Omega \left(\frac{\partial \zeta_{\mathbf{n}_0}}{\partial x_k} - n_k^0\right)(q_{k1}^0 \sin y_3 - q_{k2}^0\cos y_3) x_i dx = 2\lambda_3 q_{i3}^0.
\end{aligned}
\tag{3.10}
$$

Step 3. Let $\mathbf{u}_0 \in \mathcal{U}(1, \mathbf{n}_0)$. From Lemma 3.1,

$$
\mathbf{u}_0 = \sum_{j=1}^3 c_j \mathbf{U}_j,
$$

where

$$
\begin{aligned}
\mathbf{U}_j &= (U_{j1}, U_{j2}, U_{j3})^\top,\\
U_{1k} &= y_1(q_{k1}^0 \sin y_3 - q_{k2}^0 \cos y_3) + q_{k3}^0 \cos y_3,\\
U_{2k} &= y_2(q_{k1}^0 \sin y_3 - q_{k2}^0 \cos y_3) + q_{k3}^0 \sin y_3,\\
U_{3k} &= q_{k1}^0 \sin y_3 - q_{k2}^0 \cos y_3.
\end{aligned}
$$

Recall that $y_j = \sum_{i=1}^3 q_{ij}^0 x_i$ and $\mathbf{U}_j \cdot \mathbf{n}_0 = 0$. We use (3.10) to compute

$$
\begin{aligned}
-\int_\Omega \nabla\zeta_{\mathbf{n}_0} \cdot \mathbf{U}_1 dx &= -\int_\Omega (\nabla\zeta_{\mathbf{n}_0} - \mathbf{n}_0)\cdot \mathbf{U}_1 dx\\
&= -\sum_{k=1}^3 \int_\Omega \left(\frac{\partial \zeta_{\mathbf{n}_0}}{\partial x_k} - n_k^0\right)[y_1(q_{k1}^0 \sin y_3 - q_{k2}^0 \cos y_3) + q_{k3}^0 \cos y_3] dx\\
&= -\sum_{i,k=1}^3 \int_\Omega \left(\frac{\partial \zeta_{\mathbf{n}_0}}{\partial x_i} - n_i^0\right) q_{i1}^0 x_i (q_{k1}^0 \sin y_3 - q_{k2}^0 \cos y_3) dx\\
&\quad -\sum_{k=1}^3 q_{k3}^0 \int_\Omega \left(\frac{\partial \zeta_{\mathbf{n}_0}}{\partial x_k} - n_k^0\right)\cos y_3 dx\\
&= -\sum_{i=1}^3 2q_{i1}^0 \lambda_3 q_{i3}^0 + \sum_{k=1}^3 q_{k3}^0 (2\lambda_1 q_{k1}^0 + \lambda_4 q_{k2}^0)\\
&= 0;
\end{aligned}
$$

$$
\begin{aligned}
-\int_\Omega \nabla\zeta_{\mathbf{n}_0} \cdot \mathbf{U}_2 dx &= -\int_\Omega (\nabla\zeta_{\mathbf{n}_0} - \mathbf{n}_0)\cdot \mathbf{U}_2 dx\\
&= -\sum_{k=1}^3 \int_\Omega \left(\frac{\partial \zeta_{\mathbf{n}_0}}{\partial x_k} - n_k^0\right)[y_2(q_{k1}^0 \sin y_3 - q_{k2}^0 \cos y_3) + q_{k3}^0 \sin y_3] dx\\
&= -\sum_{i,k=1}^3 \int_\Omega \left(\frac{\partial \zeta_{\mathbf{n}_0}}{\partial x_i} - n_i^0\right) q_{i2}^0 x_i (q_{k1}^0 \sin y_3 - q_{k2}^0 \cos y_3) dx\\
&\quad -\sum_{k=1}^3 q_{k3}^0 \int_\Omega \left(\frac{\partial \zeta_{\mathbf{n}_0}}{\partial x_k} - n_k^0\right)\sin y_3 dx
\end{aligned}
$$

$$= -\sum_{i=1}^{3} 2q_{i2}^0\lambda_3 q_{i3}^0 + \sum_{k=1}^{3} q_{k3}^0(2\lambda_2 q_{k2}^0 + \lambda_4 q_{k1}^0)$$
$$= 0;$$

$$\begin{aligned}
-\int_\Omega \nabla\zeta_{\mathbf{n}_0}\cdot \mathbf{U}_3 dx &= -\int_\Omega (\nabla\zeta_{\mathbf{n}_0} - \mathbf{n}_0)\cdot \mathbf{U}_3 dx \\
&= -\sum_{i=1}^{3}\int_\Omega \left(\frac{\partial \zeta_{\mathbf{n}_0}}{\partial x_i} - n_i^0\right)(q_{i1}^0\sin y_3 - q_{i2}^0\cos y_3)dx \\
&= -\sum_{i=1}^{2} q_{i1}^0\int_\Omega \left(\frac{\partial \zeta_{\mathbf{n}_0}}{\partial x_i} - n_i^0\right)\sin y_3 dx \\
&\quad + \sum_{i=1}^{2} q_{i2}^0\int_\Omega \left(\frac{\partial \zeta_{\mathbf{n}_0}}{\partial x_i} - n_i^0\right)\cos y_3 dx \\
&= \sum_{i=1}^{3} q_{i1}^0(2\lambda_2 q_{i2}^0 + \lambda_4 q_{i1}^0) - \sum_{i=1}^{3} q_{i2}^0(2\lambda_1 q_{i1}^0 + \lambda_4 q_{i2}^0) \\
&= \lambda_4 - \lambda_4 = 0.
\end{aligned}$$

Thus, (3.8) is true. $\square$

4. Estimates of $\mathbf{{}_{c_3}}$ and energy upper bound for small . In this section we shall give an estimate of Q_{c_3} for small κ and establish an energy upper bound. For simplicity of notation, we write $\kappa = \varepsilon$. Throughout this section we assume that

(4.1) $\qquad K_1$, K_2, and τ are fixed positive constants, and $\kappa = \varepsilon > 0$ is small.

We write the functional $\mathcal{G}$ as

$$\mathcal{G}_\varepsilon[\psi, \mathbf{n}] = \int_\Omega \left\{ |\nabla_{q\mathbf{n}}\psi|^2 + \frac{\varepsilon^2}{2}(1-|\psi|^2)^2 + K_1|\text{div } \mathbf{n}|^2 + K_2|\text{curl } \mathbf{n} + \tau\mathbf{n}|^2 \right\} dx.$$

Then

$$C(K_1, K_2, \varepsilon, \tau, q) = \inf_{(\psi,\mathbf{n})\in W^{1,2}(\Omega,\mathbb{C})\times V(\Omega,\mathbb{S}^2)} \mathcal{G}_\varepsilon[\psi, \mathbf{n}].$$

The existence of minimizers has been proved in [P1].

THEOREM 4.1. *Assume that conditions* (1.2) *and* (4.1) *hold. We have, as* $\varepsilon \to 0$,

$$\frac{\varepsilon}{\sqrt{\omega_*(\tau)}} - L\varepsilon^3 \le Q_{c_3}(K_1, K_2, \varepsilon, \tau) = \frac{\varepsilon}{\sqrt{\omega_*(\tau)}} + o(\varepsilon), \tag{4.2}$$

where $\omega_*(\tau)$ *was defined in* (1.8), *and* $L > 0$ *depends only on* Ω *and* τ.

Proof. Step 1. Choose $\mathbf{n}_0 \in \mathcal{C}_*(\tau)$ such that

$$\|\zeta_{\mathbf{n}_0}\|_{L^2(\Omega)} = \min_{\mathbf{n}\in\mathcal{C}_*(\tau)} \|\zeta_{\mathbf{n}}\|_{L^2(\Omega)}.$$

Let

$$L_* = \frac{1-\omega_*(\tau)}{2\omega_*(\tau)^{5/2}|\Omega|}\|\zeta_{\mathbf{n}_0}\|^2_{L^2(\Omega)}.$$

Then L_* depends only on Ω and τ. Fix any positive number $L > L_*$. We show that there exists $\varepsilon_0 > 0$ depending only on Ω, τ, and L such that

$$Q_{c_3}(K_1, K_2, \varepsilon, \tau) \geq \frac{\varepsilon}{\sqrt{\omega_*(\tau)}} - L\varepsilon^3 \quad \text{for all } 0 < \varepsilon \leq \varepsilon_0. \tag{4.3}$$

Let us choose a number a_ε satisfying

$$0 < a_\varepsilon < \frac{1}{\sqrt{\omega_*(\tau)}} - L\varepsilon^2. \tag{4.4}$$

For the vector field $\mathbf{n}_0$ chosen above, we have $\omega(\mathbf{n}_0) = \omega_*(\tau)$, and from Lemma 2.2,

$$\begin{aligned}
\mu(a_\varepsilon \varepsilon \mathbf{n}_0) &\leq \omega_*(\tau) a_\varepsilon^2 \varepsilon^2 + \frac{1 - \omega_*(\tau)}{|\Omega|} \|\zeta_{\mathbf{n}_0}\|^2_{L^2(\Omega)} a_\varepsilon^4 \varepsilon^4 + O(\varepsilon^6) \\
&\leq \varepsilon^2 - \left[2L\sqrt{\omega_*(\tau)} - \frac{1 - \omega_*(\tau)}{\omega_*(\tau)^2 |\Omega|} \|\zeta_{\mathbf{n}_0}\|^2_{L^2(\Omega)} \right] \varepsilon^4 + O(\varepsilon^6) \\
&= \varepsilon^2 - 2(L - L_*)\sqrt{\omega_*(\tau)}\varepsilon^4 + O(\varepsilon^6).
\end{aligned}$$

So

$$\varepsilon^2 - \mu(a_\varepsilon \varepsilon \mathbf{n}_0) \geq 2(L - L_*)\sqrt{\omega_*(\tau)}\varepsilon^4 + O(\varepsilon^6) \quad \text{for all small } \varepsilon.$$

Let ϕ_0 be the eigenfunction of (2.7) for $\mathbf{u} = \mathbf{n}_0$ associated with the lowest eigenvalue $\mu(a\varepsilon \mathbf{n}_0)$. For $q = a_\varepsilon \varepsilon$ we take $\psi = t\phi_0$ and $\mathbf{n} = \mathbf{n}_0$ as test functions for $\mathcal{G}_\varepsilon$ and find that, for sufficiently small $t > 0$,

$$\begin{aligned}
\mathcal{G}_\varepsilon[t\phi_0, \mathbf{n}_0] &= \frac{\varepsilon^2|\Omega|}{2} - t^2 \int_\Omega \left\{ (\varepsilon^2 - \mu(a_\varepsilon \varepsilon \mathbf{n}_0))|\phi_0|^2 - \frac{t^2}{2}|\phi_0|^4 \right\} dx \\
&< \frac{\varepsilon^2|\Omega|}{2}.
\end{aligned}$$

Thus the minimizers are nontrivial, and hence $Q_{c_3}(K_1, K_2, \varepsilon, \tau) \geq a_\varepsilon \varepsilon$ for all small ε. So (4.3) is true.

Step 2. Now we prove an upper bound

$$\limsup_{\varepsilon \to 0} \frac{1}{\varepsilon} Q_{c_3}(K_1, K_2, \varepsilon, \tau) \leq \frac{1}{\sqrt{\omega_*(\tau)}}.$$

Let us choose a_ε such that

$$0 < a_\varepsilon < \frac{1}{\varepsilon} Q_{c_3}(K_1, K_2, \varepsilon, \tau).$$

We may assume $a_\varepsilon \to a_0 > 0$ as $\varepsilon \to 0$. Then $\mathcal{G}_\varepsilon$ has a nontrivial minimizer $(\psi_\varepsilon, \mathbf{n}_\varepsilon)$ for $q = a_\varepsilon \varepsilon$. Recall that $(0, \mathbf{N}_\tau)$ is a trivial critical point of $\mathcal{G}_\varepsilon$, and

$$\mathcal{G}_\varepsilon[0, \mathbf{N}_\tau] = \frac{\varepsilon^2|\Omega|}{2}.$$

Thus

$$\begin{aligned}
&\int_\Omega \left\{ |\nabla_{a_\varepsilon \varepsilon \mathbf{n}_\varepsilon} \psi_\varepsilon|^2 + \frac{\varepsilon^2}{2}(1 - |\psi_\varepsilon|^2)^2 + K_1|\text{div } \mathbf{n}_\varepsilon|^2 + K_2|\text{curl } \mathbf{n}_\varepsilon + \tau \mathbf{n}_\varepsilon|^2 \right\} dx \\
&\quad < \frac{\varepsilon^2|\Omega|}{2}.
\end{aligned} \tag{4.5}$$

Since $|\mathbf{n}_\varepsilon| = 1$, from (4.5) we see that $\{\mathbf{n}_\varepsilon\}$ is bounded in $V(\Omega, \mathbb{R}^3)$. Using Lemma 2.3, we find that there exist a subsequence, still denoted by $\mathbf{n}_\varepsilon$, and a vector field $\mathbf{n}_0 \in V(\Omega, \mathbb{R}^3)$ such that, as $\varepsilon \to 0$,

$$\begin{aligned}\mathbf{n}_\varepsilon \to \mathbf{n}_0 \quad & \text{weakly in } L^p(\Omega, \mathbb{R}^3) \text{ for all } 1 \le p < \infty, \text{ weakly in } W^{1,2}_{\mathrm{loc}}(\Omega, \mathbb{R}^3), \\ & \text{and strongly in } L^4_{\mathrm{loc}}(\Omega, \mathbb{R}^3).\end{aligned}$$

Then $|\mathbf{n}_0(x)| = 1$ a.e. in Ω. Since $|\mathbf{n}_\varepsilon| = |\mathbf{n}_0| = 1$, we see that

$$\mathbf{n}_\varepsilon \to \mathbf{n}_0 \quad \text{strongly in } L^p(\Omega, \mathbb{R}^3) \text{ for all } 1 \le p < \infty.$$

Using (4.5) again we see that $\operatorname{div} \mathbf{n}_0 = 0$ and $\operatorname{curl} \mathbf{n}_0 + \tau \mathbf{n}_0 = \mathbf{0}$ a.e. in Ω. Thus $\mathbf{n}_0 \in \mathcal{C}(\tau)$.

Let $\mathbf{w}_\varepsilon = \mathbf{n}_\varepsilon - \mathbf{n}_0$. Then

$$\mathbf{w}_\varepsilon \to \mathbf{0} \quad \text{strongly in } L^p(\Omega, \mathbb{R}^3) \text{ as } \varepsilon \to 0, \text{ for all } 1 \le p < \infty.$$

Using (4.5) we have

$$\begin{aligned}\mathcal{G}_\varepsilon[\psi_\varepsilon, \mathbf{n}_\varepsilon] = \int_\Omega \Big\{ & |\nabla_{a_\varepsilon \varepsilon(\mathbf{n}_0 + \mathbf{w}_\varepsilon)} \psi_\varepsilon|^2 + \frac{\varepsilon^2}{2}(|\psi_\varepsilon|^2 - 1)^2 \\ & + K_1 |\operatorname{div} \mathbf{w}_\varepsilon|^2 + K_2 |\operatorname{curl} \mathbf{w}_\varepsilon + \tau \mathbf{w}_\varepsilon|^2 \Big\} dx \\ < {} & \frac{\varepsilon^2 |\Omega|}{2}.\end{aligned}$$

In particular,

$$\int_\Omega \left\{ |\nabla_{a_\varepsilon \varepsilon(\mathbf{n}_0 + \mathbf{w}_\varepsilon)} \psi_\varepsilon|^2 - \varepsilon^2 |\psi_\varepsilon|^2 \right\} dx < 0.$$

So the lowest eigenvalue of $-\nabla^2_{a_\varepsilon \varepsilon(\mathbf{n}_0 + \mathbf{w}_\varepsilon)}$ (see (2.7)) is less than ε^2:

$$\mu\big(a_\varepsilon \varepsilon(\mathbf{n}_0 + \mathbf{w}_\varepsilon)\big) < \varepsilon^2.$$

Let ϕ^ε be the associated eigenfunction such that $\|\phi^\varepsilon\|_{L^2(\Omega)} = 1$. Then

$$\|\nabla_{a_\varepsilon \varepsilon(\mathbf{n}_0 + \mathbf{w}_\varepsilon)} \phi^\varepsilon\|^2_{L^2(\Omega)} = \mu\big(a_\varepsilon \varepsilon(\mathbf{n}_0 + \mathbf{w}_\varepsilon)\big) \|\phi^\varepsilon\|^2_{L^2(\Omega)} < \varepsilon^2 \|\phi^\varepsilon\|^2_{L^2(\Omega)}.$$

Using the Sobolev imbedding theorem and Kato's inequality, we have

$$\begin{aligned}\|\phi^\varepsilon\|_{L^4(\Omega)} & \le C(\Omega) \left\{ \|\nabla|\phi^\varepsilon|\|_{L^2(\Omega)} + \|\phi^\varepsilon\|_{L^2(\Omega)} \right\} \\ & \le C(\Omega) \left\{ \|\nabla_{a_\varepsilon \varepsilon(\mathbf{n}_0 + \mathbf{w}_\varepsilon)} \phi^\varepsilon\|_{L^2(\Omega)} + \|\phi^\varepsilon\|_{L^2(\Omega)} \right\} \\ & < C(\Omega)(1 + \varepsilon) \|\phi^\varepsilon\|_{L^2(\Omega)}.\end{aligned} \tag{4.6}$$

So

$$\begin{aligned}
&\varepsilon^2 \int_\Omega |\phi^\varepsilon|^2 dx > \mu\big(a_\varepsilon\varepsilon(\mathbf{n}_0+\mathbf{w}_\varepsilon)\big) \int_\Omega |\phi^\varepsilon|^2 dx \\
&= \int_\Omega |\nabla_{a_\varepsilon\varepsilon(\mathbf{n}_0+\mathbf{w}_\varepsilon)}\phi_\varepsilon|^2 dx \\
&= \int_\Omega \left\{|\nabla_{a_\varepsilon\varepsilon\mathbf{n}_0}\phi^\varepsilon|^2 - 2a_\varepsilon\varepsilon\mathbf{w}_\varepsilon\Im[\bar\phi^\varepsilon\nabla_{a_\varepsilon\varepsilon(\mathbf{n}_0+\mathbf{w}_\varepsilon)}\phi^\varepsilon] - a_\varepsilon^2\varepsilon^2|\mathbf{w}_\varepsilon\phi^\varepsilon|^2\right\} dx \\
&\geq \int_\Omega |\nabla_{a_\varepsilon\varepsilon\mathbf{n}_0}\phi^\varepsilon|^2 dx - 2a_\varepsilon\varepsilon\|\mathbf{w}_\varepsilon\|_{L^4(\Omega)}\|\phi^\varepsilon\|_{L^4(\Omega)}\|\nabla_{a_\varepsilon\varepsilon(\mathbf{n}_0+\mathbf{w}_\varepsilon)}\phi^\varepsilon\|_{L^2(\Omega)} \\
&\qquad - a_\varepsilon^2\varepsilon^2\|\mathbf{w}_\varepsilon\|^2_{L^4(\Omega)}\|\phi^\varepsilon\|^2_{L^4(\Omega)} \\
&\geq \int_\Omega |\nabla_{a_\varepsilon\varepsilon\mathbf{n}_0}\phi^\varepsilon|^2 dx - Ca_\varepsilon\varepsilon^2\|\mathbf{w}_\varepsilon\|_{L^4(\Omega)}\|\phi^\varepsilon\|^2_{L^2(\Omega)} - Ca_\varepsilon^2\varepsilon^2\|\mathbf{w}_\varepsilon\|^2_{L^4(\Omega)}\|\phi^\varepsilon\|^2_{L^2(\Omega)},
\end{aligned}$$

where we have used (4.6). As $\varepsilon \to 0$, $a_\varepsilon \to a_0$, and $\mathbf{w}_\varepsilon \to \mathbf{0}$ in $L^4(\Omega, \mathbb{R}^3)$, so

$$\begin{aligned}
\omega(\mathbf{n}_0)a_0^2\varepsilon^2 + o(\varepsilon^2) \leq \mu(a_\varepsilon\varepsilon\mathbf{n}_0) &\leq \frac{\int_\Omega |\nabla_{a_\varepsilon\varepsilon\mathbf{n}_0}\phi^\varepsilon|^2 dx}{\|\phi^\varepsilon\|^2_{L^2(\Omega)}} \\
&\leq \varepsilon^2 + Ca_\varepsilon\varepsilon^2\|\mathbf{w}_\varepsilon\|_{L^4(\Omega)} + Ca_\varepsilon^2\varepsilon^2\|\mathbf{w}_\varepsilon\|^2_{L^4(\Omega)} \\
&= \varepsilon^2(1+o(1)).
\end{aligned}$$

So $\omega(\mathbf{n}_0)a_0^2 \leq 1$; that is,

$$0 < a_0 \leq \frac{1}{\sqrt{\omega(\mathbf{n}_0)}} \leq \frac{1}{\sqrt{\omega_*(\tau)}}.$$

Thus, the upper bound is true. ∎

Next we look for an energy upper bound of the minimizers. For a vector field $\mathbf{n}_0 \in \mathcal{C}_*(\tau)$ and $a > 0$, we define a function of t by

$$f_{a,\mathbf{n}_0}(t) = a^2\omega(\mathbf{n}_0)t^2 + \frac{1}{2}(1-t^2)^2. \tag{4.7}$$

Obviously,

$$\min_{t\geq 0} f_{a,\mathbf{n}_0}(t) = \begin{cases} \frac{1}{2} & \text{if } a \geq \frac{1}{\sqrt{\omega(\mathbf{n}_0)}}, \\ a^2\omega(\mathbf{n}_0)[1-\frac{1}{2}a^2\omega(\mathbf{n}_0)] & \text{if } 0 < a < \frac{1}{\sqrt{\omega(\mathbf{n}_0)}}, \end{cases}$$

and the minimum is attained at

$$t = \sqrt{[1-a^2\omega(\mathbf{n}_0)]_+};$$

here we use the notation $a_+ = \max\{a, 0\}$. Let

$$d_0(\varepsilon) = a_\varepsilon^2\omega_*(\tau)\left[1 - \frac{1}{2}a_\varepsilon^2\omega_*(\tau)\right]|\Omega|. \tag{4.8}$$

THEOREM 4.2. *Assume the conditions* (1.2) *and* (4.1). *For small* $\kappa = \varepsilon > 0$ *and* $q = a_\varepsilon\varepsilon$, *where* a_ε *satisfies* (4.4), *we have*

$$C(K_1, K_2, \varepsilon, \tau, a_\varepsilon\varepsilon) \leq d_0(\varepsilon)\varepsilon^2 - d(\varepsilon)\varepsilon^4 + o(\varepsilon^4), \tag{4.9}$$

where $d_0(\varepsilon)$ is given in (4.8), *and there exists a constant $c > 0$ independent of ε such that $-c \le d(\varepsilon) < +\infty$.*

The proof of (4.9) is lengthy and will be divided into several steps. Note that (4.9) is a rough estimate, and better upper bounds can be obtained; see the discussions below.

Step 1. *A general computation.* To begin with, we consider test functions in the form

$$\psi^\varepsilon = c^\varepsilon[1 + i\varepsilon a_\varepsilon(\zeta + \varepsilon\rho\varphi^\varepsilon)], \qquad \mathbf{n}^\varepsilon = \mathbf{n}_0 + \varepsilon\rho\mathbf{u}^\varepsilon, \tag{4.10}$$

where $\mathbf{n}_0 \in \mathcal{C}(\tau)$, $\mathbf{u}^\varepsilon \in V(\Omega, \mathbb{R}^3)$, such that

$$|\mathbf{n}_0(x) + \varepsilon\rho\mathbf{u}^\varepsilon(x)| = 1 \qquad \text{for a.e. } x \in \Omega, \tag{4.11}$$

c^ε is a real number, and ρ is a positive number. ζ and φ^ε are complex-valued functions such that

$$\int_\Omega \zeta dx = 0, \qquad \int_\Omega \varphi^\varepsilon dx = 0. \tag{4.12}$$

ρ and ζ may depend on ε. Let us define a set $V(\mathbf{n}_0, b)$ by

$$V(\mathbf{n}_0, b) = \left\{ \mathbf{u} \in V(\Omega, \mathbb{R}^3) : \ \mathbf{u}(x) \cdot \mathbf{n}_0(x) = -\frac{b}{2}|\mathbf{u}(x)|^2 \quad \text{a.e. in } \Omega \right\}.$$

From (4.11), $\mathbf{u}^\varepsilon \in V(\mathbf{n}_0, \varepsilon\rho)$. We compute

$$\begin{aligned}
|\nabla_{a_\varepsilon\varepsilon\mathbf{n}^\varepsilon}\psi^\varepsilon|^2 &= \varepsilon^2 a_\varepsilon^2 |c^\varepsilon|^2 |\nabla\zeta - \mathbf{n}_0 + \varepsilon\rho(\nabla\varphi^\varepsilon - \mathbf{u}^\varepsilon) - i\varepsilon a_\varepsilon(\zeta + \varepsilon\rho\varphi^\varepsilon)\mathbf{n}^\varepsilon|^2 \\
&= \varepsilon^2 a_\varepsilon^2 |c^\varepsilon|^2 \Big\{ |\nabla\zeta - \mathbf{n}_0|^2 + 2\varepsilon\rho\Re[(\nabla\zeta - \mathbf{n}_0)\cdot(\nabla\bar{\varphi}^\varepsilon - \mathbf{u}^\varepsilon)] + \varepsilon^2\rho^2|\nabla\varphi^\varepsilon - \mathbf{u}^\varepsilon|^2 \\
&\qquad - 2\varepsilon a_\varepsilon \Im[\bar{\zeta}\mathbf{n}_0 \cdot (\nabla\zeta - \mathbf{n}_0)] \\
&\qquad - 2\varepsilon^2\rho a_\varepsilon \Im[\bar{\zeta}\mathbf{n}_0 \cdot (\nabla\varphi^\varepsilon - \mathbf{u}^\varepsilon) + (\nabla\zeta - \mathbf{n}_0)\cdot(\bar{\zeta}\mathbf{u}^\varepsilon + \bar{\varphi}^\varepsilon\mathbf{n}_0)] \\
&\qquad - 2\varepsilon^3\rho^2 a_\varepsilon \Im[(\nabla\varphi^\varepsilon - \mathbf{u}^\varepsilon)\cdot(\bar{\zeta}\mathbf{u}^\varepsilon + \bar{\varphi}^\varepsilon\mathbf{n}^\varepsilon) + \bar{\varphi}^\varepsilon\mathbf{u}^\varepsilon\cdot(\nabla\zeta - \mathbf{n}_0)] \\
&\qquad + \varepsilon^2 a_\varepsilon^2 |\zeta + \varepsilon\rho\varphi^\varepsilon|^2 \Big\},
\end{aligned}$$

$$\begin{aligned}
(1 - |\psi^\varepsilon|^2)^2 &= \Big[1 - |c^\varepsilon|^2 + 2\varepsilon a_\varepsilon |c^\varepsilon|^2 \Im(\zeta + \varepsilon\rho\varphi^\varepsilon) - \varepsilon^2 a_\varepsilon^2 |c^\varepsilon|^2 |\zeta + \varepsilon\rho\varphi^\varepsilon|^2 \Big]^2 \\
&= (1 - |c^\varepsilon|^2)^2 + 4\varepsilon a_\varepsilon |c^\varepsilon|^2 (1 - |c^\varepsilon|^2)\Im(\zeta + \varepsilon\rho\varphi^\varepsilon) \\
&\quad - 2\varepsilon^2 a_\varepsilon^2 |c^\varepsilon|^2 (1 - |c^\varepsilon|^2)|\zeta + \varepsilon\rho\varphi^\varepsilon|^2 \\
&\quad + \varepsilon^2 a_\varepsilon^2 |c^\varepsilon|^4 \Big[2\Im(\zeta + \varepsilon\rho\varphi^\varepsilon) - \varepsilon a_\varepsilon |\zeta + \varepsilon\rho\varphi^\varepsilon|^2 \Big]^2.
\end{aligned}$$

Since $\mathcal{I}[\mathbf{n}_0] = 0$, we have

$$\mathcal{I}[\mathbf{n}_0 + \varepsilon\rho\mathbf{u}^\varepsilon] = \varepsilon^2\rho^2\mathcal{I}[\mathbf{u}^\varepsilon].$$

So

(4.13)

$$
\begin{aligned}
&\varepsilon^{-2}\mathcal{G}_\varepsilon[\psi^\varepsilon, \mathbf{n}^\varepsilon] \\
&= \int_\Omega \left\{ a_\varepsilon^2 |c^\varepsilon|^2 |\nabla\zeta - \mathbf{n}_0|^2 + \frac{1}{2}(1-|c^\varepsilon|^2)^2 \right\} dx \\
&+ \rho^2 \mathcal{I}[\mathbf{u}^\varepsilon] \\
&+ 2\varepsilon\rho a_\varepsilon^2 |c^\varepsilon|^2 \Re \int_\Omega (\nabla\zeta - \mathbf{n}_0)\cdot(\nabla\bar{\varphi}^\varepsilon - \mathbf{u}^\varepsilon) dx \\
&+ 2\varepsilon a_\varepsilon |c^\varepsilon|^2 \Im \int_\Omega \big\{ (1-|c^\varepsilon|^2)\zeta - a_\varepsilon^2 \bar{\zeta}\mathbf{n}_0 \cdot (\nabla\zeta - \mathbf{n}_0) \big\} dx \\
&+ \varepsilon^2\rho^2 a_\varepsilon^2 |c^\varepsilon|^2 \int_\Omega |\nabla\varphi^\varepsilon - \mathbf{u}^\varepsilon|^2 dx \\
&+ 2\varepsilon^2\rho a_\varepsilon |c^\varepsilon|^2 \Im \int_\Omega \big\{ (1-|c^\varepsilon|^2)\varphi^\varepsilon - a_\varepsilon^2 [\bar{\zeta}\mathbf{n}_0 \cdot (\nabla\varphi^\varepsilon - \mathbf{u}^\varepsilon) + (\nabla\zeta - \mathbf{n}_0)\cdot(\bar{\zeta}\mathbf{u}^\varepsilon + \bar{\varphi}^\varepsilon \mathbf{n}_0)] \big\} dx \\
&+ \varepsilon^2 a_\varepsilon^2 |c^\varepsilon|^2 (a_\varepsilon^2 - 1 + |c^\varepsilon|^2) \int_\Omega |\zeta + \varepsilon\rho\varphi^\varepsilon|^2 dx \\
&- 2\varepsilon^3\rho^2 a_\varepsilon^3 |c^\varepsilon|^2 \Im \int_\Omega \{ (\nabla\varphi^\varepsilon - \mathbf{u}^\varepsilon)\cdot(\bar{\zeta}\mathbf{u}^\varepsilon + \bar{\varphi}^\varepsilon \mathbf{n}^\varepsilon) + \bar{\varphi}^\varepsilon \mathbf{u}^\varepsilon \cdot (\nabla\zeta - \mathbf{n}_0) \} dx \\
&+ \frac{1}{2}\varepsilon^2 a_\varepsilon^2 |c^\varepsilon|^4 \int_\Omega \big[2\Im(\zeta + \varepsilon\rho\varphi^\varepsilon) - \varepsilon a_\varepsilon |\zeta + \varepsilon\rho\varphi^\varepsilon|^2 \big]^2 dx.
\end{aligned}
$$

Checking the first line in the right-hand side of (4.13), we choose

$$
\mathbf{n}_0 \in \mathcal{C}_*(\tau), \qquad \zeta = \zeta_{\mathbf{n}_0}, \qquad c^\varepsilon = \sqrt{1 - a_\varepsilon^2 \omega_*(\tau)}, \tag{4.14}
$$

where $\zeta_{\mathbf{n}_0}$ is determined by (1.7) for $\mathbf{n}_0$. Then the first line in the right-hand side of (4.13) becomes

$$
a_\varepsilon^2 (1 - a_\varepsilon^2 \omega_*(\tau))\omega_*(\tau)|\Omega| + \frac{1}{2}(a_\varepsilon^2 \omega_*(\tau))^2 |\Omega| = d_0(\varepsilon).
$$

Applying (2.3) for $\mathbf{u} = \mathbf{n}_0$ we have

$$
\int_\Omega (\nabla\zeta_{\mathbf{n}_0} - \mathbf{n}_0)\cdot \nabla\bar{\varphi}^\varepsilon dx = 0.
$$

So the third line in the right-hand side of (4.13) becomes

$$
-2\varepsilon\rho a_\varepsilon^2 |c^\varepsilon|^2 \int_\Omega (\nabla\zeta_{\mathbf{n}_0} - \mathbf{n}_0)\cdot \mathbf{u}^\varepsilon \, dx.
$$

Since $\zeta_{\mathbf{n}_0}$ is real-valued, the fourth line in the right-hand side of (4.13) vanishes. From (4.12), the sixth line in the right-hand side of (4.13) becomes

$$
-2\varepsilon^2\rho a_\varepsilon^3 |c^\varepsilon|^2 \Im \int_\Omega \mathbf{n}_0 \cdot (\zeta_{\mathbf{n}_0}\nabla\varphi^\varepsilon + \bar{\varphi}^\varepsilon \nabla\zeta_{\mathbf{n}_0}) dx.
$$

Let us define a functional $T_{\varepsilon,\rho}$ by

$$
T_{\varepsilon,\rho}[\mathbf{u}] = \rho^2 \mathcal{I}[\mathbf{u}] - 2\varepsilon\rho a_\varepsilon^2 |c^\varepsilon|^2 \int_\Omega (\nabla\zeta_{\mathbf{n}_0} - \mathbf{n}_0)\cdot \mathbf{u}\, dx. \tag{4.15}
$$

Then we can rewrite (4.13) as follows:

(4.16)
$$\begin{aligned}
&\varepsilon^{-2}\mathcal{G}_\varepsilon[\psi^\varepsilon,\mathbf{n}^\varepsilon]-d_0(\varepsilon)-T_{\varepsilon,\rho}[\mathbf{u}^\varepsilon]\\
&=\varepsilon^2\rho^2a_\varepsilon^2|c^\varepsilon|^2\int_\Omega|\nabla\varphi^\varepsilon-\mathbf{u}^\varepsilon|^2dx\\
&\quad-2\varepsilon^2\rho a_\varepsilon^3|c^\varepsilon|^2\Im\int_\Omega\mathbf{n}_0\cdot(\zeta_{\mathbf{n}_0}\nabla\varphi^\varepsilon+\bar\varphi^\varepsilon\nabla\zeta_{\mathbf{n}_0})dx\\
&\quad+\varepsilon^2a_\varepsilon^2|c^\varepsilon|^2(a_\varepsilon^2-1+|c^\varepsilon|^2)\int_\Omega|\zeta_{\mathbf{n}_0}+\varepsilon\rho\varphi^\varepsilon|^2dx\\
&\quad-2\varepsilon^3\rho^2a_\varepsilon^3|c^\varepsilon|^2\Im\int_\Omega\big\{\zeta_{\mathbf{n}_0}\mathbf{u}^\varepsilon\cdot\nabla\varphi^\varepsilon+\bar\varphi^\varepsilon\big[\mathbf{u}^\varepsilon\cdot(\nabla\zeta_{\mathbf{n}_0}-\mathbf{n}_0)+\mathbf{n}^\varepsilon\cdot(\nabla\varphi^\varepsilon-\mathbf{u}^\varepsilon)\big]\big\}dx\\
&\quad+\frac{1}{2}\varepsilon^4a_\varepsilon^2|c^\varepsilon|^4\int_\Omega\big[2\rho\Im(\varphi^\varepsilon)-a_\varepsilon|\zeta_{\mathbf{n}_0}+\varepsilon\rho\varphi^\varepsilon|^2\big]^2dx.
\end{aligned}$$

Step 2. *A variational problem in* $W(\mathbf{n}_0)$. Equation (4.16) suggests that $\mathbf{u}^\varepsilon$ and ρ should be chosen to minimize $T_{\varepsilon,\rho}$ asymptotically; namely,

$$T_{\varepsilon,\rho}[\mathbf{u}^\varepsilon]=T_v(\varepsilon)+o(1)\quad\text{as }\varepsilon\to0,$$

where

$$T_v(\varepsilon)=\inf_{0<\rho<+\infty}\ \inf_{\mathbf{u}\in V(\mathbf{n}_0,\varepsilon\rho)}T_{\varepsilon,\rho}[\mathbf{u}]. \tag{4.17}$$

In order to get an upper bound estimate of $T_v(\varepsilon)$, one may first look at the following variational problem, which can be derived as a formal limit of (4.17):

$$\begin{aligned}
t_v(\varepsilon,\rho)&=\inf_{\mathbf{u}\in V(\mathbf{n}_0)}T_{\varepsilon,\rho}[\mathbf{u}],\\
t_v(\varepsilon)&=\inf_{0<\rho<+\infty}t_v(\varepsilon,\rho),
\end{aligned}\tag{4.18}$$

where

$$V(\mathbf{n}_0)=\{\mathbf{u}\in V(\Omega,\mathbb{R}^3):\ \mathbf{u}(x)\cdot\mathbf{n}_0(x)=0\quad\text{a.e. in }\Omega\}.$$

We may try to use the minimizing sequence of (4.18) to construct test functions for $T_v(\varepsilon)$. However, to avoid technical complexity, instead of working on $V(\mathbf{n}_0)$, we consider the variational problem in a subset of the Sobolev space $W^{1,2}(\Omega,\mathbb{R}^3)$:

$$W(\mathbf{n}_0)=\{\mathbf{u}\in W^{1,2}(\Omega,\mathbb{R}^3):\ \mathbf{u}(x)\cdot\mathbf{n}_0(x)=0\quad\text{a.e. in }\Omega\}.$$

Define

$$\begin{aligned}
t_w(\varepsilon,\rho)&=\inf_{\mathbf{u}\in W(\mathbf{n}_0)}T_{\varepsilon,\rho}[\mathbf{u}],\\
t_w(\varepsilon)&=\inf_{0<\rho<+\infty}t_w(\varepsilon,\rho).
\end{aligned}\tag{4.19}$$

On $W(\mathbf{n}_0)$ we can write

$$T_{\varepsilon,\rho}[\mathbf{u}]=\rho^2\mathcal{I}[\mathbf{u}]-2\varepsilon\rho a_\varepsilon^2|c^\varepsilon|^2\int_\Omega\nabla\zeta_{\mathbf{n}_0}\cdot\mathbf{u}dx.$$

As a subspace of $W^{1,2}(\Omega,\mathbb{R}^3)$, the set $W(\mathbf{n}_0)$ has a decomposition

$$W(\mathbf{n}_0) = \mathcal{U}(\tau,\mathbf{n}_0) \oplus \mathcal{U}_w^{\perp}(\tau,\mathbf{n}_0),$$

where $\mathcal{U}_w^{\perp}(\tau,\mathbf{n}_0)$ is the orthogonal complement of $\mathcal{U}(\tau,\mathbf{n}_0)$ in $W(\mathbf{n}_0)$. Since $\mathbf{n}_0 \in \mathcal{C}_*(\tau)$, using Proposition 3.3 we see that

$$T_{\varepsilon,\rho}[\mathbf{u}] = 0 \quad \text{for all } \mathbf{u} \in \mathcal{U}(\tau,\mathbf{n}_0).$$

Thus

$$t_w(\varepsilon,\rho) = \inf_{\mathbf{u}\in\mathcal{U}_w^{\perp}(\tau,\mathbf{n}_0)} T_{\varepsilon,\rho}[\mathbf{u}].$$

It is interesting that the variational problem (4.19) is linked to the following number:

$$\mu_w(\mathbf{n}_0) = \sup_{\mathbf{v}\in W(\mathbf{n}_0),\ \mathcal{I}[\mathbf{v}]>0} \frac{\int_\Omega \nabla\zeta_{\mathbf{n}_0}\cdot\mathbf{v}dx}{\sqrt{\mathcal{I}[\mathbf{v}]}}. \tag{4.20}$$

The advantage of choosing $W(\mathbf{n}_0)$ as the admissible set is that we know $C^1(\bar\Omega,\mathbb{R}^3)$ is dense in $W(\mathbf{n}_0)$. Since $\mathcal{I}[\mathbf{v}] \neq 0$ for nonzero elements $\mathbf{v} \in \mathcal{U}_w^{\perp}(\tau,\mathbf{n}_0)$, and using the homogeneity of the ratio, we have

$$\begin{aligned}\mu_w(\mathbf{n}_0) &= \sup\left\{\frac{\int_\Omega \nabla\zeta_{\mathbf{n}_0}\cdot\mathbf{v}dx}{\sqrt{\mathcal{I}[\mathbf{v}]}} :\ \mathbf{v}\in\mathcal{U}_w^{\perp}(\tau,\mathbf{n}_0)\cap C^1(\bar\Omega,\mathbb{R}^3)\right\}\\ &= \sup\left\{\frac{\int_\Omega \nabla\zeta_{\mathbf{n}_0}\cdot\mathbf{v}dx}{\sqrt{\mathcal{I}[\mathbf{v}]}} :\ \mathbf{v}\in\mathcal{U}_w^{\perp}(\tau,\mathbf{n}_0)\cap C^1(\bar\Omega,\mathbb{R}^3),\ \|\mathbf{v}\|_{C^1(\bar\Omega,\mathbb{R}^3)}=1\right\}.\end{aligned}$$

Note that, if $\{\mathbf{v}_j\} \subset W(\mathbf{n}_0)$ is a maximizing sequence of $\mu_w(\mathbf{n}_0)$, then for any $\mathbf{u}_j \in \mathcal{U}(\tau,\mathbf{n}_0)$, $\{\mathbf{u}_j+\mathbf{v}_j\}$ is also a maximizing sequence of $\mu_w(\mathbf{n}_0)$ in $W(\mathbf{n}_0)$.

LEMMA 4.3. *Let $\tau>0$ and $\mathbf{n}_0\in\mathcal{C}_*(\tau)$.*

(1) $0<\mu_w(\mathbf{n}_0)\le+\infty$.

(2) *Assume $\mu_w(\mathbf{n}_0)<+\infty$. If $\mu_w(\mathbf{n}_0)$ is achieved, we can choose $\rho^\varepsilon>0$, $\rho^\varepsilon = O(\varepsilon)$ as $\varepsilon\to0$, such that*

$$t_w(\varepsilon) = t_w(\varepsilon,\rho^\varepsilon) = -\varepsilon^2 a_\varepsilon^4|c^\varepsilon|^4\mu_w(\mathbf{n}_0)^2. \tag{4.21}$$

If $\mu_w(\mathbf{n}_0)$ is not achieved, we can choose $\rho^\varepsilon>0$, $\rho^\varepsilon \ll \varepsilon|\log\varepsilon|$ as $\varepsilon\to0$, such that

$$-\varepsilon^2 a_\varepsilon^4|c^\varepsilon|^4\mu_w(\mathbf{n}_0)^2 \le t_w(\varepsilon) \le t_w(\varepsilon,\rho^\varepsilon) = -\varepsilon^2 a_\varepsilon^4|c^\varepsilon|^4\mu_w(\mathbf{n}_0)^2 + o(\varepsilon^2). \tag{4.22}$$

(3) *Assume $\mu_w(\mathbf{n}_0)=+\infty$. Then $t_w(\varepsilon)=-\infty$ for any $\varepsilon>0$.*

Proof of (1). Suppose $\mu_w(\mathbf{n}_0)=0$. Then we use Proposition 3.3 to conclude that

$$\int_\Omega \nabla\zeta_{\mathbf{n}_0}\cdot\mathbf{u}dx = 0 \qquad \text{for all } \mathbf{u}\in W(\mathbf{n}_0). \tag{4.23}$$

Let us choose

$$\mathbf{v}_0 = \nabla\zeta_{\mathbf{n}_0} - (\mathbf{n}_0\cdot\nabla\zeta_{\mathbf{n}_0})\mathbf{n}_0.$$

Then $\mathbf{v}_0 \in W(\mathbf{n}_0)$, and

$$\int_\Omega \nabla\zeta_{\mathbf{n}_0} \cdot \mathbf{v}_0 dx = \int_\Omega \{|\nabla\zeta_{\mathbf{n}_0}|^2 - (\mathbf{n}_0 \cdot \nabla\zeta_{\mathbf{n}_0})^2\} dx.$$

From (4.23), the right side of the above equality is zero. From this and the fact $|\mathbf{n}_0 \cdot \nabla\zeta_{\mathbf{n}_0}| \le |\nabla\zeta_{\mathbf{n}_0}|$, we must have

$$\nabla\zeta_{\mathbf{n}_0} = g\mathbf{n}_0 \quad \text{on } \Omega,$$

where $g = \mathbf{n}_0 \cdot \nabla\zeta_{\mathbf{n}_0}$. Applying the operator curl to the above equality we get

$$\mathbf{0} = \text{curl } (\nabla\zeta_{\mathbf{n}_0}) = \text{curl } (g\mathbf{n}_0) = g\text{curl } \mathbf{n}_0 + \nabla g \times \mathbf{n}_0 = -\tau g\mathbf{n}_0 + \nabla g \times \mathbf{n}_0.$$

So $\tau g\mathbf{n}_0 = \nabla g \times \mathbf{n}_0$, and hence $g = \frac{1}{\tau}\mathbf{n}_0 \cdot (\nabla g \times \mathbf{n}_0) = 0$. Thus $\nabla\zeta_{\mathbf{n}_0} = g\mathbf{n}_0 = \mathbf{0}$, which is a contradiction.

Proof of (2). Assume $\mu_w(\mathbf{n}_0) < +\infty$. For any $\rho > 0$, if $\mathbf{u} \in \mathcal{U}(\tau, \mathbf{n}_0)$, then $T_{\varepsilon,\rho}[\mathbf{u}] = 0$. Now if $\mathbf{u} \in W(\mathbf{n}_0) \setminus \mathcal{U}(\tau, \mathbf{n}_0)$, then

$$\begin{aligned} T_{\varepsilon,\rho}[\mathbf{u}] &= \rho^2\mathcal{I}[\mathbf{u}] - 2a_\varepsilon^2|c^\varepsilon|^2\varepsilon\rho\int_\Omega \nabla\zeta_{\mathbf{n}_0} \cdot \mathbf{u} dx \\ &= \left[\rho\sqrt{\mathcal{I}[\mathbf{u}]} - a_\varepsilon^2|c^\varepsilon|^2\varepsilon\frac{\int_\Omega \nabla\zeta_{\mathbf{n}_0} \cdot \mathbf{u} dx}{\sqrt{\mathcal{I}[\mathbf{u}]}}\right]^2 - \varepsilon^2 a_\varepsilon^4|c^\varepsilon|^4\left[\frac{\int_\Omega \nabla\zeta_{\mathbf{n}_0} \cdot \mathbf{u} dx}{\sqrt{\mathcal{I}[\mathbf{u}]}}\right]^2 \\ &\ge -\varepsilon^2 a_\varepsilon^4|c^\varepsilon|^4\mu_w(\mathbf{n}_0)^2. \end{aligned}$$

Hence

$$t_w(\varepsilon) \ge -\varepsilon^2 a_\varepsilon^4|c^\varepsilon|^4\mu_w(\mathbf{n}_0)^2.$$

Case 1. $\mu_w(\mathbf{n}_0)$ is achieved. We fix a maximizer $\mathbf{v}_0$ of $\mu_w(\mathbf{n}_0)$ with $\|\mathbf{v}_0\|_{C^1(\bar\Omega,\mathbb{R}^3)} = 1$, and let

$$\rho^\varepsilon = \frac{\varepsilon a_\varepsilon^2|c^\varepsilon|^2\mu_w(\mathbf{n}_0)}{\mathcal{I}[\mathbf{v}_0]}.$$

We have

$$T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}_0] = -a_\varepsilon^4|c^\varepsilon|^4\mu_w(\mathbf{n}_0)^2.$$

From the choice of a_ε and c^ε, we have $\rho^\varepsilon = O(\varepsilon)$ as $\varepsilon \to 0$.

Case 2. $\mu_w(\mathbf{n}_0)$ is not achieved. We can show that, if $\mu_w(\mathbf{n}_0)$ is not achieved, then the following conclusions are true:

(i) if $\{\mathbf{v}_j\} \subset W(\mathbf{n}_0)$ is a maximizing sequence of $\mu_w(\mathbf{n}_0)$ and if it is bounded in $W^{1,2}(\Omega, \mathbb{R}^3)$, then

$$\mathcal{I}[\mathbf{v}_j] \to 0 \quad \text{as } j \to \infty;$$

(ii) if in addition $\{\mathbf{v}_j\} \subset \mathcal{U}_w^\perp(\tau, \mathbf{n}_0)$, then

$$\mathbf{v}_j \to \mathbf{0} \quad \text{weakly in } W^{1,2}(\Omega, \mathbb{R}^3) \text{ as } j \to \infty.$$

To prove (i) and (ii), let $\{\mathbf{v}_j\} \subset W(\mathbf{n}_0)$ be a maximizing sequence of $\mu_w(\mathbf{n}_0)$ that is bounded in $W^{1,2}(\Omega, \mathbb{R}^3)$. We pass to a subsequence and assume that $\mathbf{v}_j \to \mathbf{v}_0$ weakly in $W^{1,2}(\Omega,\mathbb{R}^3)$. Since $\mu_w(\mathbf{n}_0)$ is not achieved, we have $\mathcal{I}[\mathbf{v}_0]=0$, because otherwise $\mathbf{v}_0$ achieves $\mu_w(\mathbf{n}_0)$. Hence $\mathcal{I}[\mathbf{v}_j]\to 0$. Now if $\{\mathbf{v}_j\}\subset\mathcal{U}_w^{\perp}(\tau,\mathbf{n}_0)$, then $\mathbf{v}_0\in\mathcal{U}_w^{\perp}(\tau,\mathbf{n}_0)$. Since $\mathcal{I}[\mathbf{v}_0]=0$, we have $\mathbf{v}_0\in\mathcal{U}(\tau,\mathbf{n}_0)$. So $\mathbf{v}_0\in\mathcal{U}(\tau,\mathbf{n}_0)\cap\mathcal{U}_w^{\perp}(\tau,\mathbf{n}_0)=\{\mathbf{0}\}$, and $\mathbf{v}_j\to\mathbf{0}$ weakly in $W^{1,2}(\Omega,\mathbb{R}^3)$ as $j\to\infty$.

To finish the proof of (2) of Lemma 4.3 in Case 2, note that we may choose a maximizing sequence in $\mathcal{U}_w^{\perp}(\tau,\mathbf{n}_0)$ that approaches $\mathbf{0}$ weakly as slowly as possible. So let us choose a maximizing sequence $\{\mathbf{v}^\varepsilon\}\subset\mathcal{U}_w^{\perp}(\tau,\mathbf{n}_0)$ such that $\|\mathbf{v}^\varepsilon\|_{C^1(\bar\Omega,\mathbb{R}^3)}=1$, and

$$\mathcal{I}[\mathbf{v}^\varepsilon]\gg\frac{1}{|\log\varepsilon|}\quad\text{as }\varepsilon\to 0.$$

Let

$$\rho^\varepsilon=\frac{\varepsilon a_\varepsilon^2|c^\varepsilon|^2\mu_w(\mathbf{n}_0)}{\mathcal{I}[\mathbf{v}^\varepsilon]}.$$

Then $\rho^\varepsilon\ll\varepsilon|\log\varepsilon|$ as $\varepsilon\to 0$, and

$$T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon]=-\varepsilon^2a_\varepsilon^4|c^\varepsilon|^4\left[\frac{\int_\Omega\nabla\zeta_{\mathbf{n}_0}\cdot\mathbf{v}^\varepsilon dx}{\sqrt{\mathcal{I}[\mathbf{v}^\varepsilon]}}\right]^2=-\varepsilon^2a_\varepsilon^4|c^\varepsilon|^4\mu_w(\mathbf{n}_0)^2+o(\varepsilon^2).$$

Hence (2) of Lemma 4.3 is true.

Proof of (3). Assume $\mu_w(\mathbf{n}_0)=+\infty$. Choose $\mathbf{v}_j\in\mathcal{U}_w^{\perp}(\tau,\mathbf{n}_0)$ such that $\|\mathbf{v}_j\|_{C^1(\bar\Omega,\mathbb{R}^3)}$ $=1$ and

$$\frac{\int_\Omega\nabla\zeta_{\mathbf{n}_0}\cdot\mathbf{v}_j dx}{\sqrt{\mathcal{I}[\mathbf{v}_j]}}\equiv M_j\to+\infty\quad\text{as }j\to\infty.$$

Fix $\varepsilon>0$ and let

$$\rho_j=\frac{\varepsilon a_\varepsilon^2|c^\varepsilon|^2M_j}{\mathcal{I}[\mathbf{v}_j]}.$$

Then

$$t_v(\varepsilon)\le t_v(\varepsilon,\rho_j)\le T_{\varepsilon,\rho_j}[\mathbf{v}_j]=-\varepsilon^2a_\varepsilon^4|c^\varepsilon|^4M_j^2\to-\infty\quad\text{as }j\to\infty.\qquad\square$$

Note that, when $\mu_w(\mathbf{n}_0)<+\infty$, the estimate on ρ^ε given in Lemma 4.3 (conclusion (2)) is not optimal. In fact, we do not exclude the existence of a sequence $\tilde\rho^\varepsilon$ such that $t_w(\varepsilon)=t_w(\varepsilon,\tilde\rho^\varepsilon)$ holds but $\tilde\rho^\varepsilon\not\to 0$ as $\varepsilon\to 0$.

Step 3. *Upper bound of* $T_v(\varepsilon)$. Now we can prove an upper bound estimate of $T_v(\varepsilon)$ using Lemma 4.3. Let us write

$$W(\mathbf{n}_0,b)=\left\{\mathbf{u}\in W^{1,2}(\Omega,\mathbb{R}^3):\ \mathbf{u}(x)\cdot\mathbf{n}_0(x)=-\frac{b}{2}|\mathbf{u}(x)|^2\text{ a.e. in }\Omega\right\},$$
$$W_1(\mathbf{n}_0,\mathbf{v}^\varepsilon,b)=\left\{\mathbf{w}\in W^{1,2}(\Omega,\mathbb{R}^3):\ \mathbf{w}(x)\cdot\mathbf{n}_0(x)=-\frac{1}{2}|\mathbf{v}^\varepsilon(x)+b\mathbf{w}(x)|^2\text{ on }\Omega\right\}.$$

Note that $\mathbf{v}^\varepsilon+\varepsilon\rho^\varepsilon\mathbf{w}\in W(\mathbf{n}_0,\varepsilon\rho^\varepsilon)$ if and only if $\mathbf{w}\in W_1(\mathbf{n}_0,\mathbf{v}^\varepsilon,\varepsilon\rho^\varepsilon)$.

Lemma 4.4. *Let $\tau > 0$ and $\mathbf{n}_0 \in \mathcal{C}_*(\tau)$. Assume $\mu_w(\mathbf{n}_0) < +\infty$. There exist $\rho^\varepsilon > 0$, $\mathbf{v}^\varepsilon \in W(\mathbf{n}_0) \cap C^1(\bar\Omega, \mathbb{R}^3)$, and $\mathbf{w}^\varepsilon \in W^{1,2}(\Omega, \mathbb{R}^3)$ such that*

$$\begin{aligned} &\mathbf{v}^\varepsilon + \varepsilon\rho^\varepsilon \mathbf{w}^\varepsilon \in W(\mathbf{n}_0, \varepsilon\rho^\varepsilon), \\ &\rho^\varepsilon \ll \varepsilon|\log\varepsilon|, \\ &\|\mathbf{v}^\varepsilon\|_{C^1(\bar\Omega,\mathbb{R}^3)} = 1, \\ &\|\mathbf{w}^\varepsilon\|_{W^{1,2}(\Omega,\mathbb{R}^3)} \le C; \end{aligned} \tag{4.24}$$

and as $\varepsilon \to 0$,

$$T_v(\varepsilon) \le T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon + \varepsilon\rho^\varepsilon\mathbf{w}^\varepsilon] = -\varepsilon^2 a_\varepsilon^4 |c^\varepsilon|^4 \mu_w(\mathbf{n}_0)^2 + o(\varepsilon^2). \tag{4.25}$$

Moreover, if $\mu_w(\mathbf{n}_0)$ is achieved, then we can choose ρ^ε such that $\rho^\varepsilon = O(\varepsilon)$ as $\varepsilon \to 0$.

Proof. From the proof of Lemma 4.3 (conclusion (2)) we see that for any $\varepsilon > 0$ there exist $\rho^\varepsilon > 0$ and $\mathbf{v}^\varepsilon \in W(\mathbf{n}_0) \cap C^1(\bar\Omega, \mathbb{R}^3)$ such that, as $\varepsilon \to 0$,

$$\begin{aligned} &\|\mathbf{v}^\varepsilon\|_{C^1(\bar\Omega,\mathbb{R}^3)} = 1, \\ &0 < \rho^\varepsilon \ll \varepsilon|\log\varepsilon|, \\ &\mathcal{I}[\mathbf{v}^\varepsilon] \gg \frac{1}{|\log\varepsilon|}, \\ &T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon] = t_w(\varepsilon) + o(\varepsilon^2) = -\varepsilon^2 a_\varepsilon^4 |c^\varepsilon|^4 \mu_w(\mathbf{n}_0)^2 + o(\varepsilon^2). \end{aligned} \tag{4.26}$$

Moreover, if $\mu_w(\mathbf{n}_0)$ is achieved, then ρ^ε can be chosen such that $\rho^\varepsilon = O(\varepsilon)$. We have

$$\begin{aligned} T_v(\varepsilon) &\le \inf_{\mathbf{u}\in V(\mathbf{n}_0,\varepsilon\rho^\varepsilon)} T_{\varepsilon,\rho^\varepsilon}[\mathbf{u}] \le \inf_{\mathbf{u}\in W(\mathbf{n}_0,\varepsilon\rho^\varepsilon)} T_{\varepsilon,\rho^\varepsilon}[\mathbf{u}] \\ &\le \inf_{\mathbf{w}\in W_1(\mathbf{n}_0,\mathbf{v}^\varepsilon,\varepsilon\rho^\varepsilon)} T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon + \varepsilon\rho^\varepsilon\mathbf{w}]. \end{aligned}$$

Let

$$\begin{aligned} &\mathbf{w}^\varepsilon(x) = -f^\varepsilon(x)\mathbf{n}_0(x), \\ &\text{where} \quad f^\varepsilon(x) = \frac{|\mathbf{v}^\varepsilon(x)|^2}{1 + \sqrt{1 - (\varepsilon\rho^\varepsilon)^2|\mathbf{v}^\varepsilon(x)|^2}}. \end{aligned} \tag{4.27}$$

It is direct to verify that $\mathbf{w}^\varepsilon \in W_1(\mathbf{n}_0, \mathbf{v}^\varepsilon, \varepsilon\rho^\varepsilon)$. Moreover, since $\|\mathbf{v}^\varepsilon\|_{C^1(\bar\Omega,\mathbb{R}^3)} = 1$, we can show that there exists a constant C such that for all small ε,

$$\|\mathbf{w}^\varepsilon\|_{W^{1,2}(\Omega,\mathbb{R}^3)} \le C. \tag{4.28}$$

Now we choose $\mathbf{v}^\varepsilon + \varepsilon\rho^\varepsilon\mathbf{w}^\varepsilon$ as a test function for $T_v(\varepsilon)$. We use (4.28) to compute

$$\begin{aligned} &T_v(\varepsilon) \le T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon + \varepsilon\rho^\varepsilon\mathbf{w}^\varepsilon] \\ &= T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon] + \varepsilon(\rho^\varepsilon)^3 \int_\Omega \{K_1 \text{div } \mathbf{v}^\varepsilon \text{div } \mathbf{w}^\varepsilon + K_2(\text{curl } \mathbf{v}^\varepsilon + \tau\mathbf{v}^\varepsilon) \cdot (\text{curl } \mathbf{w}^\varepsilon + \tau\mathbf{w}^\varepsilon)\} dx \\ &\quad + \varepsilon^2(\rho^\varepsilon)^4 \mathcal{I}[\mathbf{w}^\varepsilon] - 2\varepsilon^2(\rho^\varepsilon)^2 a_\varepsilon^2 |c^\varepsilon|^2 \int_\Omega (\nabla\zeta_{\mathbf{n}_0} - \mathbf{n}_0) \cdot \mathbf{w}^\varepsilon dx \\ &= T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon] + O\left(\varepsilon(\rho^\varepsilon)^3 \|\mathbf{v}^\varepsilon\|_{W^{1,2}(\Omega,\mathbb{R}^3)} \|\mathbf{w}^\varepsilon\|_{W^{1,2}(\Omega,\mathbb{R}^3)}\right) + O\big(\varepsilon^2(\rho^\varepsilon)^4 \|\mathbf{w}^\varepsilon\|^2_{W^{1,2}(\Omega,\mathbb{R}^3)}\big) \\ &\quad + O\left(\varepsilon^2(\rho^\varepsilon)^2 \|\mathbf{w}^\varepsilon\|_{W^{1,2}(\Omega,\mathbb{R}^3)}\right) \\ &= T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon] + O\left(\varepsilon(\rho^\varepsilon)^3\right) + O\left(\varepsilon^2(\rho^\varepsilon)^4\right) + O\left(\varepsilon^2(\rho^\varepsilon)^2\right) \\ &= T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon] + O\left(\varepsilon^4|\log\varepsilon|^3\right) \\ &= T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon] + o(\varepsilon^3). \end{aligned}$$

Inequality (4.25) follows from this and the last line in (4.26). $\square$

LEMMA 4.5. *Let $\tau > 0$ and $\mathbf{n}_0 \in \mathcal{C}_*(\tau)$. Assume that $\mu_w(\mathbf{n}_0) = +\infty$. Let δ_ε satisfy for any ε*

$$\varepsilon^2 \ll \delta_\varepsilon \le \delta_0,$$

where $0 < \delta_0 < 1$ is independent of ε. There exist $\rho^\varepsilon > 0$, $\mathbf{v}^\varepsilon \in W(\mathbf{n}_0) \cap C^1(\bar{\Omega}, \mathbb{R}^3)$, and $\mathbf{w}^\varepsilon \in W^{1,2}(\Omega, \mathbb{R}^3)$ such that

$$\begin{aligned} &\mathbf{v}^\varepsilon + \varepsilon\rho^\varepsilon \mathbf{w}^\varepsilon \in W(\mathbf{n}_0, \varepsilon\rho^\varepsilon),\\ &\rho_\varepsilon \le \frac{\delta_\varepsilon}{\varepsilon},\\ &\|\mathbf{v}^\varepsilon\|_{C^1(\bar{\Omega},\mathbb{R}^3)} = 1,\\ &\|\mathbf{w}^\varepsilon\|_{W^{1,2}(\Omega,\mathbb{R}^3)} \le C; \end{aligned} \tag{4.29}$$

and as $\varepsilon \to 0$,

$$T_v(\varepsilon) \le T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon + \varepsilon\rho^\varepsilon\mathbf{w}^\varepsilon] = -\varepsilon^2 a_\varepsilon^4 |c^\varepsilon|^4 L^\varepsilon + O\left(\frac{\delta_\varepsilon^3}{\varepsilon^2}\right), \tag{4.30}$$

where $L^\varepsilon \to +\infty$.

Proof. Let

$$W_\varepsilon(\mathbf{n}_0) = \left\{ \mathbf{v} \in \mathcal{U}_w^\perp(\tau, \mathbf{n}_0) \cap C^1(\bar{\Omega}, \mathbb{R}^3) : \ \frac{\int_\Omega \nabla\zeta_{\mathbf{n}_0} \cdot \mathbf{v}\,dx}{\mathcal{I}[\mathbf{v}]^{3/2}} \le \frac{\delta_\varepsilon}{a_\varepsilon^2 |c^\varepsilon|^2 \varepsilon^2} \right\},$$

and define

$$\mu_\varepsilon(\mathbf{n}_0) = \sup_{\mathbf{v} \in W_\varepsilon(\mathbf{n}_0)} \frac{\int_\Omega \nabla\zeta_{\mathbf{n}_0} \cdot \mathbf{v}\,dx}{\sqrt{\mathcal{I}[\mathbf{v}]}}. \tag{4.31}$$

We have

$$\mu_\varepsilon(\mathbf{n}_0) = \sup \left\{ \frac{\int_\Omega \nabla\zeta_{\mathbf{n}_0} \cdot \mathbf{v}\,dx}{\sqrt{\mathcal{I}[\mathbf{v}]}} : \ \mathbf{v} \in W_\varepsilon(\mathbf{n}_0), \ \|\mathbf{v}\|_{C^1(\bar{\Omega},\mathbb{R}^3)} = 1 \right\}.$$

We claim that, if $\mu_w(\mathbf{n}_0) = +\infty$, then

$$\lim_{\varepsilon \to 0} \mu_\varepsilon(\mathbf{n}_0) = +\infty. \tag{4.32}$$

To prove (4.32), let v_j and M_j be given in the proof of Lemma 4.3 (conclusion (3)). Note that, as $\varepsilon \to 0$, a_ε and $|c^\varepsilon|$ remain bounded, and $\delta_\varepsilon/\varepsilon^2 \to +\infty$. Let us choose $j(\varepsilon)$ such that $j(\varepsilon_2) \ge j(\varepsilon_1)$ if $0 < \varepsilon_2 < \varepsilon_1$, and

$$\frac{M_{j(\varepsilon)}}{\mathcal{I}[\mathbf{v}_{j(\varepsilon)}]} \le \frac{\delta_\varepsilon}{a_\varepsilon^2 |c^\varepsilon|^2 \varepsilon^2}.$$

Then $\mathbf{v}_{j(\varepsilon)} \in W_\varepsilon(\mathbf{n}_0)$. Hence

$$\mu_\varepsilon(\mathbf{n}_0) \ge \frac{\int_\Omega \nabla\zeta_{\mathbf{n}_0} \cdot \mathbf{v}_{j(\varepsilon)}\,dx}{\sqrt{\mathcal{I}[\mathbf{v}_{j(\varepsilon)}]}} = M_{j(\varepsilon)} \to +\infty \quad \text{as } \varepsilon \to 0.$$

So (4.32) is true.

Now we choose $\mathbf{v}^\varepsilon \in W_\varepsilon(\mathbf{n}_0) \cap C^1(\bar{\Omega}, \mathbb{R}^3)$ such that $\|\mathbf{v}^\varepsilon\|_{C^1(\bar{\Omega},\mathbb{R}^3)} = 1$ and

$$\frac{\int_\Omega \nabla\zeta_{\mathbf{n}_0} \cdot \mathbf{v}^\varepsilon dx}{\sqrt{\mathcal{I}[\mathbf{v}^\varepsilon]}} = \mu_\varepsilon(\mathbf{n}_0) + o\left(\frac{\varepsilon^2}{\mu_\varepsilon(\mathbf{n}_0)}\right).$$

Let

$$\rho^\varepsilon = \frac{\varepsilon a_\varepsilon^2 |c^\varepsilon|^2 \mu_\varepsilon(\mathbf{n}_0)}{\mathcal{I}[\mathbf{v}^\varepsilon]}.$$

Then $\varepsilon\rho^\varepsilon \le \delta_\varepsilon$, and

$$T_{\varepsilon,\rho_\varepsilon}[\mathbf{v}^\varepsilon] = -\varepsilon^2 a_\varepsilon^4 |c^\varepsilon|^4 \mu_\varepsilon(\mathbf{n}_0)^2 + o(\varepsilon^4).$$

For these $\mathbf{v}^\varepsilon$ and ρ^ε we define $\mathbf{w}^\varepsilon$ by (4.27). Then $\mathbf{w}^\varepsilon \in W_1(\mathbf{n}_0, \mathbf{v}^\varepsilon, \varepsilon\rho^\varepsilon)$ and satisfies (4.28). By computation as in the proof of Lemma 4.4 we have

$$T_v(\varepsilon) \le T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon + \varepsilon\rho^\varepsilon \mathbf{w}^\varepsilon] = T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon] + O(\varepsilon(\rho^\varepsilon)^3) + O(\varepsilon^2(\rho^\varepsilon)^4) + O(\varepsilon^2(\rho^\varepsilon)^2).$$

Since $\rho^\varepsilon \le \delta_\varepsilon/\varepsilon$ and $\delta_\varepsilon \gg \varepsilon^2$, the error terms in the above inequality can be controlled by

$$O(\delta_\varepsilon^2) + O\left(\frac{\delta_\varepsilon^3}{\varepsilon^2}\right) = O\left(\frac{\delta_\varepsilon^3}{\varepsilon^2}\right).$$

Estimate (4.30) follows from this and the last line in (4.29) if we set $L^\varepsilon = \mu_\varepsilon(\mathbf{n}_0)^2$. $\square$

Note that, in Lemma 4.5 we do not have control on the growth rate of L^ε, but we have control on ρ^ε by $\rho^\varepsilon \ll \delta_\varepsilon/\varepsilon$, which is needed in order to get (4.28). In order to make (4.30) useful, we choose δ_ε such that

$$\varepsilon^2 \ll \delta_\varepsilon \ll \varepsilon^2|\log\varepsilon|.$$

Hence $\rho^\varepsilon \ll \varepsilon|\log\varepsilon|$.

Now we return to the construction of the test functions for $\mathcal{G}_\varepsilon$. Let

$$\mu_*(\tau) = \sup_{\mathbf{n}_0 \in \mathcal{C}_*(\tau)} \mu_w(\mathbf{n}_0). \tag{4.33}$$

Step 4. *Constructing test functions for $\mathcal{G}_\varepsilon$ when $\mu_*(\tau) < +\infty$.* When $\mu_w(\mathbf{n}_0) < +\infty$, from Lemma 4.4, we can choose $\mathbf{v}^\varepsilon$, $\mathbf{w}^\varepsilon$, and ρ^ε satisfying (4.24) such that (4.25) holds. In particular, $\rho^\varepsilon \ll \varepsilon|\log\varepsilon|$. Moreover, $\rho^\varepsilon = O(\varepsilon)$ if $\mu_w(\mathbf{n}_0)$ is achieved. We set

$$\mathbf{u}^\varepsilon = \mathbf{v}^\varepsilon + \varepsilon\rho^\varepsilon \mathbf{w}^\varepsilon, \qquad \rho = \rho^\varepsilon, \qquad \varphi^\varepsilon = \varphi_1^\varepsilon + \frac{ia_\varepsilon}{\rho^\varepsilon} f^\varepsilon, \tag{4.34}$$

where φ_1^ε and f_ε are real-valued functions bounded in $W^{1,2}(\Omega)$ as $\varepsilon \to 0$, and

$$\int_\Omega \varphi_1^\varepsilon dx = 0, \qquad \int_\Omega f^\varepsilon dx = 0.$$

From (4.16) we have

(4.35)

$$
\begin{aligned}
&\varepsilon^{-2}\mathcal{G}_\varepsilon[\psi^\varepsilon, \mathbf{n}^\varepsilon] - d_0(\varepsilon) - T_{\varepsilon,\rho^\varepsilon}[\mathbf{u}^\varepsilon] \\
&= \varepsilon^2 a_\varepsilon^4 |c^\varepsilon|^2 \int_\Omega \{|\nabla f^\varepsilon|^2 - 2\mathbf{n}_0 \cdot (\zeta_{\mathbf{n}_0}\nabla f^\varepsilon - f^\varepsilon \nabla\zeta_{\mathbf{n}_0})\}dx \\
&\quad + \varepsilon^2 a_\varepsilon^2 |c^\varepsilon|^2 (a_\varepsilon^2 - 1 + |c^\varepsilon|^2) \int_\Omega \{|\zeta_{\mathbf{n}_0} + \varepsilon\rho^\varepsilon \varphi_1^\varepsilon|^2 + \varepsilon^2 a_\varepsilon^2 |f^\varepsilon|^2\}dx \\
&\quad + (\varepsilon\rho^\varepsilon)^2 a_\varepsilon^2 |c^\varepsilon|^2 \int_\Omega |\nabla\varphi_1^\varepsilon - \mathbf{u}^\varepsilon|^2 dx \\
&\quad - 2\varepsilon^3\rho^\varepsilon a_\varepsilon^4 |c^\varepsilon|^2 \int_\Omega \{\nabla f^\varepsilon \cdot (\zeta_{\mathbf{n}_0}\mathbf{u}^\varepsilon + \varphi_1^\varepsilon \mathbf{n}^\varepsilon) - f^\varepsilon[\mathbf{u}^\varepsilon(\nabla\zeta_{\mathbf{n}_0} - \mathbf{n}_0) + \mathbf{n}^\varepsilon(\nabla\varphi_1^\varepsilon - \mathbf{u}^\varepsilon)]\}dx \\
&\quad + \frac{1}{2}\varepsilon^4 a_\varepsilon^4 |c^\varepsilon|^4 \int_\Omega [2f^\varepsilon - |\zeta_{\mathbf{n}_0} + \varepsilon\rho^\varepsilon\varphi_1^\varepsilon|^2 - \varepsilon^2 a_\varepsilon^2 |f^\varepsilon|^2]^2 dx \\
&= -\varepsilon^2 a_\varepsilon^2 |c^\varepsilon|^2 (1 - |c^\varepsilon|^2) \int_\Omega |\zeta_{\mathbf{n}_0}|^2 dx \\
&\quad + \varepsilon^2 a_\varepsilon^4 |c^\varepsilon|^2 \int_\Omega \{|\nabla f^\varepsilon - \zeta_{\mathbf{n}_0}\mathbf{n}_0|^2 + 2(\mathbf{n}_0 \cdot \nabla\zeta_{\mathbf{n}_0}) f^\varepsilon\}dx \\
&\quad + (\varepsilon\rho^\varepsilon)^2 a_\varepsilon^2 |c^\varepsilon|^2 \int_\Omega |\nabla\varphi_1^\varepsilon - \mathbf{u}^\varepsilon|^2 dx \\
&\quad + O(\varepsilon^3\rho^\varepsilon) + O(\varepsilon^4),
\end{aligned}
$$

where we have used the fact that $\mathbf{v}^\varepsilon$ and $\mathbf{w}^\varepsilon$ are bounded in $W^{1,2}(\Omega, \mathbb{R}^3)$, φ_1^ε and f^ε are bounded in $W^{1,2}(\Omega)$, and $\rho^\varepsilon \ll \varepsilon|\log\varepsilon|$. Write

$$F_{\mathbf{n}_0}[f] = \int_\Omega \{|\nabla f - \zeta_{\mathbf{n}_0}\mathbf{n}_0|^2 + 2(\mathbf{n}_0 \cdot \nabla\zeta_{\mathbf{n}_0}) f\}dx.$$

Checking the right side of (4.35), we choose f^ε to be the minimizer of the variational problem

$$\min \left\{ F_{\mathbf{n}_0}[f] : \ f \in W^{1,2}(\Omega), \ \int_\Omega f dx = 0 \right\}. \tag{4.36}$$

It is easy to verify that (4.36) has a unique minimizer, which is denoted by $f_{\mathbf{n}_0}$; and for any $g \in W^{1,2}(\Omega)$ satisfying $\int_\Omega g dx = 0$ we have

$$\int_\Omega \{(\nabla f_{\mathbf{n}_0} - \zeta_{\mathbf{n}_0}\mathbf{n}_0) \cdot \nabla g + (\mathbf{n}_0 \cdot \nabla\zeta_{\mathbf{n}_0}) g\}dx = 0. \tag{4.37}$$

Thus $f_{\mathbf{n}_0}$ is the solution of the following equation:

$$\begin{cases} \Delta f_{\mathbf{n}_0} = 2\mathbf{n}_0 \cdot \nabla\zeta_{\mathbf{n}_0} - l(\mathbf{n}_0) & \text{in } \Omega, \\ \dfrac{\partial f_{\mathbf{n}_0}}{\partial\nu} = \zeta_{\mathbf{n}_0}\mathbf{n}_0 \cdot \nu & \text{on } \partial\Omega, \\ \displaystyle\int_\Omega f_{\mathbf{n}_0} dx = 0, \end{cases} \tag{4.38}$$

where

$$l(\mathbf{n}_0) = \fint_\Omega |\nabla\zeta_{\mathbf{n}_0}|^2 dx = \fint_\Omega \mathbf{n}_0 \cdot \nabla\zeta_{\mathbf{n}_0} dx.$$

Using (4.37) with $g = f_{\mathbf{n}_0}$ we find

$$F_{n_0}[f_{\mathbf{n}_0}] = \int_\Omega \{|\nabla f_{\mathbf{n}_0} - \zeta_{\mathbf{n}_0}\mathbf{n}_0|^2 - 2(\nabla f_{\mathbf{n}_0} - \zeta_{\mathbf{n}_0}\mathbf{n}_0)\cdot\nabla f_{\mathbf{n}_0}\}dx = \int_\Omega (|\zeta_{\mathbf{n}_0}|^2 - |\nabla f_{\mathbf{n}_0}|^2)dx. \tag{4.39}$$

From the above discussion and (4.10), (4.14), and (4.34), the test functions for $\mathcal{G}_\varepsilon$ are chosen to be

$$\begin{aligned} &\psi^\varepsilon = c^\varepsilon\{1 + i\varepsilon a_\varepsilon\zeta_{\mathbf{n}_0} - \varepsilon^2 a_\varepsilon^2 f_{\mathbf{n}_0} + i\varepsilon^2\rho^\varepsilon a_\varepsilon \varphi_1^\varepsilon\},\\ &\mathbf{n}^\varepsilon = \mathbf{n}_0 + \varepsilon\rho^\varepsilon\mathbf{u}^\varepsilon,\\ &\text{where}\quad \mathbf{n}_0\in\mathcal{C}_*(\tau),\quad \mathbf{u}^\varepsilon = \mathbf{v}^\varepsilon + \varepsilon\rho^\varepsilon\mathbf{w}^\varepsilon,\quad \varphi_1^\varepsilon = \zeta_{\mathbf{u}^\varepsilon}, \end{aligned} \tag{4.40}$$

where $\zeta_{\mathbf{u}^\varepsilon}$ is the solution of (1.7) for $\mathbf{u} = \mathbf{u}^\varepsilon$. Now we explain how φ_1^ε is chosen.

Case 1. ρ^ε satisfies

$$\varepsilon \ll \rho^\varepsilon \ll \varepsilon|\log\varepsilon| \quad \text{as } \varepsilon\to 0. \tag{4.41}$$

Checking the term of order $O((\varepsilon\rho^\varepsilon)^2)$ in the right-hand side of (4.35), we naturally choose $\varphi_1^\varepsilon = \zeta_{\mathbf{u}^\varepsilon}$. Then from (4.35) and (4.39) we find

$$\begin{aligned} &\varepsilon^{-2}\mathcal{G}_\varepsilon[\psi^\varepsilon,\mathbf{n}^\varepsilon] - d_0(\varepsilon) - T_{\varepsilon,\rho^\varepsilon}[\mathbf{u}^\varepsilon]\\ &= -\varepsilon^2 a_\varepsilon^2|c^\varepsilon|^2(1-|c^\varepsilon|^2)\int_\Omega|\zeta_{\mathbf{n}_0}|^2dx + \varepsilon^2 a_\varepsilon^4|c^\varepsilon|^2\int_\Omega(|\zeta_{\mathbf{n}_0}|^2 - |\nabla f_{\mathbf{n}_0}|^2)dx\\ &\quad + (\varepsilon\rho^\varepsilon)^2 a_\varepsilon^2|c^\varepsilon|^2\omega(\mathbf{u}^\varepsilon)|\Omega| + O(\varepsilon^3\rho^\varepsilon)\\ &= \varepsilon^2 a_\varepsilon^2|c^\varepsilon|^2\left\{(a_\varepsilon^2 - 1 + |c^\varepsilon|^2)\|\zeta_{\mathbf{n}_0}\|^2_{L^2(\Omega)} - a_\varepsilon^2\|\nabla f_{\mathbf{n}_0}\|^2_{L^2(\Omega)}\right\}\\ &\quad + (\varepsilon\rho^\varepsilon)^2 a_\varepsilon^2|c^\varepsilon|^2\omega(\mathbf{u}^\varepsilon)|\Omega| + O(\varepsilon^3\rho^\varepsilon) + O(\varepsilon^4). \end{aligned}$$

From (4.14), (4.25), and (4.40) we have, under the condition (4.41),

$$\begin{aligned} &\varepsilon^{-2}\mathcal{G}_\varepsilon[\psi^\varepsilon,\mathbf{n}^\varepsilon]\\ &= d_0(\varepsilon) + T_{\varepsilon,\rho^\varepsilon}[\mathbf{u}^\varepsilon] + \varepsilon^2 a_\varepsilon^2|c^\varepsilon|^2\left\{(a_\varepsilon^2 - 1 + |c^\varepsilon|^2)\|\zeta_{\mathbf{n}_0}\|^2_{L^2(\Omega)} - a_\varepsilon^2\|\nabla f_{\mathbf{n}_0}\|^2_{L^2(\Omega)}\right\}\\ &\quad + (\varepsilon\rho^\varepsilon)^2 a_\varepsilon^2|c^\varepsilon|^2\omega(\mathbf{u}^\varepsilon)|\Omega| + O(\varepsilon^3\rho^\varepsilon)\\ &= d_0(\varepsilon) - \varepsilon^2[a_\varepsilon^4(1 - a_\varepsilon^2\omega_*(\tau))^2\mu_w(\mathbf{n}_0)^2 + d_1(\varepsilon,\mathbf{n}_0)] + o(\varepsilon^2), \end{aligned}$$

where

$$d_1(\varepsilon,\mathbf{n}_0) = a_\varepsilon^4(1 - a_\varepsilon^2\omega_*(\tau))\left\{\|\nabla f_{\mathbf{n}_0}\|^2_{L^2(\Omega)} - (1-\omega_*(\tau))\|\zeta_{\mathbf{n}_0}\|^2_{L^2(\Omega)}\right\}. \tag{4.42}$$

Thus, under the condition (4.41) we have

$$\mathcal{G}_\varepsilon[\psi^\varepsilon,\mathbf{n}^\varepsilon] = d_0(\varepsilon)\varepsilon^2 - d_1(\varepsilon)\varepsilon^4 + o(\varepsilon^4). \tag{4.43}$$

Case 2. $\rho^\varepsilon = O(\varepsilon)$. In this case, (4.43) remains valid if we also choose $\varphi_1^\varepsilon = \zeta_{\mathbf{u}^\varepsilon}$. However, we mention that, if we wish to get a better estimate in the higher order terms, we may compute all the terms of order $O(\varepsilon^4)$ in the right-hand side of (4.35) to determine a better choice of φ_1^ε.

Summarizing the above discussions we have the following proposition.

Proposition 4.6. *Assume $\mu_*(\tau) < +\infty$. We have the energy upper bound estimate*

$$C(K_1, K_2, \varepsilon, \tau, a_\varepsilon\varepsilon) \le d_0(\varepsilon)\varepsilon^2 - \varepsilon^4[a_\varepsilon^4(1 - a_\varepsilon^2\omega_*(\tau))^2\mu_*(\tau)^2 + d_1(\varepsilon)] + o(\varepsilon^4), \tag{4.44}$$

where

$$d_1(\varepsilon) = \inf_{\mathbf{n}_0 \in \mathcal{C}_*(\tau)} d_1(\varepsilon, \mathbf{n}_0). \tag{4.45}$$

Step 5. *Constructing test functions for $\mathcal{G}_\varepsilon$ when $\mu_*(\tau) = +\infty$.* When $\mu_w(\mathbf{n}_0) = +\infty$, we use Lemma 4.5 for $\delta_\varepsilon = \varepsilon^2|\log\varepsilon|$ to find $\mathbf{v}^\varepsilon$, $\mathbf{w}^\varepsilon$, and ρ^ε satisfying (4.29) such that (4.30) holds, that is,

$$T_{\varepsilon,\rho^\varepsilon}[\mathbf{v}^\varepsilon + \varepsilon\rho^\varepsilon\mathbf{w}^\varepsilon] = -\varepsilon^2 a_\varepsilon^4|c^\varepsilon|^4 L^\varepsilon + O(\varepsilon^4|\log\varepsilon|^3),$$

where $L^\varepsilon \to +\infty$. Then we define the test functions ψ^ε and $\mathbf{n}^\varepsilon$ by (4.40) with $\varphi_1^\varepsilon = \zeta_{\mathbf{u}^\varepsilon}$. Similar to the computations in Step 5, we have

$$\begin{aligned}
&\varepsilon^{-2}\mathcal{G}_\varepsilon[\psi^\varepsilon, \mathbf{n}^\varepsilon] \\
&= d_0(\varepsilon) + T_{\varepsilon,\rho^\varepsilon}[\mathbf{u}^\varepsilon] + \varepsilon^2 a_\varepsilon^2|c^\varepsilon|^2\left\{(a_\varepsilon^2 - 1 + |c^\varepsilon|^2)\|\zeta_{\mathbf{n}_0}\|^2_{L^2(\Omega)} - a_\varepsilon^2\|\nabla f_{\mathbf{n}_0}\|^2_{L^2(\Omega)}\right\} \\
&\quad + (\varepsilon\rho^\varepsilon)^2 a_\varepsilon^2|c^\varepsilon|^2\omega(\mathbf{u}^\varepsilon)|\Omega| + O(\varepsilon^3\rho^\varepsilon) \\
&\le d_0(\varepsilon) - \varepsilon^2 L^\varepsilon a_\varepsilon^4(1 - a_\varepsilon^2\omega_*(\tau))^2 - \varepsilon^2 d_1(\varepsilon, \mathbf{n}_0) + o(\varepsilon^2).
\end{aligned}$$

So we have the following proposition.

Proposition 4.7. *Assume $\mu_*(\tau) = +\infty$. Then we have the energy upper bound estimate*

$$C(K_1, K_2, \varepsilon, \tau, a_\varepsilon\varepsilon) \le d_0(\varepsilon)\varepsilon^2 - \varepsilon^4 L^\varepsilon a_\varepsilon^4(1 - a_\varepsilon^2\omega_*(\tau))^2 - \varepsilon^4 d_{1,\infty}(\varepsilon) + o(\varepsilon^4), \tag{4.46}$$

where

$$d_{1,\infty}(\varepsilon) = \inf_{\mathbf{n}_0 \in \mathcal{C}_*(\tau),\ \mu_w(\mathbf{n}_0) = +\infty} d_1(\varepsilon, \mathbf{n}_0). \tag{4.47}$$

Proof of Theorem 4.2. The proof follows from Propositions 4.6 and 4.7. ∎

5. Asymptotics of minimizers for small . In this section we investigate the behavior of the minimizers for small κ. We always assume the condition (1.2), and let ε denote κ. In the following we write the minimizers of $\mathcal{G}$ by $(\psi_\varepsilon, \mathbf{n}_\varepsilon)$. Note that they depend also on the wave number q.

Theorem 5.1. *Assume* (1.2) *and* (4.1) *hold. For small $\varepsilon > 0$, let $q = a_\varepsilon\varepsilon$, where*

$$\begin{gathered}
0 < a_\varepsilon < \frac{1}{\varepsilon}Q_{c_3}(K_1, K_2, \varepsilon, \tau), \\
\lim_{\varepsilon \to 0^+} a_\varepsilon = a_0 \le \frac{1}{\sqrt{\omega_*(\tau)}}.
\end{gathered} \tag{5.1}$$

We have the following:

(1) $C(K_1, K_2, \varepsilon, \tau, a_\varepsilon\varepsilon) = m(a_0)\varepsilon^2 + o(\varepsilon^2)$, *where*

$$m(a_0) = \begin{cases} \frac{1}{2}|\Omega| & \text{if } a_0 = \frac{1}{\sqrt{\omega_*(\tau)}}, \\ a_0^2\omega_*(\tau)[1 - \frac{1}{2}a_0^2\omega_*(\tau)]|\Omega| & \text{if } 0 < a_0 < \frac{1}{\sqrt{\omega_*(\tau)}}. \end{cases}$$

(2) *Let* $(\psi_\varepsilon, \mathbf{n}_\varepsilon)$ *be the minimizers of* $\mathcal{G}$ *for* $q = a_\varepsilon \varepsilon$. *Then there exist a subsequence and* $\mathbf{n}_0 \in \mathcal{C}_*(\tau)$ *such that*

$$\mathbf{n}_\varepsilon \to \mathbf{n}_0 \text{ weakly in } W^{1,2}_{loc}(\Omega, \mathbb{R}^3) \text{ and strongly in } L^p(\Omega, \mathbb{R}^3) \text{ for all } 1 \le p < \infty. \tag{5.2}$$

Moreover, the following expansions hold:

$$\begin{aligned} \psi_\varepsilon &= c_\varepsilon[1 + i a_\varepsilon \varepsilon(\zeta_{\mathbf{n}_0} + \varepsilon \rho_\varepsilon \varphi_\varepsilon)], \\ \mathbf{n}_\varepsilon &= \mathbf{n}_0 + \varepsilon \rho_\varepsilon \mathbf{u}_\varepsilon, \end{aligned} \tag{5.3}$$

where, as $\varepsilon \to 0$,

(i) $\rho_\varepsilon \ge 0$ *and* $\varepsilon \rho_\varepsilon \to 0$;

(ii) $\|\mathbf{u}_\varepsilon\|_{L^2(\Omega)} = 1$ *and* $\rho_\varepsilon^2 \mathcal{I}[\mathbf{u}_\varepsilon] \to 0$;

(iii) $\int_\Omega \varphi_\varepsilon dx = 0$ *and* $\varepsilon \rho_\varepsilon \varphi_\varepsilon \to 0$ *strongly in* $W^{1,2}(\Omega, \mathbb{C})$;

(iv) c_ε *is a complex number and*

$$|c_\varepsilon| \to \sqrt{[1 - a_0^2 \omega_*(\tau)]_+}. \tag{5.4}$$

Proof. The upper bound of energy follows from Theorem 4.2. Now we prove the lower bound. Let $q = a_\varepsilon \varepsilon$, where a_ε satisfies (5.1). From (4.9) we find

$$\begin{aligned} &\int_\Omega \Big\{ |\nabla_{a_\varepsilon \varepsilon \mathbf{n}_\varepsilon} \psi_\varepsilon|^2 + \frac{\varepsilon^2}{2}(1 - |\psi_\varepsilon|^2)^2 + K_1 |\text{div } \mathbf{n}_\varepsilon|^2 + K_2 |\text{curl } \mathbf{n}_\varepsilon + \tau \mathbf{n}_\varepsilon|^2 \Big\} dx \\ &\le d_0(\varepsilon)\varepsilon^2 + O(\varepsilon^4). \end{aligned} \tag{5.5}$$

As in the proof of Theorem 4.1 (Step 2) we can show that there exists a subsequence, still denoted by $(\psi_\varepsilon, \mathbf{n}_\varepsilon)$, such that (5.2) holds, where $\mathbf{n}_0 \in \mathcal{C}(\tau)$. From (5.5),

$$\int_\Omega |\nabla_{a_\varepsilon \varepsilon \mathbf{n}_\varepsilon} \psi_\varepsilon|^2 dx = O(\varepsilon^2).$$

Since $|\psi_\varepsilon(x)| \le 1$, we have

$$\|\nabla \psi_\varepsilon\|_{L^2(\Omega)} \le \|\nabla_{a_\varepsilon \varepsilon \mathbf{n}_\varepsilon} \psi_\varepsilon\|_{L^2(\Omega)} + a_\varepsilon \varepsilon \|\mathbf{n}_\varepsilon \psi_\varepsilon\|_{L^2(\Omega)} = O(\varepsilon)$$

and

$$\psi_\varepsilon \to c_0 \quad \text{strongly in } W^{1,2}(\Omega, \mathbb{C}) \text{ as } \varepsilon \to 0,$$

where c_0 is a constant. Let

$$c_\varepsilon = \fint_\Omega \psi_\varepsilon dx.$$

Then $c_\varepsilon \to c_0$ as $\varepsilon \to 0$.

CLAIM 1. $c_\varepsilon \ne 0$ *for all small* ε.

Suppose Claim 1 were not true. Passing to a subsequence if necessary, we have $c_\varepsilon = 0$ for all small ε. Then $\psi_\varepsilon \to 0$ in $W^{1,2}(\Omega, \mathbb{C})$ as $\varepsilon \to 0$. We write

$$\psi_\varepsilon = \varepsilon b_\varepsilon \phi_\varepsilon,$$

where $\int_\Omega \phi_\varepsilon dx = 0$, and $\|\phi_\varepsilon\|_{L^2(\Omega)} = 1$, $b_\varepsilon > 0$, and $\varepsilon b_\varepsilon \to 0$. Thus

$$\|\nabla \phi_\varepsilon\|^2_{L^2(\Omega)} \ge \mu_1 \|\phi_\varepsilon\|^2_{L^2(\Omega)} = \mu_1,$$

where

$$\mu_1 = \inf\left\{\|\nabla\phi\|^2_{L^2(\Omega)} :\ \phi \in W^{1,2}(\Omega),\ \|\phi\|_{L^2(\Omega)} = 1,\ \int_\Omega \phi dx = 0\right\}.$$

We have

$$\begin{aligned}
\varepsilon^{-2}\mathcal{G}_\varepsilon[\psi_\varepsilon, \mathbf{n}_\varepsilon] &\geq \frac{1}{\varepsilon^2}\int_\Omega \left\{|\nabla_{a_\varepsilon\varepsilon\mathbf{n}_\varepsilon}\psi_\varepsilon|^2 + \frac{1}{2}(1-\varepsilon^2 b_\varepsilon^2|\phi_\varepsilon|^2)^2\right\} dx \\
&= \frac{|\Omega|}{2} + b_\varepsilon^2 \int_\Omega |\nabla_{a_\varepsilon\varepsilon\mathbf{n}_\varepsilon}\phi_\varepsilon|^2 dx + O(\varepsilon^2 b_\varepsilon^2) \\
&= \frac{|\Omega|}{2} + b_\varepsilon^2 \left\{\int_\Omega |\nabla\phi_\varepsilon|^2 dx + O(\varepsilon)\right\} \\
&\geq \frac{|\Omega|}{2} + b_\varepsilon^2(\mu_1 + O(\varepsilon)) \\
&> \frac{|\Omega|}{2},
\end{aligned}$$

which contradicts the obvious upper bound $\mathcal{G}_\varepsilon[\psi_\varepsilon, \mathbf{n}_\varepsilon] < \frac{|\Omega|}{2}$. Thus Claim 1 is true.

Now we write

$$\psi_\varepsilon = c_\varepsilon(1 + ia_\varepsilon\varepsilon\phi_\varepsilon), \qquad \mathbf{n}_\varepsilon = \mathbf{n}_0 + \varepsilon\rho_\varepsilon\mathbf{u}_\varepsilon,$$

where $\int_\Omega \phi_\varepsilon dx = 0$ and $\|\mathbf{u}_\varepsilon\|_{L^2(\Omega)} = 1$. Since $c_\varepsilon \to c_0$, $\varepsilon\phi_\varepsilon \to 0$ in $W^{1,2}(\Omega, \mathbb{C})$, and $\mathbf{n}_\varepsilon \to \mathbf{n}_0$ in $L^4(\Omega, \mathbb{R}^3)$ as $\varepsilon \to 0$, we find

(5.6)

$$\begin{aligned}
\varepsilon^{-2}\mathcal{G}_\varepsilon[\psi_\varepsilon, \mathbf{n}_\varepsilon] = \int_\Omega \Bigg\{ & a_\varepsilon^2|c_\varepsilon|^2|\nabla\phi_\varepsilon - \mathbf{n}_\varepsilon - ia_\varepsilon\varepsilon\phi_\varepsilon\mathbf{n}_\varepsilon|^2 + \frac{1}{2}(1-|c_\varepsilon|^2|1+ia_\varepsilon\varepsilon\phi_\varepsilon|^2)^2 \\
&+ \rho_\varepsilon^2 K_1|\mathrm{div}\ \mathbf{u}_\varepsilon|^2 + \rho_\varepsilon^2 K_2|\mathrm{curl}\ \mathbf{u}_\varepsilon + \tau\mathbf{u}_\varepsilon|^2\Bigg\} dx \\
= \int_\Omega & \left\{a_0^2|c_\varepsilon|^2|\nabla\phi_\varepsilon - \mathbf{n}_0|^2 + \frac{1}{2}(1-|c_\varepsilon|^2)^2\right\} dx + \rho_\varepsilon^2\mathcal{I}[\mathbf{u}_\varepsilon] + o(1).
\end{aligned}$$

Let $f_{a_0,\mathbf{n}_0}(t)$ be the function defined in (4.7). From (2.1), (5.5), and (5.6), we have

$$\begin{aligned}
d_0(\varepsilon) - \rho_\varepsilon^2\mathcal{I}[\mathbf{u}_\varepsilon] + O(\varepsilon^2) &\geq \int_\Omega \left\{a_0^2|c_\varepsilon|^2|\nabla\phi_\varepsilon - \mathbf{n}_0|^2 + \frac{1}{2}(1-|c_\varepsilon|^2)^2\right\} dx + o(1) \\
&\geq \left\{a_0^2|c_\varepsilon|^2\omega(\mathbf{n}_0) + \frac{1}{2}(1-|c_\varepsilon|^2)^2\right\}|\Omega| + o(1) \\
&= \left\{a_0^2|c_0|^2\omega(\mathbf{n}_0) + \frac{1}{2}(1-|c_0|^2)^2\right\}|\Omega| + o(1) \\
&= f_{a_0,\mathbf{n}_0}(|c_0|)|\Omega| + o(1) \\
&\geq |\Omega| \min_{\mathbf{n}\in\mathcal{C}(\tau)} \min_{c\geq 0} f_{a_0,\mathbf{n}}(c) + o(1) \\
&= m(a_0) + o(1).
\end{aligned}$$

Note that $d_0(\varepsilon) = m_0(a_0) + o(1)$. Thus

$$
\begin{aligned}
&f_{a_0,\mathbf{n}_0}(|c_0|) = \frac{m(a_0)}{|\Omega|} = \min_{c\geq 0} f_{a_0,\mathbf{n}_0}(c) = f_{a_0,\mathbf{n}_0}\Big(\sqrt{[1-a_0^2\omega(\mathbf{n}_0)]_+}\Big),\\
&\omega(\mathbf{n}_0) = \omega_*(\tau),\\
&\lim_{\varepsilon\to 0}|c_\varepsilon| = |c_0| = \sqrt{[1-a_0^2\omega(\mathbf{n}_0)]_+},\\
&⨍_\Omega |\nabla\phi_\varepsilon - \mathbf{n}_0|^2 dx = \omega(\mathbf{n}_0) + o(1);
\end{aligned}
\tag{5.7}
$$

and

$$
\begin{aligned}
&\mathcal{G}_\varepsilon[\psi_\varepsilon, \mathbf{n}_\varepsilon] = m(a_0)\varepsilon^2 + o(\varepsilon^2),\\
&\rho_\varepsilon^2 \mathcal{I}[\mathbf{u}_\varepsilon] = o(1).
\end{aligned}
$$

From the last equality in (5.7) we find that, as $\varepsilon \to 0$,

$$\nabla\phi_\varepsilon = \nabla\zeta_{\mathbf{n}_0} + o(1) \quad \text{in } L^2(\Omega).$$

Since $\int_\Omega \phi_\varepsilon dx = 0$, we have

$$\phi_\varepsilon = \zeta_{\mathbf{n}_0} + o(1) \quad \text{in } W^{1,2}(\Omega).$$

Summarizing the above discussions, we get the conclusions of Theorem 5.1. ☐

Proof of Theorem 1.1. The proof follows from Theorems 4.1 and 5.1.

Appendix A. Proof of Lemma 2.3. We follow the idea of Evans in the proof of the div-curl lemma; see [E, p. 54]. Without loss of generality we assume $D_2 = \Omega$ and $\mathbf{u} \in V(\Omega, \mathbb{R}^3) \cap C^2(\bar\Omega, \mathbb{R}^3)$. Let D denote D_1. Let $\mathbf{w}$ be the solution of

$$\Delta\mathbf{w} = \mathbf{u} \quad \text{in } \Omega, \quad \mathbf{w} = \mathbf{0} \quad \text{on } \partial\Omega. \tag{A.1}$$

There exists a constant $C_1(\Omega)$ such that

$$\|\mathbf{w}\|_{W^{2,2}(\Omega,\mathbb{R}^3)} \leq C_1(\Omega)\|\mathbf{u}\|_{L^2(\Omega)}.$$

Let $\xi = \text{div } \mathbf{w}$ and $\mathbf{v} = \mathbf{u} - \nabla\xi = \Delta\mathbf{w} - \nabla\text{div } \mathbf{w} = -\text{curl }^2\mathbf{w}$. Then

$$\mathbf{u} = \mathbf{v} + \nabla\xi.$$

From (A.1), we have

$$\Delta(\text{curl } \mathbf{w}) = \text{curl } \mathbf{u} \quad \text{in } \Omega.$$

Applying the interior elliptic estimate we have

$$
\begin{aligned}
\|\text{curl } \mathbf{w}\|_{W^{2,2}(D,\mathbb{R}^3)} &\leq C_2(D,\Omega)\{\|\text{curl } \mathbf{w}\|_{W^{1,2}(\Omega,\mathbb{R}^3)} + \|\text{curl } \mathbf{u}\|_{L^2(\Omega)}\}\\
&\leq C_2(D,\Omega)\{\|\mathbf{w}\|_{W^{2,2}(\Omega,\mathbb{R}^3)} + \|\text{curl } \mathbf{u}\|_{L^2(\Omega)}\}\\
&\leq C_3(D,\Omega)\{\|\mathbf{u}\|_{L^2(\Omega)} + \|\text{curl } \mathbf{u}\|_{L^2(\Omega)}\}.
\end{aligned}
$$

Thus

$$
\begin{aligned}
\|\mathbf{v}\|_{W^{1,2}(D,\mathbb{R}^3)} = \|\text{curl}^2\mathbf{w}\|_{W^{1,2}(D,\mathbb{R}^3)} &\leq C_4\|\text{curl } \mathbf{w}\|_{W^{2,2}(D,\mathbb{R}^3)}\\
&\leq C_5(D,\Omega)\{\|\mathbf{u}\|_{L^2(\Omega)} + \|\text{curl } \mathbf{u}\|_{L^2(\Omega)}\}.
\end{aligned}
$$

Using (A.1) again, we have

$$\Delta(\text{div } \mathbf{w}) = \text{div } \mathbf{u} \quad \text{in } \Omega.$$

Applying the interior elliptic estimate again, we have

$$\begin{aligned}\|\text{div } \mathbf{w}\|_{W^{2,2}(D)} &\le C_6(D,\Omega)\{\|\text{div } \mathbf{w}\|_{W^{1,2}(\Omega)} + \|\text{div } \mathbf{u}\|_{L^2(\Omega)}\} \\ &\le C_7(D,\Omega)\{\|\mathbf{w}\|_{W^{2,2}(\Omega,\mathbb{R}^3)} + \|\text{div } \mathbf{u}\|_{L^2(\Omega)}\} \\ &\le C_8(D,\Omega)\{\|\mathbf{u}\|_{L^2(\Omega)} + \|\text{curl } \mathbf{u}\|_{L^2(\Omega)}\}.\end{aligned}$$

Hence

$$\begin{aligned}\|\nabla\xi\|_{W^{1,2}(D)} = \|\nabla(\text{div } \mathbf{w})\|_{W^{1,2}(D)} &\le \|\text{div } \mathbf{w}\|_{W^{2,2}(D)} \\ &\le C_8(D,\Omega)\{\|\mathbf{u}\|_{L^2(\Omega)} + \|\text{curl } \mathbf{u}\|_{L^2(\Omega)}\}.\end{aligned}$$

Now we have

$$\begin{aligned}\|\mathbf{u}\|_{W^{1,2}(D,\mathbb{R}^3)} &\le \|\mathbf{v}\|_{W^{1,2}(D,\mathbb{R}^3)} + \|\nabla\xi\|_{W^{1,2}(D)} \\ &\le C_9(D,\Omega)\{\|\mathbf{u}\|_{L^2(\Omega)} + \|\text{div } \mathbf{u}\|_{L^2(\Omega)} + \|\text{curl } \mathbf{u}\|_{L^2(\Omega)}\}. \qquad \square\end{aligned}$$

REFERENCES

[ADN] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions,* I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[AG] P. Aviles and Y. Giga, *A mathematical problem related to the physical theory of liquid crystal configurations*, in Miniconference on Geometry and Partial Differential Equations, Proc. Centre Math. Anal. Austral. Nat. Univ. 12, 1987, pp. 1–16.

[B] H. Brezis, *Liquid crystals and energy estimates for S^2-valued maps*, in Theory and Applications of Liquid Crystals, IMA Vol. Math. Appl. 5, J. L. Ericksen and D. Kinderlehrer, eds., Springer-Verlag, New York, 1987, pp. 31–52.

[BCLP] P. Bauman, M. Calderer, C. Liu, and D. Phillips, *The phase transition between chiral nematic and smectic A^* liquid crystals*, Arch. Ration. Mech. Anal., 165 (2002), pp. 161–186.

[C] M. C. Calderer, *Studies of layering and chirality of smectic A^* liquid crystals*, Math. Comput. Modelling, 34 (2001), pp. 1273–1288.

[ChL] J.-H. Chen and T. Lubensky, *Landau–Ginzburg mean-field theory for the nematic to smectic-C and nematic to smectic-A phase transitions*, Phys. Rev. A, 14 (1976), pp. 1202–1207.

[dG] P. G. de Gennes, *An analogy between superconductors and smectics A*, Solid State Comm., 10 (1972), pp. 753–756.

[dGP] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, 2nd ed., Oxford Science Publications, Oxford, UK, 1993.

[DL] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 3, Springer-Verlag, New York, 1990.

[E] L. C. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, Regional Conference Series in Mathematics 74, AMS, Providence, RI, 1990.

[GL] V. Ginzburg and L. Landau, *On the theory of superconductivity*, Zh. Eksper. Teoret. Fiz., 20 (1950), pp. 1064–1082 (in Russian); English translation in L. D. Landau, Collected Papers, Gordon and Breach, New York, 1967, pp. 546–568.

[GT] D. Gilberg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, New York, 1983.

[HK] R. Hardt and D. Kinderlehrer, *Mathematical questions of liquid crystal theory*, in Theory and Applications of Liquid Crystals, IMA Vol. Math. Appl. 5, J. L. Ericksen and D. Kinderlehrer, eds., Springer-Verlag, New York, 1987, pp. 151–184.

[HKL] R. Hardt, D. Kinderlehrer, and F. H. Lin, *Existence and partial regularity of static liquid crystal configurations*, Comm. Math. Phys., 105 (1986), pp. 547–570.

[L] F. H. Lin, *Nonlinear theory of defects in nematic liquid crystals, phase transitions and flow phenomena*, Comm. Pure Appl. Math., 42 (1989), pp. 789–814.

[LM] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Springer, Berlin, 1972.

[LR] T. Lubensky and S. Renn, *Twist-grain-boundary phases near the nematic–smectic-A–smectic-A–smectic-C point in liquid crystals*, Phys. Rev. A, 41 (1990), pp. 4392–4401.

[P1] X. B. Pan, *Landau–de Gennes model of liquid crystals and critical wave number*, Comm. Math. Phys., 239 (2003), pp. 343–382.

[P2] X. B. Pan, *Superconductivity near critical temperature*, J. Math. Phys., 44 (2003), pp. 2639–2678.

[RL] S. Renn and T. Lubensky, *Abrikosov dislocation lattice in a model of the cholesteric–to–smectic-A transition*, Phys. Rev. A, 38 (1988), pp. 2132–2147.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1649–1656

ON THE REGULARITY CONDITIONS FOR THE DISSIPATIVE QUASI-GEOSTROPHIC EQUATIONS*

DONGHO CHAE†

Abstract. We obtain regularity conditions for solutions of the dissipative quasi-geostrophic equation. The first one imposes on the integrability of the magnitude of the temperature gradient, and corresponds to the Serrin type of condition in the theory of Navier–Stokes equations. The other one incorporates the direction of normals to the level curves and the magnitude of the temperature gradient simultaneously. For the proof of the second result, in particular, we use geometric properties of the nonlinear term as well as the estimates using the Triebel–Lizorkin type of norms.

Key words. quasi-geostrophic equations, regularity conditions, Triebel–Lizorkin spaces

AMS subject classifications. 35Q35, 76B03, 76D03, 76D09

DOI. 10.1137/040616954

1. Introduction.

We are concerned with the regularity of the quasi-geostrophic equation with a dissipation term:

$$\frac{\partial\theta}{\partial t} + (v\cdot\nabla)\theta = -\kappa\Lambda^\alpha\theta, \tag{1.1}$$

$$v(x,t) = -\nabla^\perp(-\Delta)^{-\frac{1}{2}}\theta = -\int_{\mathbb{R}^2}\frac{\nabla^\perp\theta(x+y,t)}{|y|}dy, \tag{1.2}$$

$$\theta(x,0)=\theta_0(x), \tag{1.3}$$

where $\theta(x,t)$ is a scalar function representing temperature, $v(x,t)$ is the velocity field of the fluid, $\kappa\geq 0$ is the diffusion constant, $\Lambda^\alpha=(-\Delta)^{\frac{\alpha}{2}}$, and $\nabla^\perp=(-\partial_{x_2},\partial_{x_1})$. See, e.g., [5, 3, 10] for the instructive discussions and the physical and mathematical motivations of the study of (1.1)–(1.3), in particular of the inviscid case ($\kappa=0$). For $\alpha>1$ the global regularity of the solution of (1.1)–(1.3) is well known (see [6]). On the other hand, for $0\leq\alpha\leq 1$, the question of global regularity/finite time singularity is still a challenging open problem (see, e.g., [2, 4, 7, 8, 9, 15, 16, 17] for related studies). In particular, the critical dissipation case ($\alpha=1$) has similar features to the three-dimensional (3D) Navier–Stokes equations, and could be considered as its model problem. In order to see the similarities to the 3D Navier–Stokes equations (with fractional powers of Laplacian) more apparently we apply the operation $\nabla^\perp$ to (1.1) to obtain

$$\frac{\partial\nabla^\perp\theta}{\partial t} + (v\cdot\nabla)\nabla^\perp\theta = (\nabla^\perp\theta\cdot\nabla)v - \kappa\Lambda^\alpha\nabla^\perp\theta. \tag{1.4}$$

*Received by the editors October 13, 2004; accepted for publication June 1, 2005; published electronically January 27, 2006. This work was supported by Korea Research Foundation grant KRF-2002-015-CS0003.
http://www.siam.org/journals/sima/37-5/61695.html

†Department of Mathematics, Sungkyunkwan University, Suwon 440-746, Korea (chae@skku.edu). Current address: Center for Scientific Computation and Mathematical Modeling, CSIC Bldg. No. 406, Paint Branch Drive, University of Maryland, College Park, MD 20742-3289 (dchae@cscamm.umd.edu).

Then, we observe that $\nabla^\perp\theta$ has the role of vorticity, and (1.2) corresponds to the Biot–Savart law for the 3D Navier–Stokes equations. In this note we are concerned with the sufficient conditions to guarantee regularity of solutions of the quasi-geostrophic equations, which have been studied by many authors for the case of the 3D Navier–Stokes equations since Prodi [11] and Serrin [12]. To the author's knowledge the only regularity condition available in the literature for (1.1)–(1.3) is the following one obtained by Constantin, Majda, and Tabak [5] (except its refinement in [1]):

$$\limsup_{t\nearrow T}\|\theta(t)\|_{H^m}<\infty \quad \text{if and only if} \quad \int_0^T\|\nabla^\perp\theta(t)\|_{L^\infty}dt<\infty, \tag{1.5}$$

where $m>2$, which holds for solutions of both viscous and inviscid ($\kappa=0$) equations. Our first theorem generalizes this as follows.

THEOREM 1.1. *Let $\theta(x,t)$ be a solution of the quasi-geostrophic equation* (1.1)–(1.3) *with $\alpha\in(0,1]$, $\kappa>0$, and its derivative $\nabla^\perp\theta$ satisfies*

$$\nabla^\perp\theta\in L^r(0,T;L^p(\mathbb{R}^2)) \quad \textit{for some } p,r \textit{ with} \quad \frac{2}{p}+\frac{\alpha}{r}\le\alpha,\quad \frac{2}{\alpha}<p<\infty; \tag{1.6}$$

then there is no singularity up to T.

Remark 1.1. We observe that $p=\infty$, $r=1$ corresponds to (1.5) for any $\alpha\in(0,1]$.

Remark 1.2. We note that the system (1.1)–(1.2) has symmetry under the scaling transform

$$\theta(x,t)\to\theta^\lambda(x,t)=\lambda^{\alpha-1}\theta(\lambda x,\lambda^\alpha t).$$

Under this scaling transform we have the invariance of the norms,

$$\|\nabla^\perp\theta\|_{L^r(0,T;L^p(\mathbb{R}^2))}=\|\nabla^\perp\theta^\lambda\|_{L^r(0,\lambda^\alpha T;L^p(\mathbb{R}^2))} \quad \text{if} \quad \frac{2}{p}+\frac{\alpha}{r}=\alpha.$$

In this sense the condition (1.6) is optimal.

For the statement of our next theorem we introduce a function space. Given $0<\alpha<1$, $1\le p\le\infty$, $1\le q\le\infty$, the function space $\dot{\mathcal{F}}^s_{p,q}$ is defined by the seminorm

$$\|f\|_{\dot{\mathcal{F}}^s_{p,q}}=\begin{cases}\left\|\left(\int_{\mathbb{R}^n}\frac{|f(x+y)-f(x)|^q}{|y|^{n+sq}}dy\right)^{\frac{1}{q}}\right\|_{L^p(\mathbb{R}^n,dx)} & \text{if } 1\le p\le\infty,\ 1\le q<\infty,\\ \left\|ess\sup_{|y|\neq0}\frac{|f(x+y)-f(x)|}{|y|^s}\right\|_{L^p(\mathbb{R}^n,dx)} & \text{if } 1\le p\le\infty,\ q=\infty.\end{cases}$$

Observe that, in particular, $\dot{\mathcal{F}}^s_{\infty,\infty}\cong C^s$, the usual Hölder seminormed space. In order to compare this space with other more classical function spaces let us introduce the Banach space $\mathcal{F}^s_{p,q}$ by defining its norm,

$$\|f\|_{\mathcal{F}^s_{p,q}}=\|f\|_{L^p}+\|f\|_{\dot{\mathcal{F}}^s_{p,q}}.$$

We note that for $0<s<1$, $2\le p<\infty$, and $q=2$, $\mathcal{F}^s_{p,2}\cong L^p_s(\mathbb{R}^n)=(1-\Delta)^{-\frac{s}{2}}L^p(\mathbb{R}^n)$, the fractional order Sobolev space (or the Bessel potential space); see [13, p. 163]. On the other hand, if $\frac{n}{\min\{p,q\}} < s < 1$, $n < p < \infty$, and $n < q \le \infty$, then $\mathcal{F}^s_{p,q}$ coincides with the usual Triebel–Lizorkin space $F^s_{p,q}$ (see [14, p. 101]). The following is our second main theorem.

THEOREM 1.2. *Let $\theta(x,t)$ be a solution of the dissipative quasi-geostrophic equation* (1.1)–(1.3) *with $\kappa > 0$. Let $\xi(x,t)$ be its direction field, $\xi(x,t) = \nabla^\perp\theta(x)/|\nabla^\perp\theta(x)|$ defined for $\nabla^\perp\theta(x,t) \neq 0$. Suppose there exist $s \in (0,1)$, $q \in (\frac{2}{2-s}, \infty]$, $p_1 \in (1,\infty]$, $p_2 \in (1, \frac{2}{s})$ satisfying $\frac{s}{2} < \frac{1}{p_1} + \frac{1}{p_2} < \frac{s+\alpha}{2}$, and $r_1, r_2 \in [1,\infty]$ such that the following hold:*

$$\xi(x,t) \in L^{r_1}(0,T;\dot{\mathcal{F}}^s_{p_1,q}) \quad \textit{and} \quad \nabla^\perp\theta(x,t) \in L^{r_2}(0,T;L^{p_2}(\mathbb{R}^2))$$

$$\textit{with} \qquad \frac{2}{p_1} + \frac{2}{p_2} + \frac{\alpha}{r_1} + \frac{\alpha}{r_2} \le \alpha + s. \tag{1.7}$$

Then, the solution $\theta(x,t)$ is regular up to T.

Remark 1.3. The condition (1.7) describes quantitatively that by assuming regularity of the direction field $\xi(x,t)$ we can have regularity of solution with assumption of lower integrability of $|\nabla^\perp\theta(x,t)|$ than required by Theorem 1.1.

2. Proof of the main theorems.

Proof of Theorem 1.1. We plan to show that our integrability condition for $\nabla^\perp\theta$ in Theorem 1.1 implies

$$\int_0^T \|\nabla^\perp\theta(t)\|_{L^\infty} dt < \infty, \tag{2.1}$$

thus guaranteeing the desired regularity until T by (1.5). We take $L^2(\mathbb{R}^2)$ inner product of (1.4) with $\Delta\nabla^\perp\theta$ and integrate it by parts to obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\Lambda^2\theta(t)\|^2_{L^2} + \kappa\|\Lambda^{2+\frac{\alpha}{2}}\theta\|^2_{L^2} &= \int_{\mathbb{R}^2}(v\cdot\nabla)\nabla^\perp\theta\cdot\Delta\nabla^\perp\theta dx\\ &\quad - \int_{\mathbb{R}^2}(\nabla^\perp\theta\cdot\nabla)v\cdot\Delta\nabla^\perp\theta dx = I + J.\end{aligned} \tag{2.2}$$

Integrating by parts, we have

$$\begin{aligned} I &= -\int_{\mathbb{R}^2}\nabla[(v\cdot\nabla)\nabla^\perp\theta]\cdot\nabla\nabla^\perp\theta dx\\ &= -\int_{\mathbb{R}^2}(\nabla v)(\nabla\nabla^\perp\theta)\cdot\nabla\nabla^\perp\theta dx - \int_{\mathbb{R}^2}(v\cdot\nabla)\nabla\nabla^\perp\theta\cdot\nabla\nabla^\perp\theta dx\\ &= I_1 + I_2.\end{aligned}$$

Integrating by parts again, and using the fact that $\operatorname{div} v = 0$, we derive

$$I_2 = -\frac{1}{2}\int_{\mathbb{R}^2}(v\cdot\nabla)|\nabla\nabla^\perp\theta|^2 dx = \frac{1}{2}\int_{\mathbb{R}^2}(\operatorname{div} v)|\nabla\nabla^\perp\theta|^2 dx = 0. \tag{2.3}$$

In the case $\frac{4}{\alpha} \le p < \infty$ we estimate

$$\begin{aligned} I_1 &\le \int_{\mathbb{R}^2}|\nabla v||\nabla\nabla^\perp\theta|^2 dx\\ &\le \|\nabla\nabla^\perp\theta\|_{L^2}\|\nabla v\|_{L^p}\|\nabla\nabla^\perp\theta\|_{L^{\frac{2p}{p-2}}} \quad \text{(Hölder's inequality)}\end{aligned}$$

$$
\begin{aligned}
&\leq C\|\nabla\nabla^\perp\theta\|_{L^2}\|\nabla v\|_{L^p}\|\nabla\nabla^\perp\theta\|_{L^2}^{1-\frac{4}{\alpha p}}\\
&\quad\times\|\Lambda^{\frac{\alpha}{2}}\nabla\nabla^\perp\theta\|_{L^2}^{\frac{4}{\alpha p}} \quad \text{(Gagliardo–Nirenberg's inequality)}\\
&\leq C\|\Lambda^2\theta\|_{L^2}^{2-\frac{4}{\alpha}}\|\nabla^\perp\theta\|_{L^p}\|\Lambda^{2+\frac{\alpha}{2}}\theta\|_{L^2}^{\frac{4}{\alpha}} \quad \text{(Calderon–Zygmund's inequality)}\\
&\leq \frac{\kappa}{4}\|\Lambda^{2+\frac{\alpha}{2}}\theta\|_{L^2}^2 + C\|\nabla^\perp\theta\|_{L^p}^{\frac{\alpha p}{\alpha p-2}}\|\Lambda^2\theta\|_{L^2}^2 \quad \text{(Young's inequality)}, \qquad (2.4)
\end{aligned}
$$

while in the case $\frac{2}{\alpha} < p < \frac{4}{\alpha}$, we estimate

$$
\begin{aligned}
I_1 &\leq \|\nabla\nabla^\perp\theta\|_{L^{\frac{4}{2-\alpha}}}\|\nabla v\|_{L^p}\|\nabla\nabla^\perp\theta\|_{L^{\frac{4p}{2p-4+\alpha p}}} \quad \text{(Hölder's inequality)}\\
&\leq C\|\Lambda^{\frac{\alpha}{2}}\nabla\nabla^\perp\theta\|_{L^2}\|\nabla v\|_{L^p}\|\nabla\nabla^\perp\theta\|_{L^2}^{2-\frac{4}{\alpha p}}\|\Lambda^{\frac{\alpha}{2}}\nabla\nabla^\perp\theta\|_{L^2}^{\frac{4}{\alpha p}-1}\\
&\quad \text{(Sobolev's and Gagliardo–Nirenberg's inequalities)}\\
&\leq C\|\Lambda^2\theta\|_{L^2}^{2-\frac{4}{\alpha}}\|\nabla^\perp\theta\|_{L^p}\|\Lambda^{2+\frac{\alpha}{2}}\theta\|_{L^2}^{\frac{4}{\alpha}} \quad \text{(Calderon–Zygmund's inequality)}\\
&\leq \frac{\kappa}{4}\|\Lambda^{2+\frac{\alpha}{2}}\theta\|_{L^2}^2 + C\|\nabla^\perp\theta\|_{L^p}^{\frac{\alpha p}{\alpha p-2}}\|\Lambda^2\theta\|_{L^2}^2 \quad \text{(Young's inequality)}. \qquad (2.5)
\end{aligned}
$$

In order to estimate J we first integrate by parts:

$$
\begin{aligned}
J &= \int_{\mathbb{R}^2} \nabla[(\nabla^\perp\theta\cdot\nabla)v]\cdot\nabla\nabla^\perp\theta dx\\
&= \int_{\mathbb{R}^2} \nabla\nabla^\perp\theta\cdot\nabla v\cdot\nabla\nabla^\perp\theta dx + \int_{\mathbb{R}^2}(\nabla^\perp\theta\cdot\nabla)\nabla v\cdot\nabla\nabla^\perp\theta dx.
\end{aligned}
$$

Since $\|\nabla\nabla v\|_{L^q} \leq C\|\nabla\nabla^\perp\theta\|_{L^q}$, $1<q<\infty$, due to the Calderon–Zygmund inequality, we observe that the estimate of J is the same as the estimate of I_1, and we have

$$
J \leq \frac{\kappa}{4}\|\Lambda^{2+\frac{\alpha}{2}}\theta\|_{L^2}^2 + C\|\nabla^\perp\theta\|_{L^p}^{\frac{\alpha p}{\alpha p-2}}\|\Lambda^2\theta\|_{L^2}^2. \tag{2.6}
$$

Combining the estimates (2.3)–(2.6) and absorbing the diffusion term into the left-hand side, we obtain

$$
\frac{d}{dt}\|\Lambda^2\theta\|_{L^2}^2 + \kappa\|\Lambda^{2+\frac{\alpha}{2}}\theta\|_{L^2}^2 \leq C\|\nabla^\perp\theta\|_{L^p}^{\frac{\alpha p}{\alpha p-2}}\|\Lambda^2\theta\|_{L^2}^2. \tag{2.7}
$$

By Gronwall's lemma,

$$
\|\Lambda^2\theta(t)\|_{L^2} \leq \|\Lambda^2\theta_0\|_{L^2}\exp\left[C\int_0^T \|\nabla^\perp\theta(t)\|_{L^p}^{\frac{\alpha p}{\alpha p-2}}dt\right] \quad \forall t\in[0,T].
$$

Hence, $\|\Lambda^2\theta(t)\|_{L^2} \in L^\infty(0,T)$. Integrating (2.7) over $[0,T]$, we have

$$
\begin{aligned}
&\|\Lambda^2\theta(t)\|_{L^2}^2 + \kappa\int_0^T \|\Lambda^{2+\frac{\alpha}{2}}\theta(t)\|_{L^2}^2 dt\\
&\qquad \leq C\int_0^T \|\nabla^\perp\theta(t)\|_{L^p}^{\frac{\alpha p}{\alpha p-2}}dt \sup_{0\leq t\leq T}\|\Lambda^2\theta(t)\|_{L^2}^2 + \|\Lambda^2\theta_0\|_{L^2}^2,
\end{aligned}
$$

which implies $\int_0^T \|\Lambda^{2+\frac{\alpha}{2}}\theta(t)\|_{L^2}^2 dt < \infty$. Applying the Gagliardo–Nirenberg inequality,

$$
\|\nabla f\|_{L^\infty} \leq C\|f\|_{L^2}^{\frac{\alpha}{4+\alpha}}\|\Lambda^{2+\frac{\alpha}{2}}f\|_{L^2}^{\frac{4}{4+\alpha}},
$$

in $\mathbb{R}^2$, we have

$$\int_0^T \|\nabla\theta(t)\|_{L^\infty} dt \le C \int_0^T \|\theta(t)\|_{L^2}^{\frac{\alpha}{4+\alpha}} \|\Lambda^{2+\frac{\alpha}{2}}\theta(t)\|_{L^2}^{\frac{4}{4+\alpha}} dt$$
$$\le C\|\theta_0\|_{L^2}^{\frac{\alpha}{4+\alpha}} T^{\frac{2+\alpha}{4+\alpha}} \left(\int_0^T \|\Lambda^{2+\frac{\alpha}{2}}\theta\|_{L^2}^2 dt\right)^{\frac{2}{4+\alpha}} < \infty.$$

Hence, (2.1) is proved. ◻

Proof of Theorem 1.2. Let p be an integer of the form $p = 2^m$, where m is a positive integer, and satisfy

$$\frac{2}{\alpha} \le p < \infty. \tag{2.8}$$

Taking $L^2(\mathbb{R}^2)$ inner product of (1.4) with $\nabla^\perp\theta(x,t)|\nabla^\perp\theta(x,t)|^{p-2}$ and substituting v from (1.2) into it, we have after integration by parts

$$\begin{aligned}
&\frac{1}{p}\frac{d}{dt}\|\nabla^\perp\theta(t)\|_{L^p}^p + \kappa \int_{\mathbb{R}^2} (\Lambda^\alpha \nabla^\perp\theta)\cdot\nabla^\perp\theta|\nabla^\perp\theta|^{p-2}dx\\
&= \int_{\mathbb{R}^2} (\nabla^\perp\theta\cdot\nabla)v\cdot\nabla^\perp\theta|\nabla^\perp\theta|^{p-2}dx\\
&= \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} [\nabla\theta(x,t)\cdot\hat{y}][\nabla^\perp\theta(x+y,t)\cdot\nabla\theta(x,t)]\frac{dy}{|y|^2}|\nabla^\perp\theta(x,t)|^{p-2}dx\\
&:= I,
\end{aligned}\tag{2.9}$$

where the integral with respect to y in the right-hand side is in the sense of principal value. We start estimating the dissipation term

$$\begin{aligned}
\kappa \int_{\mathbb{R}^2} (\nabla^\perp\theta\cdot\nabla)v\cdot\nabla^\perp\theta|\nabla^\perp\theta|^{p-2}dx &\ge \frac{\kappa}{p}\int_{\mathbb{R}^2} \left|\Lambda^{\frac{\alpha}{2}}|\nabla^\perp\theta|^{\frac{p}{2}}\right|^2 dx\\
&\ge \frac{\kappa C_\alpha}{p}\left(\int_{\mathbb{R}^2} |\nabla^\perp\theta|^{\frac{2p}{2-\alpha}}dx\right)^{\frac{2-a}{2}} = \frac{\kappa C_\alpha}{p}\|\nabla^\perp\theta\|_{L^{\frac{2p}{2-\alpha}}}^p,
\end{aligned}\tag{2.10}$$

where we used Lemma 2.4 of [8] for the estimate of the fractional derivative in the first inequality, and the Sobolev imbedding, $L^2_{\frac{\alpha}{2}}(\mathbb{R}^2) \hookrightarrow L^{\frac{4}{2-\alpha}}(\mathbb{R}^2)$, in the second inequality. Next, we estimate I as follows:

$$\begin{aligned}
I &= \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} (\xi^\perp(x,t)\cdot\hat{y})[\xi(x+y,t)\cdot\xi^\perp(x,t)]|\nabla^\perp\theta(x+y,t)|\frac{dy}{|y|^2}|\nabla^\perp\theta(x,t)|^p dx\\
&= \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} (\xi^\perp(x,t)\cdot\hat{y})[\xi(x+y,t)-\xi(x,t)]\cdot\xi^\perp(x,t)|\nabla^\perp\theta(x+y,t)|\frac{dy}{|y|^2}|\nabla^\perp\theta(x,t)|^p dx\\
&\le \int_{\mathbb{R}^2}\int_{\mathbb{R}^2} |\xi(x+y,t)-\xi(x,t)||\nabla^\perp\theta(x+y,t)|\frac{dy}{|y|^2}|\nabla^\perp\theta(x,t)|^p dx\\
&\le \int_{\mathbb{R}^2}\left(\int_{\mathbb{R}^2}\frac{|\xi(x+y,t)-\xi(x,t)|^q}{|y|^{2+sq}}dy\right)^{\frac{1}{q}}\left(\int_{\mathbb{R}^2}\frac{|\nabla^\perp\theta(x+y,t)|^{q'}}{|y|^{2-sq'}}dy\right)^{\frac{1}{q'}}|\nabla^\perp\theta|^p dx\\
&\le \|\xi\|_{\dot{\mathcal{F}}^s_{p_1,q}}\left\|\{I_{\alpha q'}(|\nabla^\perp\theta|^{q'})\}^{\frac{1}{q'}}\right\|_{L^{\tilde{p}_2}}\|\nabla^\perp\theta\|_{L^{p_3}}^p,
\end{aligned}\tag{2.11}$$

where we used the fact that $\xi(x,t)\cdot\xi^\perp(x,t)=0$ in the second equality, and Hölder's inequality in the second and the third inequalities with the exponents satisfying

$$\frac{1}{p_1}+\frac{1}{\tilde p_2}+\frac{p}{p_3}=1,\qquad \frac{1}{q}+\frac{1}{q'}=1, \tag{2.12}$$

and $I_\sigma(\cdot)$, $0<\sigma<2$, is the operator defined by the Riesz potential as follows:

$$I_\sigma(f)(x)=\gamma(\sigma)\int_{\mathbb{R}^2}\frac{f(x+y)}{|y|^{2-\sigma}}dy,\quad \gamma(\sigma)=2^\sigma\pi\frac{\Gamma(\frac{\sigma}{2})}{\Gamma(\frac{2-\sigma}{2})}.$$

From the well-defined property of the Riesz operator we have the restriction $0<sq'<2$, which gives us $q\in(\frac{2}{2-s},\infty]$ due to the second equation of (2.12). Using the Hardy–Littlewood–Sobolev inequality [13], we estimate

$$\begin{aligned}\left\|\{I_{sq'}(|\nabla^\perp\theta|^{q'})\}^{\frac{1}{q'}}\right\|_{L^{p_2}}&=\|I_{sq'}(|\nabla^\perp\theta|^{q'})\|^{\frac{1}{q'}}_{L^{\frac{\tilde p_2}{q'}}}\\&\le C\||\nabla^\perp\theta|^{q'}\|^{\frac{1}{q'}}_{L^r}=C\|\nabla^\perp\theta\|_{L^{rq'}}=C\|\nabla^\perp\theta\|_{L^{\frac{2\tilde p_2}{2+s\tilde p_2}}},\end{aligned}\tag{2.13}$$

where we used the relation $\frac{1}{r}=\frac{q'}{\tilde p_2}+\frac{sq'}{2}$, and hence $rq'=\frac{2\tilde p_2}{2+s\tilde p_2}$. On the other hand, using the L^p interpolation inequality, we estimate

$$\begin{aligned}\|\nabla^\perp\theta\|^p_{L^{p_3}}&\le\|\nabla^\perp\theta\|^{p(1+\frac{2p}{\alpha p_3}-\frac{2}{\alpha})}_{L^p}\|\nabla^\perp\theta\|^{p(\frac{2}{\alpha}-\frac{2p}{\alpha p_3})}_{L^{\frac{2p}{2-\alpha}}}\\&=\|\nabla^\perp\theta\|^{p-\frac{2p}{\alpha p_1}-\frac{2p}{\alpha\tilde p_2}}_{L^p}\|\nabla^\perp\theta\|^{\frac{2p}{\alpha p_1}+\frac{2p}{\alpha\tilde p_2}}_{L^{\frac{2p}{2-\alpha}}}.\end{aligned}\tag{2.14}$$

Note that use of the interpolation inequality in (2.14) requires that $p<p_3<\frac{2p}{2-\alpha}$, which, in turn, gives us the condition

$$0<\frac{1}{p_1}+\frac{1}{\tilde p_2}<\frac{\alpha}{2}\tag{2.15}$$

due to the first equation of (2.12). Combining (2.11) with (2.13) and (2.14), we derive

$$\begin{aligned}I&\le C\|\xi\|_{\dot{\mathcal{F}}^s_{p_1,q}}\|\nabla^\perp\theta\|_{L^{\frac{2\tilde p_2}{2+s\tilde p_2}}}\|\nabla^\perp\theta\|^{p-\frac{2p}{\alpha p_1}-\frac{2p}{\alpha\tilde p_2}}_{L^p}\|\nabla^\perp\theta\|^{\frac{2p}{\alpha p_1}+\frac{2p}{\alpha\tilde p_2}}_{L^{\frac{2p}{2-\alpha}}}\\&\le C\left(\|\xi\|_{\dot{\mathcal{F}}^s_{p_1,q}}\|\nabla^\perp\theta\|_{L^{\frac{2\tilde p_2}{2+s\tilde p_2}}}\right)^{\frac{\alpha p_1\tilde p_2}{\alpha p_1\tilde p_2-2p_1-2\tilde p_2}}\|\nabla^\perp\theta\|^p_{L^p}+\frac{\kappa C_\alpha}{2}\|\nabla^\perp\theta\|^p_{L^{\frac{2p}{2-\alpha}}},\end{aligned}\tag{2.16}$$

where we used Young's inequality, $ab\le\frac{a^u}{u}+\frac{b^{u'}}{u'}$, with

$$a=\|\xi\|_{\dot{\mathcal{F}}^s_{p_1,q}}\|\nabla^\perp\theta\|_{L^{\frac{2\tilde p_2}{2+s\tilde p_2}}}\|\nabla^\perp\theta\|^{p-\frac{2p}{\alpha p_1}-\frac{2p}{\alpha\tilde p_2}}_{L^p},\quad b=\|\nabla^\perp\theta\|^{\frac{2p}{\alpha p_1}+\frac{2p}{\alpha\tilde p_2}}_{L^{\frac{2p}{2-\alpha}}},$$

and

$$u=\frac{\alpha p_1\tilde p_2}{2(p_1+\tilde p_2)},\quad u'=\frac{\alpha p_1\tilde p_2}{\alpha p_1\tilde p_2-2(p_1+\tilde p_2)}.$$

Setting $\frac{2\tilde p_2}{2+s\tilde p_2}=p_2$, we have $\tilde p_2=\frac{2p_2}{2-sp_2}$. We observe here that there is the restriction $p_2<\frac{2}{s}$ due to positiveness of $\tilde p_2$. Substituting this value of $\tilde p$ into (2.16), we obtain

$$I\le C\|\xi\|^Q_{\dot{\mathcal{F}}^s_{p_1,q}}\|\nabla^\perp\theta\|^Q_{L^{p_2}}\|\nabla^\perp\theta\|^p_{L^p}+\frac{\kappa C_\alpha}{2}\|\nabla^\perp\theta\|^p_{L^{\frac{2p}{2-\alpha}}},\tag{2.17}$$

where we set

$$Q = \frac{\alpha p_1 p_2}{(\alpha + s)p_1 p_2 - 2p_1 - 2p_2}.$$

We note that the restriction (2.15) becomes

$$\frac{s}{2} < \frac{1}{p_1} + \frac{1}{p_2} < \frac{s}{2} + \frac{\alpha}{2} \tag{2.18}$$

in terms of p_1, p_2. We substitute the estimate (2.17) into (2.9), and combine this with (2.10). Then, after absorbing $\frac{\kappa C_\alpha}{2}\|\nabla^\perp\theta\|^p_{L^{\frac{2p}{2-\alpha}}}$ to the left-hand side, we obtain

$$\begin{aligned} &\frac{d}{dt}\|\nabla^\perp\theta(t)\|^p_{L^p} + \frac{\kappa C_\alpha}{2}\|\nabla^\perp\theta(t)\|^p_{L^{\frac{2p}{2-\alpha}}} \\ &\quad \le C\|\xi(t)\|^Q_{\dot{\mathcal{F}}^s_{p_1,q}}\|\nabla^\perp\theta(t)\|^Q_{L^{p_2}}\|\nabla^\perp\theta(t)\|^p_{L^p}. \end{aligned} \tag{2.19}$$

Now the inequality condition (1.7) becomes

$$\frac{1}{r_1} + \frac{1}{r_2} \le \frac{1}{Q}$$

in terms of Q. Thanks to the Gronwall lemma and Hölder's inequality we estimate

$$\begin{aligned} &\|\nabla^\perp\theta(t)\|_{L^p} \le \|\nabla^\perp\theta_0\|_{L^p}\exp\left[C\int_0^T \|\xi(t)\|^Q_{\dot{\mathcal{F}}^s_{p_1,q}}\|\nabla^\perp\theta(t)\|^Q_{L^{p_2}}\,dt\right] \\ &\le \|\nabla^\perp\theta_0\|_{L^p}\exp\left[C\left(\int_0^T \|\xi(t)\|^{r_1}_{\dot{\mathcal{F}}^s_{p_1,q}}\,dt\right)^{\frac{Q}{r_1}}\left(\int_0^T \|\nabla^\perp\theta(t)\|^{r_2}_{L^{p_p}}\,dt\right)^{\frac{Q}{r_2}} T^{\left(1-\frac{Q}{r_1}-\frac{Q}{r_2}\right)}\right] \end{aligned} \tag{2.20}$$

for all $t \in [0,T]$. Hence, $\nabla^\perp\theta \in L^\infty(0,T;L^p(\mathbb{R}^2))$. Integrating (2.19) over $[0,T]$ we have

$$\begin{aligned} &\|\nabla^\perp\theta(t)\|^p_{L^p} + \frac{\kappa C_\alpha}{2}\|\nabla^\perp\theta\|^p_{L^{\frac{2p}{2-\alpha}}} \\ &\quad \le C\int_0^T \|\xi(t)\|^Q_{\dot{\mathcal{F}}^s_{p_1,q}}\|\nabla^\perp\theta(t)\|^Q_{L^{p_2}}\,dt \sup_{0\le t\le T}\|\nabla^\perp\theta(t)\|^p_{L^p} + \|\nabla^\perp\theta_0\|^p_{L^p} < \infty \end{aligned}$$

for all $t \in [0,T]$, and

$$\int_0^T \|\nabla^\perp\theta\|^p_{L^{\frac{2p}{2-\alpha}}}\,dt < \infty.$$

Since $\nabla^\perp\theta \in L^p(0,T;L^{\frac{2p}{2-\alpha}}(\mathbb{R}^2))$, and our choice of p in (2.8) implies

$$2\cdot\frac{(2-\alpha)}{2p} + \frac{\alpha}{p} \le \alpha,$$

the solution $\theta(x,t)$ is regular on $[0,T]$ by applying Theorem 1.1. $\square$

REFERENCES

[1] D. Chae, *The quasi-geostrophic equation in the Triebel-Lizorkin spaces*, Nonlinearity, 16 (2003), pp. 479–495.

[2] D. Chae and J. Lee, *Global well-posedness in the super critical dissipative quasi-geostrophic equations*, Comm. Math. Phys., 233 (2003), pp. 297–311.

[3] P. Constantin, *Geometric statistics in turbulence*, SIAM Rev., 36 (1994), pp. 73–98.

[4] P. Constantin, D. Córdoba, and J. Wu, *On the critical dissipative quasi-geostrophic equation*, Indiana Univ. Math. J., 50 (2001), pp. 97–107.

[5] P. Constantin, A. Majda, and E. Tabak, *Formation of strong fronts in the 2-D quasi-geostrophic thermal active scalar*, Nonlinearity, 7 (1994), pp. 1459–1533.

[6] P. Constantin and J. Wu, *Behavior of solutions of 2D quasi-geostrophic equations*, SIAM J. Math. Anal., 30 (1999), pp. 937–948.

[7] D. Córdoba, *Nonexistence of simple hyperbolic blow-up for the quasi-geostrophic equation*, Ann. of Math. (2), 148 (1998), pp. 1135–1152.

[8] A. Córdoba and D. Córdoba, *A maximum principle applied to quasi-geostrophic equations*, Comm. Math. Phys., 249 (2004), pp. 511–528.

[9] D. Córdoba and C. Fefferman, *Growth of solutions for QG and 2D Euler equations*, J. Amer. Math. Soc., 15 (2002), pp. 665–670.

[10] A. Majda and A. Bertozzi, *Vorticity and Incompressible Flow*, Cambridge University Press, Cambridge, UK, 2002.

[11] G. Prodi, *Un teorema di unicità per le equazioni di Navier-Stokes*, Ann. Mat. Pura Appl. (4), 48 (1959), pp. 173–182.

[12] J. Serrin, *On the interior regularity of weak solutions of the Navier-Stokes equations*, Arch. Ration. Mech. Anal., 9 (1962), pp. 187–191.

[13] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[14] H. Triebel, *Theory of Function Spaces*, Birkäuser Verlag, Boston, 1983.

[15] J. Wu, *The quasi-geostrophic equation and its two regularizations*, Comm. Partial Differential Equations, 27 (2002), pp. 1161–1181.

[16] J. Wu, *Dissipative quasi-geostrophic equations with L^p data*, Electron. J. Differential Equations, 56 (2001), pp. 1–13.

[17] J. Wu, *Inviscid limits and regularity estimates for the solutions of the 2-D dissipative quasi-geostrophic equations*, Indiana Univ. Math. J., 46 (1997), pp. 1113–1124.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1657–1687

GLOBAL SOLUTIONS TO THE GRADIENT FLOW EQUATION OF A NONCONVEX FUNCTIONAL*

G. BELLETTINI†, M. NOVAGA‡, AND E. PAOLINI§

Abstract. We study the L^2-gradient flow of the nonconvex functional $F_\phi(u) := \frac{1}{2}\int_{(0,1)} \phi(u_x)\,dx$, where $\phi(\xi) := \min(\xi^2, 1)$. We show the existence of a global in time possibly discontinuous solution u starting from a mixed-type initial datum u_0, i.e., when u_0 is a piecewise smooth function having derivative taking values both in the region where $\phi'' > 0$ and where $\phi'' = 0$. We show that, in general, the region where the derivative of u takes values where $\phi'' = 0$ progressively disappears while the region where ϕ'' is positive grows. We show this behavior with some numerical experiments.

Key words. nonconvex functionals, forward-backward parabolic equations, finite element method

AMS subject classifications. 35K55, 35B05

DOI. 10.1137/050625333

1. Introduction.

Let $\phi : \mathbb{R} \to [0, +\infty)$ be the nonconvex continuous function defined as

$$\phi(\xi) := \begin{cases} \xi^2 & \text{if } |\xi| \le 1, \\ 1 & \text{otherwise.} \end{cases} \tag{1.1}$$

In this paper we study the L^2-gradient flow of the nonconvex functional

$$F_\phi(u) := \frac{1}{2}\int_{(0,1)} \phi(u_x)\,dx, \qquad u \in BV(0,1), \tag{1.2}$$

where u_x stands for the absolutely continuous part of the distributional derivative of u. Note that $\phi^{**} \equiv 0$, where ϕ^{**} is the convex envelope of ϕ; hence the L^2-lower semicontinuous envelope of F_ϕ is identically zero. Note also that if the initial datum u_0 is smooth and such that $u_{0x}([0,1]) \subset (-1,1)$, it is reasonable to look for a solution of the gradient flow of F_ϕ which coincides with the usual solution of the heat equation starting from u_0. In particular, such a solution cannot coincide with the standing solution $u(x,t) \equiv u_0(x)$ obtained as the gradient flow of the lower semicontinuous envelope of F_ϕ.

The solution $u(x,t)$ of the formal gradient flow of F_ϕ should satisfy the following evolution equation:

$$\begin{cases} u_t = u_{xx}, & \text{where } |u_x| < 1, \\ u_t = 0, & \text{where } |u_x| > 1, \\ u(0) = u_0, \end{cases} \tag{1.3}$$

but the behavior of the interface $\{|u_x| = 1\}$ is not apparent.

*Received by the editors February 25, 2005; accepted for publication (in revised form) June 27, 2005; published electronically January 27, 2006.
http://www.siam.org/journals/sima/37-5/62533.html

†Dipartimento di Matematica, Università di Roma "Tor Vergata" via della Ricerca Scientifica, 00133 Roma, Italy (belletti@mat.uniroma2.it).

‡Dipartimento di Matematica, Università di Pisa, Largo B. Pontecorvo 5, 56127 Pisa, Italy (novaga@dm.unipi.it).

§Dipartimento di Matematica, Università di Firenze, viale Morgagni 67/A, 50134 Firenze, Italy (paolini@math.unifi.it).

While existence and regularity theories for solutions of gradient flow equations originated by convex energies is well established (see, for instance, [12], [31], [4], [2]), very little is known for nonconvex evolution problems. The main difficulty is due to the fact that nonconvexity of the energy density leads in general to ill-posed (i.e., backward-parabolic) problems and, as a consequence, to instabilities in the evolution. The lack of forward parabolicity of the equation shows that even the local in time existence of a solution (in some reasonable class of functions) is not straightforward, as well as uniqueness and regularity. We refer the reader to [30] and to the papers [26], [27], [32], [29], [23], [24], [5], [6] for some results in this direction and for possible regularization techniques. We point out that variational models involving (1.2) have been used in [11] in the context of image segmentation; see also [14]. See also the papers [28], [20], where other backward-forward parabolic equations, such as the Perona–Malik equation corresponding to the choice $\phi_{PM}(\xi) := \log(1+\xi^2)$, have been used to reconstruct a digital image; see [34], [33], [13], [17], [18], [7], [8], [9].

Among nonconvex energy densities, the function ϕ in (1.1) is maybe the simplest one (despite the fact that it is not of class $\mathcal{C}^1$, there are no points in $\mathbb{R} \setminus \{\pm 1\}$ where ϕ'' is negative), and this motivates our choice of studying the gradient flow of the associated functional F_ϕ.

The aim of the present paper is to prove the existence of a reasonable notion of (discontinuous) *global* solution u to the gradient flow of F_ϕ starting from u_0; we stress that u_0 will be allowed to be of mixed type, i.e., to have points where u_{0x} belongs to the locally convex region $(-1,1)$ of ϕ *and* points where u_{0x} belongs to the region $\mathbb{R} \setminus [-1,1]$. We show that, in general, the interface $\{|u_x| = 1\}$ has a velocity, and that the region where u_x takes values in $(-1,1)$ has the tendency to grow at the expenses of the remaining region, with a well determined speed. Thus we are in the presence of a free boundary problem and, in general,

(a) our solution does not coincide with the standing solution $u(x,t) \equiv u_0(x)$;

(b) our solution does not coincide with the solution of (1.3) obtained by keeping the interface $\{|u_x| = 1\}$ fixed and by imposing the condition

$$\lim_{y \to x,\, y \in \{|u_x(\cdot,t)|<1\}} u_x(y,t) = 0 \qquad \text{for } x \in \{|u_x(\cdot,t)| = 1\}, \tag{1.4}$$

i.e., zero Neumann boundary conditions from the side of $\{|u_x| < 1\}$;

(c) these behaviors appear in numerical experiments; see section 7.

Observe that the lack of forward parabolicity precludes, as far as we know, a direct way to construct global solutions based on the comparison principle, such as viscosity solutions [15] or minimal barriers [10]. Moreover, global solutions obtained by using the usual minimization methods (such as the implicit Euler scheme; see [16]) coincide with the solution $u(x,t) \equiv u_0(x)$; this is due to the fact that, in the minimization procedure, the functional F_ϕ can be equivalently replaced with its lower semicontinuous envelope.

In the present paper we restrict the analysis to periodic boundary conditions, even if the same technique can be adapted to different situations such as Neumann or Dirichlet boundary conditions. We base our approach on the study of the system of ODEs obtained as the gradient flow of the restriction $F_{\phi_{|V_N}}$ of F_ϕ to V_N, the space of continuous piecewise affine functions on a uniformly distributed grid of $[0,1]$ of size $1/N$. The function $F_{\phi_{|V_N}}$ turns out to be Lipschitz continuous; nevertheless, it is possible to give a precise notion to the equation $\dot{u} = -\nabla(F_{\phi_{|V_N}})(u)$. After solving the resulting system of ODEs, we pass to the limit as the discretization step goes to zero ($N \to +\infty$), and we identify the limit problem. This sort of regularization is particularly handleable (as a consequence of the special features of ϕ in (1.1)) since the interior of the region $\{|u_x| > 1\}$ has zero velocity, so that we can focus the attention only at the free boundary $\{|u_x| = 1\}$. This is a remarkable simplification, for instance in comparison with the Perona–Malik equation where the quick formation of microstructures in the region where $|u_x| > 1$ seems to be present.

The plan of the paper is the following. In section 2 we state the main result (Theorem 2.4). We look for a solution in the class of ϕ-admissible functions in the sense of Definition 2.1. Several comments clarify both the definition and the theorem (see, in particular, Remark 2.3 concerning condition (4) of Definition 2.1). In section 3 we motivate from a variational point of view the evolution law. In section 4 we discretize the problem and introduce the discretized operator A_u; see Definition 4.4. The rigorous analysis of the discretized scheme is performed in section 5; in particular, in Theorem 5.4 we prove the basic estimates and comparisons necessary to pass to the limit as $N \to +\infty$. In section 6 we prove Theorem 2.4. In Remark 6.16 we discuss in which sense our solution could provide a solution to the gradient flow of the Mumford–Shah functional in one dimension. In section 7 we implement our scheme and show that the numerical experiments are in agreement with Theorem 2.4. In particular, we show that the free boundary $\{|u_x(\cdot)| = 1\}$ has, in general, nonzero speed.

We conclude this introduction by observing that the analysis of the gradient flow of (1.2) could be considered as a first step toward the understanding of the behavior of the Perona–Malik equation.

2. Statement of the main results. We now state the main results of the paper (Theorem 2.4). To this purpose we need some preparation. $BV(0,1)$ stands for the space of functions with bounded variation in $(0,1)$. If $u \in BV(0,1)$ and $x \in (0,1)$, $u(x_-)$ (resp., $u(x_+)$) is the left (resp., right) limit of u at x. We always identify the function u with its representative defined pointwise everywhere as the mean value of u; i.e., $u(x) = (u(x_+)+u(x_-))/2$ for any $x \in (0,1)$. We set $u(0) := u(0_+)$ $u(1) := u(1_-)$. We denote by J_u the jump set of u.

We recall that the distributional derivative of $u \in BV(0,1)$ is represented by a measure Du, with finite total variation in $(0,1)$ (which we denote by $\|Du\|$), and that it splits into the sum of an absolutely continuous part (which we denote by u_x or by u') and a singular part. We refer the reader to [3] for the main properties of BV functions. If $u : [0,T) \to \mathbb{R}$, we indicate by $\frac{d}{dt^+}u$ the right derivative of u; i.e., $\frac{d}{dt^+}u(t) := \lim_{h\to 0^+} \frac{u(t+h)-u(t)}{h}$ for any $t \in [0,T)$, provided the limit is finite.

If u depends on $(x,t) \in (0,1) \times (0,T)$, we write $u(t)(\cdot) = u(\cdot,t) = u(t)$.

Given $B \subseteq \mathbb{R}$ we denote by $\overline{B}$ (resp., $\mathrm{int}(B)$, ∂B, $\#(B)$, $|B|$) the closure (resp., the interior part, the topological boundary, the number of elements, the Lebesgue measure) of B. We denote by $d_{\mathcal{H}}(\cdot,\cdot)$ the Hausdorff distance between sets.

Our analysis is restricted to a subset of $BV(0,1)$ given by the *ϕ-admissible* functions, according to the following definition.

Definition 2.1. *Let $u \in BV(0,1)$ with $u(0) = u(1)$. We say that u is ϕ-*admissible*, and we write $u \in \mathcal{A}_\phi(0,1)$, if there exist a natural number $m \geq 0$ and real numbers $0 < a_1 \leq b_1 < \cdots < a_m \leq b_m < 1$ such that, setting*

$$\sigma_B^\phi(u) := \bigcup_{j=1}^m [a_j, b_j] \subset (0,1), \qquad \sigma_G^\phi(u) := [0,1] \setminus \sigma_B^\phi(u), \tag{2.1}$$

we have

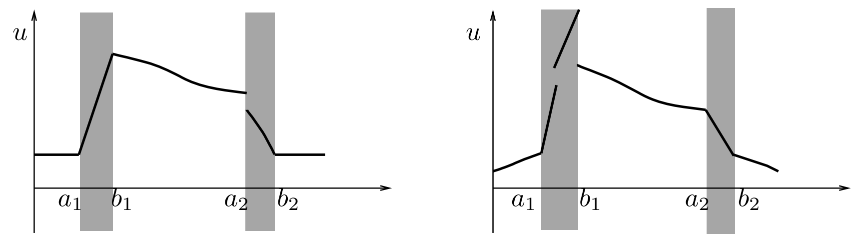


FIG. 2.1. *The gray rectangles correspond to the closed intervals of $\sigma_B^\phi(u)$. The function on the left is ϕ-admissible. The function on the right is not ϕ-admissible, because it is not monotone on the* closed *interval* $[a_1, b_1]$.

(1) $|u(x) - u(y)| \leq |x - y|$ *whenever* $[x, y] \subset \sigma_G^\phi(u)$*;*
(2) *if* $a_j = b_j$ *for some* $j \in \{1, \dots, m\}$*, then* $a_j \in J_u$*;*
(3) *if* $j \in \{1, \dots, m\}$ *and* $a_j < b_j$*, then* $|u(x) - u(y)| > |x - y|$ *whenever* $x, y \in [a_j, b_j]$*,* $x \neq y$*;*
(4) *if* $j \in \{1, \dots, m\}$ *and* $a_j < b_j$*, then* u *is monotone on* $[a_j, b_j]$*.*

Remark 2.2. Let us clarify Definition 2.1.

(a) Note that $\sigma_G^\phi(u) \neq \emptyset$ for any $u \in \mathcal{A}_\phi(0,1)$. We adopt the convention that there are no points a_i, b_j if $m = 0$; in this case, $\sigma_B^\phi(u) = \emptyset$, $\sigma_G^\phi(u) = (0,1)$ and u is one-Lipschitz in the whole of $(0,1)$. The assumptions $a_1 > 0$ and $b_m < 1$ are not restrictive, since we can always assume (up to a translation) that a one-periodic function $u \in \mathcal{C}^1(\mathbb{R})$ is such that $|u_x(0)| < 1$. Due to our periodicity assumption, the point $\{0\}$ is identified with $\{1\}$ and can be considered as belonging to the interior of $\sigma_G^\phi(u)$.
(b) In each interval I of $\sigma_G^\phi(u)$ we have that u is one-Lipschitz; hence, at almost every $x \in I$ we have that $u_x(x)$ belongs (unless $|u_x(x)| = 1$) to the set where ϕ is twice differentiable and $\phi'' > 0$, i.e., $u_x(x) \in (-1, 1)$.
(c) In each interval I of $\sigma_B^\phi(u)$ we have that

$$Du(A) \geq |A| \quad \forall\ A \subseteq I \qquad \text{or} \qquad Du(A) \leq -|A| \quad \forall\ A \subseteq I,$$

with the strict inequalities when $|A| > 0$, where A is any Borel subset of I.
(d) The class $\mathcal{A}_\phi(0,1)$ is L^2-dense in $BV(0,1)$.

The following remark shows some analogy with the entropy condition in hyperbolic conservation laws.

Remark 2.3. Condition (4) in Definition 2.1 is required on the *closed* intervals $[a_j, b_j]$. Hence, since $u(x) = (u(x_+) + u(x_-))/2$ for any $x \in (0,1)$, if u is discontinuous at some a_j and u is nondecreasing on $[a_j, b_j]$ (resp., u is nonincreasing on $[a_j, b_j]$), then $u(a_j) \leq u(a_{j_+})$ (resp., $u(a_j) \geq u(a_{j_+})$). Similarly, it happens if u is discontinuous at some b_j; see Figure 2.1. Condition (4) is fulfilled at each time by the solution that we are going to construct in Theorem 2.4 and arises naturally as a consequence of the approximation procedure through spatial discretizations. Ultimately, it can be considered as a consequence of the fact that, once a region in $\sigma_G^\phi(u_N(t))$ appears for the discretized solutions $u_N(t)$ considered in Theorem 5.4 below, it must persist (and possibly increase) with time.

Let us denote by $AC^2([0, +\infty); L^2(0,1))$ the space of absolutely continuous functions u from $[0, +\infty)$ to $L^2(0,1)$ such that $u_t \in L^2((0, +\infty) \times (0,1))$; see, for instance, [2]. Let $V_N \subset H^1(0,1)$ be the N-dimensional vector space of one-periodic continuous functions on $\mathbb{R}$ which are affine on every interval of the form $[i/N, (i+1)/N]$ with

$i = 0, \dots, N-1$. It is clear that $V_N \subset \mathcal{A}_\phi(0,1)$ and that each function in V_N is N-Lipschitz.

Let us denote by A_u the differential of $F_{\phi|V_N}$ at $u \in V_N$; the linear operator A_u is a discrete Laplace operator with zero blocks corresponding to the region $\sigma_B^\phi(u)$ and zero Neumann boundary conditions on the boundaries; see Remark 4.3 and Definition 4.4 below.

THEOREM 2.4. *Let $u_0 \in \mathcal{A}_\phi(0,1)$, and write*

$$\sigma_B^\phi(u_0) = \bigcup_{j=1}^m [a_j^0, b_j^0].$$

Then there exist a sequence of initial data $(u_0^N) \subset V_N$, a sequence (u^N) of functions taking $[0,+\infty)$ in V_N, and a function $u : (0,1) \times [0,+\infty) \to \mathbb{R}$ with the following properties:

(i) *There exist numbers $0 < a_1^{0N} \le b_1^{0N} < \cdots < a_m^{0N} \le b_m^{0N} < 1$ such that*

$$\sigma_B^\phi(u_0^N) = \bigcup_{j=1}^m [a_j^{0N}, b_j^{0N}], \tag{2.2}$$

and

$$\begin{aligned}
&\lim_{N\to+\infty} \|u_0^N - u_0\|_{L^2} = 0,\\
&\lim_{N\to+\infty} \left(\|u_0^N\|_{BV(0,1)} - \|u_0\|_{BV(0,1)}\right) = 0,\\
&\lim_{N\to+\infty} \Big(d_{\mathcal{H}}(\sigma_G^\phi(u_0^N), \sigma_G^\phi(u_0)) + d_{\mathcal{H}}(\sigma_B^\phi(u_0^N), \sigma_B^\phi(u_0))\Big) = 0,\\
&\lim_{N\to+\infty} F_\phi(u_0^N) = F_\phi(u_0).
\end{aligned} \tag{2.3}$$

(ii) *$u^N : [0,+\infty) \to V_N$ is continuous and right-differentiable, and satisfies*

$$\begin{cases} \dfrac{d}{dt^+} u^N(t) = A_{u^N(t)} u^N(t), & t \in [0,+\infty),\\ u^N(0) = u_0^N. \end{cases} \tag{2.4}$$

(iii) *$u^N, u \in L^\infty((0,+\infty); BV(0,1)) \cap AC^2([0,+\infty); L^2(0,1))$, and $u^N \rightharpoonup u$ weakly in $H^1_{\mathrm{loc}}((0,+\infty); L^2(0,1))$ and weakly* in $L^\infty((0,+\infty); BV(0,1))$ as $N \to +\infty$.*

(iv) *$u(t) \in \mathcal{A}_\phi(0,1)$ for any $t \in [0,+\infty)$.*

(v) *For any $j \in \{1,\dots,m\}$ there exist $T_j \in (0,+\infty]$ and functions $a_j, b_j : [0,T_j) \to (0,1)$ such that*

(v1) *$a_j(0) = a_j^0$, a_j is continuous and nondecreasing;*

(v2) *$b_j(0) = b_j^0$, b_j is continuous and nonincreasing;*

(v3) *$a_j \le b_j$ on $[0,T_j)$, and $\lim_{t\to T_{j\_}} a_j(t) = \lim_{t\to T_{j\_}} b_j(t)$;*

(v4) *$\overline{\bigcup_{j=1}^m (a_j(t), b_j(t))} \subseteq \sigma_B^\phi(u(t)) \subseteq \bigcup_{j=1}^m [a_j(t), b_j(t)]$ for any $t \in [0,+\infty)$, where we have set $(a_j(t), b_j(t)) = [a_j(t), b_j(t)] := \emptyset$ if $t \ge T_j$.*

(vi) *$u_{xx} \in L^2(\Gamma_u)$, where $\Gamma_u := \bigcup_{t\in(0,+\infty)} (\sigma_G^\phi(u(t)) \times \{t\})$, and u is a solution of*

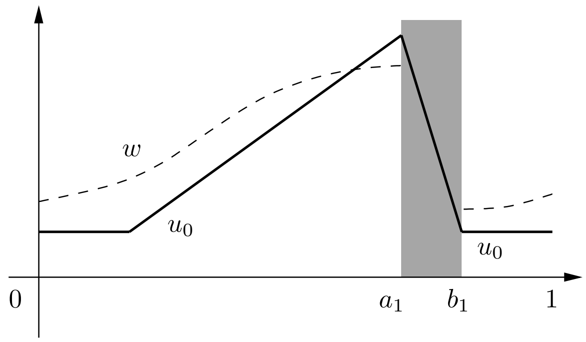


FIG. 2.2. *Remark* 2.6(b). *We construct a function w starting from u_0, such that $w \equiv u_0$ in (a_1, b_1) and that evolves according to the heat equation in $(0, a_1) \cup (b_1, 1)$ with zero Neumann boundary conditions in a_1, b_1 (dashed curve). Recall that we have periodic boundary conditions. Note that $J_{w(t)} = \{a_1, b_1\}$ for $t > 0$, and that $w(t) \notin \mathcal{A}_\phi(0,1)$ for any $t > 0$, since (4) of Definition* 2.1 *is violated at a_1, b_1.*

$$
(2.5)\quad
\begin{cases}
u_t = u_{xx}, & x \in \sigma_G^\phi(u(t)),\ t \in (0,+\infty),\\
u_t = 0, & x \in \mathrm{int}(\sigma_B^\phi(u(t))),\ t \in (0,+\infty),\\
\displaystyle\lim_{y \to x,\, y \in \sigma_G^\phi(u(t))} u_x(y,t) = 0, & x \in \partial\sigma_G^\phi(u(t)) \setminus \{0,1\},\ t \in (0,+\infty),\\
u(x,0) = u_0(x), & x \in (0,1),\\
u(0,t) = u(1,t),\ u_x(0,t) = u_x(1,t), & t \in (0,+\infty).
\end{cases}
$$

(vii) *For any $t \in (0,+\infty)$ we have*
$\sup_{\sigma_G^\phi(u(t))} |u_x(\cdot,t)| < 1$;
$\sup_{[0,1]} u(\cdot,t) \le \sup_{[0,1]} u_0$;
$\inf_{[0,1]} u(\cdot,t) \ge \inf_{[0,1]} u_0$;
$\|Du(\cdot,t)\| \le \|Du_0\|$.

The proof of Theorem 2.4 is achieved in sections 5 and 6. In particular, (i) is given by Lemma 6.1, (ii) is given by Theorem 5.4, (iii) is the content of Remark 6.5, (iv) is given by Lemma 6.12, (v) is given by Lemma 6.8, Remark 6.6, and Lemma 6.12, and (vi) is the content of Theorem 6.14. Finally, the first inequality in (vii) follows from (vi) and the maximum principle applied to u_x, while the last three inequalities in (vii) are consequences of (c) and (d) of Theorem 5.4.

Remark 2.5.

(a) In general a function u and intervals (a_j, b_j) satisfying (v) and (vi) of Theorem 2.4 are not unique: it is easy to construct a solution w of (2.5) satisfying also the requirement

$$
(2.6)\qquad \sigma_B^\phi(w(t)) = \sigma_B^\phi(u_0) \qquad \forall\, t \in (0,+\infty),
$$

and the function w in general cannot coincide with u. Indeed, $w(t) = u(t)$ for all times t for which $w(t) \in \mathcal{A}_\phi(0,1)$, but the property $w(t) \in \mathcal{A}_\phi(0,1)$ for all $t \in (0,+\infty)$ is in general violated; see Figure 2.2. In fact, condition (4) in Definition 2.1 cannot be satisfied for all times by w (cf. Remark 2.3), unless $\sigma_G^\phi(w(\cdot))$ is allowed to expand, in contrast with (2.6).

(b) If we do not require the functions a_j, b_j to be monotone (nondecreasing and

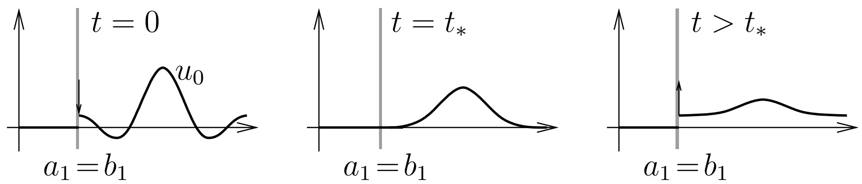


FIG. 2.3. *The "bouncing" solution discussed in Example* 1.

nonincreasing, respectively), several different solutions could be constructed; see Figure 2.4(b).

One can ask whether a function u and intervals (a_j, b_j) satisfying (iv), (v), and (vi) of Theorem 2.4 are unique. This is not the case, as shown by the following example related, in spirit, to the so-called fattening phenomenon in mean curvature flow (see [22] for similar behaviors concerning the evolution of the Mumford–Shah functional in one dimension).

Example 1. Let us construct an initial datum $u_0 \in \mathcal{A}_\phi(0,1)$ as follows:

u_0 has only one jump point $a_1 = b_1 = 1/2$;
$u_0 = 0$ in $(0, 1/2)$;
u_0 is a smooth function in $(1/2, 1)$ with the following property: $|u_{0x}| < 1$ and, if we flow $u_{0|(1/2,1)}$ by the heat equation with zero Neumann boundary conditions in $\{1/2, 1\}$, then there is a first time $t_* > 0$ for which the solution, evaluated at the point $1/2$, touches the horizontal axis with *zero vertical velocity* and then, for t immediately after t_*, becomes *positive* at $1/2$; see Figure 2.3.

Then we can exhibit two functions u_1, u_2, which coincide for $t \in [0, t_*]$ but differ for $t \in (t_*, +\infty)$, and both satisfy (iv), (v), and (vi) of Theorem 2.4. The function u_1 is defined as follows: $u_1 = 0$ in $(0, 1/2) \times [0, +\infty)$; u_1 equals, in $(1/2, 1) \times [0, t_*)$, the solution of the heat equation with zero Neumann boundary conditions in $\{1/2, 1\}$; u_1 equals, in $(0,1) \times [t_*, +\infty)$, the solution of the heat equation with zero Neumann boundary conditions in $\{0, 1\}$ starting from $u_1(t_{*-})$. Namely, immediately after the time t_* when the two graphs of the solution on the left and on the right of $1/2$ join, the evolution continues with one graph only, and the jump disappears.

The function u_2 is defined as follows: $u_2 = 0$ in $(0, 1/2) \times [0, +\infty)$; u_2 equals, in $(1/2, 1) \times [0, +\infty)$, the solution of the heat equation with zero Neumann boundary conditions in $\{1/2, 1\}$. That is, the function u_2 "bounces" at $1/2$ at time t_*, the evolutions in $(0, 1/2)$ and in $(1/2, 1)$ do not "see" each other, and $1/2$ becomes again a jump point of $u_2(t)$ for t immediately larger than t_*.

Remark 2.6.

(a) As a consequence of (v) of Theorem 2.4, the set-valued map $t \in [0, +\infty) \to \sigma_G^\phi(u(t)) \subseteq (0, 1)$ is nondecreasing up to a finite number of points (at most m), and the number of connected components with nonempty interior of $\sigma_B^\phi(u(\cdot))$ is nonincreasing. Itmay happen that at some time $\bar{t}_j \in (0, T_j)$ the interval $[a_j(\bar{t}_j), b_j(\bar{t}_j)]$ is reduced to a point not belonging to $J_{u(\bar{t}_j)}$ (recall conclusion (iv) of Theorem 2.4 and Definition 2.1(2)) but belonging to $J_{u(t)}$ for some $t \in (\bar{t}_j, T_j)$ (as it happens for the function u_2 in Example 1). At time T_j at least one of the intervals in $\sigma_B^\phi(u(\cdot))$ disappears (provided $T_j < +\infty$).

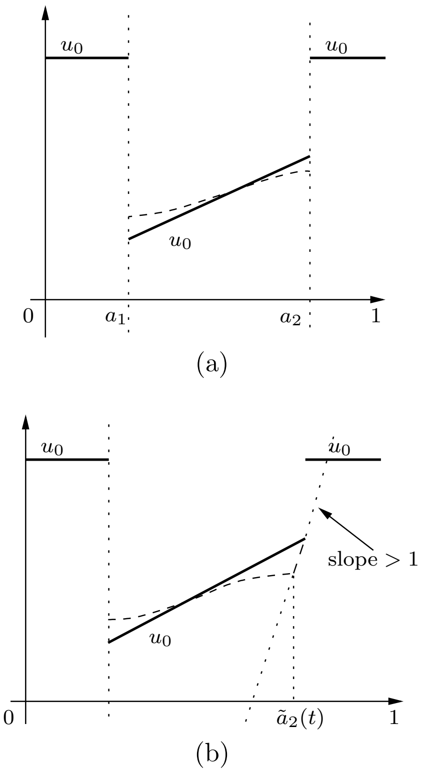


FIG. 2.4. *Remark* 2.5(b). *For the function u_0 we have $a_1^0 = b_1^0$ and $a_2^0 = b_2^0$. In* (a) *is displayed the solution u of Theorem* 2.4 *starting from u_0 for which $a_1(t) = b_1(t) \equiv a_1^0$, $a_2(t) = b_2(t) \equiv a_2^0$, u evolves according to the heat equation in $[a_1(t), a_2(t)]$ with zero Neumann boundary conditions (dashed curve), and $u(t) \equiv u_0$ in $[0,1] \setminus [a_1(t), a_2(t)]$. In* (b) *we construct a function w with $w(t) \in \mathcal{A}_\phi(0,1)$, such that w evolves according to the heat equation in $[a_1^0, \tilde{a}_2(t)]$ with zero Neumann boundary conditions, and $\tilde{a}_2(t)$ is decreasing in time, in such a way that the corresponding point $w(\tilde{a}_2(t), t)$ slides on a line with slope greater than one; hence the function w does not satisfy condition* (v1) *of Theorem* 2.4.

(b) A weak formulation of (2.5) is given by

$$\int_{(0,1)\times(0,+\infty)} u\,\psi_t\,dx\,dt - \int_{\mathrm{int}(\Gamma_u)} u_x\psi_x\,dx\,dt = 0 \tag{2.7}$$

for any $\psi \in \mathcal{C}_c^1([0,1]\times[0,+\infty))$.

(c) Solutions verifying conditions (iv), (v), and (vi) of Theorem 2.4 do not satisfy the comparison principle, in the sense that it is easy to find solutions u_1, u_2 such that $u_1(\cdot,0) \le u_2(\cdot,0)$ on $(0,1)$, but $u_1(\bar{x},\bar{t}) > u_2(\bar{x},\bar{t})$ for some $(\bar{x},\bar{t}) \in (0,1)\times(0,+\infty)$; see Figure 2.5.

Remark 2.7.

(a) Under sufficient regularity on u we can predict the speed of the free boundary $\partial\sigma_G^\phi(u(\cdot))$. For instance, assume that a_j is of class $\mathcal{C}^1$ in a neighborhood U of $\bar{t} \in (0, T_j)$ and that $a_j'(\bar{t}) \neq 0$. Assume in addition that $u(\cdot,\cdot)$ is twice differentiable in $\bigcup_{t\in U} \sigma_G^\phi(u(t)) \times \{t\}$ up to the boundary. Then from the equality

$$u(a_j(t), t) = u_0(a_j(t))$$

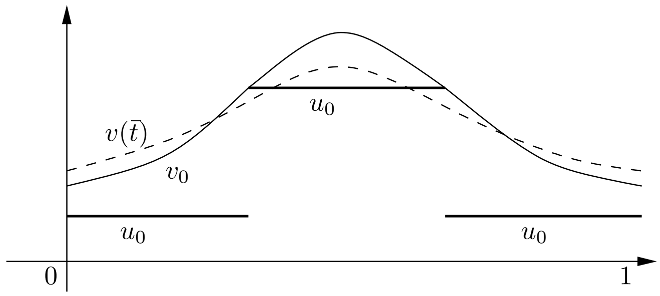


FIG. 2.5. *Remark* 2.6(c). *In general the solution* u *of Theorem* 2.4 *cannot satisfy the comparison principle. Indeed, let* u_0 *and* v_0 *be as in the figure,* $u_0 \le v_0$*, where we assume that the function* v_0 *is one-Lipschitz, so that* $\sigma_G^\phi(v_0) = (0,1)$*. Moreover,* $u(t) \equiv u_0$ *for any* $t \in (0,+\infty)$*. On the other hand, the solution* v *starting from* v_0 *given by Theorem* 2.4 *is the usual solution of the heat equation in* $(0,1)$ *with zero Neumann boundary conditions. Hence, at some time* $\bar t > 0$ *and at some* $\bar x \in (0,1)$ *it happens that* $v(\bar x, \bar t) < u(\bar x, \bar t)$*.*

valid in the neighborhood of $\bar t$ it follows, using the third equality in (2.5), that

$$u_t(a_j(t)_-, t) = \frac{d}{dt} u(a_j(t), t) = u_{0x}(a_j(t)_+) a_j'(t). \tag{2.8}$$

Hence, using the first equation in (2.5), we get

$$a_j'(\bar t) = \frac{u_{xx}(a_j(\bar t)_-, \bar t)}{u_{0x}(a_j(\bar t)_+)}. \tag{2.9}$$

Similarly, under the corresponding regularity assumptions and provided $b_j'(\bar t) \neq 0$, we get

$$b_j'(\bar t) = \frac{u_{xx}(b_j(\bar t)_+, \bar t)}{u_{0x}(b_j(\bar t)_-)}. \tag{2.10}$$

(b) We expect that if $u_0 \in \mathcal{C}^{1,1}(\sigma_G^\phi(u_0))$ and $\lim_{y\to x,\, y\in\sigma_G^\phi(u_0)} u_{0x}(y) = 0$ for any $x \in \partial\sigma_G^\phi(u_0)$, then

$$\|u_{xx}\|_{L^\infty(\sigma_G^\phi(u(t)))} \le \|u_{0xx}\|_{L^\infty(\sigma_G^\phi(u_0))}, \qquad t \ge 0. \tag{2.11}$$

Indeed, assuming we can differentiate a_j, b_j in $(0, T_j)$ and $u(\cdot, t)$ in $\sigma_G^\phi(u(t))$ up to the boundary, arguing as in (a) we get

$$\frac{u_{xx}(a_j(t)_-, t) a_j'(t)}{u_{0x}(a_j(t)_+)} \ge 0, \qquad \frac{u_{xx}(b_k(t)_+, t) b_k'(t)}{u_{0x}(b_k(t)_-)} \ge 0 \tag{2.12}$$

for any $t \ge 0$. Differentiating the equalities $u_x(a_j(t)_-, t) = u_x(b_k(t)_+, t) = 0$

with respect to t and using (2.12), we then get

$$\begin{aligned} &\frac{u_{xxx}(a_j(t)_-,t)}{u_{0x}(a_j(t)_+)} = -\frac{u_{xx}(a_j(t)_-,t)a_j'(t)}{u_{0x}(a_j(t)_+)} \le 0, \\ &u_{xxx}(a_j(t)_-,t) = 0 \quad \text{if} \quad \frac{u_{xx}(a_j(t)_-,t)}{u_{0x}(a_j(t)_+)} < 0, \\ &\frac{u_{xxx}(b_k(t)_+,t)}{u_{0x}(b_k(t)_-)} = -\frac{u_{xx}(b_k(t)_+,t)b_k'(t)}{u_{0x}(b_k(t)_-)} \le 0, \\ &u_{xxx}(b_k(t)_+,t) = 0 \quad \text{if} \quad \frac{u_{xx}(b_k(t)_+,t)}{u_{0x}(b_k(t)_-)} > 0. \end{aligned} \tag{2.13}$$

Letting $v := u_{xx}$ and differentiating (2.5) twice with respect to x, we obtain

$$\begin{cases} v_t = v_{xx}, & x \in \sigma_G^\phi(u(t)),\ t \in (0,+\infty), \\ v_t = 0, & x \in \operatorname{int}(\sigma_B^\phi(u(t))),\ t \in (0,+\infty), \\ v(x,0) = u_{0xx}(x), & x \in (0,1), \end{cases} \tag{2.14}$$

with the boundary conditions on $\partial\sigma_G^\phi(u(t))$ given by (2.13). Note that, from the third equality in (2.5), for any $t \ge 0$ it follows that

$$\int_{\sigma_G^\phi(u(t))} v(x,t)\,dx = 0 \Longrightarrow \max_{\sigma_G^\phi(u(t))} v(\cdot,t) \ge 0, \qquad \min_{\sigma_G^\phi(u(t))} v(\cdot,t) \le 0.$$

The boundary conditions (2.13) then imply that $v(\cdot,t)$ assumes its maximum and minimum in the interior of $\sigma_G^\phi(u(t))$; hence (2.11) follows from (2.14) by the maximum principle.

Let us observe that from (2.8) and (2.11) it follows that

$$\|a_j'\|_{L^\infty(0,T_j)} \le \|u_{0xx}\|_{L^\infty(\sigma_G^\phi(u_0))}, \qquad \|b_j'\|_{L^\infty(0,T_j)} \le \|u_{0xx}\|_{L^\infty(\sigma_G^\phi(u_0))}.$$

In particular, we also expect that the functions a_j and b_j are Lipschitz continuous on $[0,T_j)$.

Remark 2.8. It is clear that Theorem 2.4 holds also for the function

$$\overline{\phi}(\xi) := \min(1, \phi_{PM}(\xi)) = \min(1, \log(1+\xi^2)).$$

In the present paper, solutions u to the gradient flow of $F_{\overline{\phi}}$ are intended as those functions satisfying (iv), (v), and (vi) (with $u_t = u_{xx}$ replaced by $u_t = (\overline{\phi}'(u_x))_x$) of Theorem 2.4. These solutions could be compared with some notion of weak solutions of the gradient flow of $F_{\phi_{PM}}$; see [29]. We can observe that u is not a BV-distributional solution of the Perona–Malik equation in the sense of [29, Definition 1]; see (2.7). However, u turns out to be a Young-varifold solution of the Perona–Malik equation; see [19], [21]. We also observe that if $a = b \in (0,1)$ is a jump point of $u(t)$ and if u is sufficiently smooth in a neighborhood of a (see Remark 2.7), then as a consequence of (2.9), (2.10), we have that $a'(t) = 0$. This is consistent with [29, formula (3)], in connection with the notion of generalized solution. Finally, observe that $a'(t) = 0$ is also a consequence of the $AC_2([0,+\infty); L^2(0,1))$ regularity of u.

3. First variation. In this section we want to identify the L^2-gradient of the functional F_ϕ in (1.2) on a suitable dense subspace X of $L^2(0,1)$; see Definition 3.3. We begin by computing the first variation of F_ϕ along functions $\psi \in \mathrm{Lip}(0,1)$.

PROPOSITION 3.1. *Let $u \in \mathcal{A}_\phi(0,1)$ be such that $\sigma_B^\phi(u) = \bigcup_{j=1}^m [a_j, b_j]$, $a_j < b_j$ for any $j = 1, \dots, m$,*

$$u \in H^2(\sigma_G^\phi(u)) \qquad \text{and} \qquad \sup_{\sigma_G^\phi(u)} |u_x| < 1.$$

Then for any $\psi \in \mathrm{Lip}(0,1)$ with $\psi(0) = \psi(1)$ we have

$$\begin{aligned} \frac{d}{d\lambda} F_\phi(u+\lambda\psi)_{|\lambda=0} &= \int_{\sigma_G^\phi(u)} u_x \psi_x \, dx \\ &= -\int_{\sigma_G^\phi(u)} u_{xx} \psi \, dx \\ &\quad + \sum_{j=1}^m \big(u_x(a_{j-})\psi(a_j) - u_x(b_{j+})\psi(b_j) \big). \end{aligned} \tag{3.1}$$

Proof. Since $\sup_{\sigma_G^\phi(u)} |u_x| < 1$ and $\psi \in \mathrm{Lip}(0,1)$, we have $\sigma_G^\phi(u+\lambda\psi) = \sigma_G^\phi(u)$ for $|\lambda|$ small enough. In addition, $\sigma_B^\phi(u+\lambda\psi) = \sigma_B^\phi(u)$ for $|\lambda|$ small enough. For such λ we have

$$\begin{aligned} F_\phi(u+\lambda\psi) &= \frac{1}{2}\int_{\sigma_B^\phi(u+\lambda\psi)} 1 \, dx + \frac{1}{2}\int_{\sigma_G^\phi(u+\lambda\psi)} (u_x + \lambda\psi_x)^2 \, dx \\ &= \frac{|\sigma_B^\phi(u)|}{2} + \frac{1}{2}\int_{\sigma_G^\phi(u)} (u_x + \lambda\psi_x)^2 \, dx \\ &= \frac{|\sigma_B^\phi(u)|}{2} + \frac{1}{2}\int_{\sigma_G^\phi(u)} (u_x)^2 \, dx + \lambda \int_{\sigma_G^\phi(u)} u_x \psi_x \, dx + O(\lambda^2). \end{aligned}$$

Then (3.1) follows with an integration by parts, using the assumptions $u \in H^2(\sigma_G^\phi(u))$ and $\psi(0) = \psi(1)$. ◻

Remark 3.2. Observe that the variations $u \to u + \lambda\psi$, as in Proposition 3.1, cannot increase the number of singular points of $u \in \mathcal{A}_\phi(0,1)$.

If u is as in Proposition 3.1 it follows that

$$\begin{aligned} &\inf_{\substack{\psi \in \mathrm{Lip}(0,1),\, \psi(0)=\psi(1) \\ \|\psi\|_{L^2} \le 1}} \frac{d}{d\lambda} F_\phi(u+\lambda\psi)_{|\lambda=0} \\ &= \begin{cases} -\|u_{xx}\|_{L^2(\sigma_G^\phi(u))} & \text{if } u_x(a_{j-}) = u_x(b_{j+}) = 0, \quad 1 \le j \le m, \\ -\infty & \text{otherwise.} \end{cases} \end{aligned} \tag{3.2}$$

DEFINITION 3.3. *We denote by X the dense subset of $L^2(0,1)$ consisting of the functions u as in Proposition* 3.1 *and satisfying $u_x(a_{j-}) = u_x(b_{j+}) = 0$ for any $1 \le j \le m$.*

Once we fix $u \in X$, the right-hand side of (3.1), if considered as a function of ψ, is a linear functional defined on the Lipschitz functions ψ in $(0,1)$ with $\psi(0) = \psi(1)$ (which form a dense subset of $L^2(0,1)$), which is continuous with respect to the

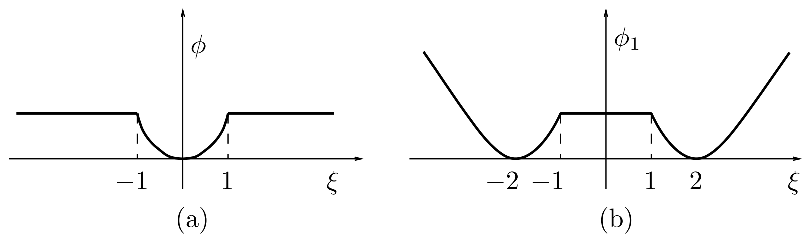


FIG. 3.1. (a) *the function ϕ considered in the present paper;* (b) *the function ϕ_1 of Remark* 3.6.

$L^2(0,1)$-norm. Therefore it can be extended on the whole of $L^2(0,1)$, thus providing a well-defined unique left-hand side of (3.1) for any $\psi \in L^2(0,1)$, and

$$\inf_{\substack{\psi\in\mathrm{Lip}(0,1),\,\psi(0)=\psi(1)\\ \|\psi\|_{L^2}\le 1}} \frac{d}{d\lambda}F_\phi(u+\lambda\psi)_{|\lambda=0} = \inf_{\substack{\psi\in L^2(0,1)\\ \|\psi\|_{L^2}\le 1}} \frac{d}{d\lambda}F_\phi(u+\lambda\psi)_{|\lambda=0}. \tag{3.3}$$

The infimum in (3.3) is attained at $\tilde\psi \in L^2(0,1)$, where

$$\tilde\psi = \begin{cases} 0 & \text{on } \sigma_B^\phi(u),\\ \|u_{xx}\|^{-1}_{L^2(\sigma_G^\phi(u))} u_{xx} & \text{on } \sigma_G^\phi(u). \end{cases}$$

It follows that the L^2-gradient flow of F_ϕ starting from $u_0 \in X$ is given by the free boundary problem (2.5).

As already observed in the introduction, in general, solutions to problem (2.5) are not unique, since the motion of the free boundary $\partial\sigma_G^\phi(u)$ is not prescribed. However, among all ϕ-admissible solutions we can look for those which most decrease the energy F_ϕ. This is expressed by the following proposition, which follows by a direct computation and recalling that $\phi \equiv 1/2$ on $\sigma_B^\phi(u)$.

PROPOSITION 3.4. *Let u be a solution of* (2.5) *satisfying* (iv) *of Theorem* 2.4. *Then for almost every $t \in (0,+\infty)$ we have*

$$\frac{d}{dt}F_\phi(u(t)) = -\frac{1}{2}\frac{d}{dt}|\sigma_G^\phi(u(t))| - \int_{\sigma_G^\phi(u(t))} |u_{xx}(x,t)|^2\,dx. \tag{3.4}$$

Remark 3.5. Proposition 3.4 implies that in order to most decrease the energy F_ϕ, the region $\sigma_G^\phi(u)$ should expand as fast as possible, compatibly with the ϕ-admissibility of u.

Remark 3.6. Our results can be extended to other integrands. Let us consider, for example, the potential in Figure 3.1(b), i.e.,

$$\phi_1(\xi) := \begin{cases} |\xi-2|^2 & \text{if } \xi \ge 1,\\ |\xi+2|^2 & \text{if } \xi \le -1,\\ 1 & \text{otherwise,} \end{cases} \tag{3.5}$$

which is related to the ones considered in [26], [1], [36], [35]. Then Definition 2.1 still makes sense, provided that we define $\sigma_B^{\phi_1}(u)$ as the finite union of closed intervals where $|u(x)-u(y)| < |x-y|$, and $\sigma_G^{\phi_1}(u) = [0,1] \setminus \sigma_B^{\phi_1}(u)$ as the finite union of intervals where either $u(x)-u(y) \ge x-y$ or $u(x)-u(y) \le -(x-y)$. Let us denote by

$\sigma^{\phi_1}_{G,+}(u)$ (resp., $\sigma^{\phi_1}_{G,-}(u)$) the subset of $\sigma^{\phi_1}_G(u)$ where u is increasing (resp., decreasing). The first variation of F_{ϕ_1} can be computed as in Proposition 3.1, and the evolution equation corresponding to (2.5) reads as

$$
\begin{cases}
u_t = u_{xx}, & x \in \sigma^{\phi_1}_G(u(t)),\ t \in (0,+\infty),\\
u_t = 0, & x \in \mathrm{int}(\sigma^{\phi_1}_B(u(t))),\ t \in (0,+\infty),\\
\displaystyle\lim_{y\to x,\, y\in\sigma^{\phi_1}_{G,\pm}(u(t))} u_x(y,t) = \pm 2, & x \in \partial\sigma^{\phi_1}_{G,\pm}(u(t)) \setminus \{0,1\},\ t \in (0,+\infty),\\
u(x,0) = u_0(x), & x \in (0,1),\\
u(0,t) = u(1,t),\ u_x(0,t) = u_x(1,t), & t \in (0,+\infty).
\end{cases}
\tag{3.6}
$$

Since equality (3.4) still holds, also in this case the region $\sigma^{\phi_1}_G(u(\cdot))$ expands as fast as possible, compatibly with (3.6). We finally observe that the analogue of Theorem 2.4 is not expected to hold in this case; cf. Remark 7.1.

Remark 3.7. Let us consider a continuous function $\phi_2 : \mathbb{R} \to [0,+\infty)$ of the form $\phi_2(\xi) = \xi^2$ for $|\xi| \in [0,1]$, and $\phi_2(\xi) = \alpha\xi + \beta$ for $|\xi| \in [1,+\infty)$, where $\alpha + \beta = 1$ and $\alpha \geq 0$. The computations leading to (3.2) can be repeated for the functional F_{ϕ_2} and give the following result:

$$
\begin{aligned}
&\inf_{\substack{\psi\in\mathrm{Lip}(0,1),\,\psi(0)=\psi(1)\\ \|\psi\|_{L^2}\leq 1}} \frac{d}{d\lambda} F_{\phi_2}(u+\lambda\psi)_{|\lambda=0}\\
&= \begin{cases}
-\|u_{xx}\|_{L^2(\sigma^{\phi_2}_G(u))} & \text{if } |u_x(a_{j-})| = |u_x(b_{j+})| = \alpha/2, \quad 1 \leq j \leq m,\\
-\infty & \text{otherwise,}
\end{cases}
\end{aligned}
\tag{3.7}
$$

where the interior Neumann boundary condition, for example in a_j, is equal to $\alpha/2$ (resp., $-\alpha/2$) if u_0 is increasing (resp., decreasing) in $[a_j, b_j]$.

In particular, the resulting PDE arising from (3.7) is different from (2.5) (since the conditions on the free boundary are different) unless $\alpha = 0$, i.e., $\phi_2 = \phi$.

4. Discretization. In this section we define the spatial discretization used to approximate problem (2.5). In particular, in Definition 4.4 we introduce the discretized operator A_v.

Let $N \in \mathbb{N}$ and $i \in \{1,\ldots,N\}$. To simplify notation, we set $i+1 = 1$ and $[i,i+1] = [0,1]$ when $i = N$, and $i-1 = N$ and $[i-1,i] = [0,1]$ when $i = 1$.

For any $i = 1,\ldots,N$ we define the *hat* function $h^i \in H^1(0,1)$ as

$$
h^i(x) := \begin{cases}
Nx - (i-1) & \text{if } Nx \in [i-1,i],\\
i+1-Nx & \text{if } Nx \in [i,i+1],\\
0 & \text{otherwise.}
\end{cases}
$$

We denote by V_N the N-dimensional vector subspace of $H^1(0,1)$ generated by $h^1,\ldots,h^N$. Each function $v \in V_N$ is Lipschitz and is the restriction to $[0,1]$ of an affine continuous periodic function defined on $\mathbb{R}$.

For any $i = 1,\ldots,N$ we define the *flat* function $k^i \in L^2(0,1)$ as

$$
k^i(x) := \begin{cases}
1 & \text{if } Nx \in (i-1,i],\\
0 & \text{otherwise.}
\end{cases}
$$

We denote by W_N the N-dimensional vector subspace of $L^2(0,1)$ generated by $k^1,\dots,k^N$. W_N is the space of all piecewise constant functions on the grid.

The spaces $\bigcup_N V_N$ and $\bigcup_N W_N$ are dense in $BV(0,1)$ with respect to the weak*-topology.

Given $v \in V_N$ (resp., $w \in W_N$) we denote with $v_1,\dots,v_N$ the coordinates of v with respect to the basis $\{h^1,\dots,h^N\}$ (resp., $\{k^1,\dots,k^N\}$), i.e.,

$$v = \sum_{i=1}^N v_i h^i, \qquad v_i = v(i/N),$$

$$w = \sum_{i=1}^N w_i k^i, \qquad w_i = w\left(\frac{i-\frac{1}{2}}{N}\right).$$

We recall that

$$\int_{(0,1)} u\,dx = \frac{1}{N}\sum_{i=1}^N u_i, \qquad u \in V_N \cup W_N.$$

We define the scalar product $\langle\cdot,\cdot\rangle$ on V_N and on W_N as

$$\langle v,\overline{v}\rangle = \frac{1}{N}\sum_{i=1}^N v_i\overline{v}_i, \quad \langle w,\overline{w}\rangle = \frac{1}{N}\sum_{i=1}^N w_i\overline{w}_i, \qquad v,\overline{v}\in V_N,\ w,\overline{w}\in W_N.$$

Recall that

$$\langle w,\overline{w}\rangle = \int_{(0,1)} w\overline{w}\,dx = \frac{1}{N}\sum_{i=1}^N w_i\overline{w}_i, \qquad w,\overline{w}\in W_N.$$

Given $v \in V_N$ we define

$$\begin{aligned}
\|v\|_{L^\infty} &:= \max\{|v_i| \colon i = 1,\dots,N\},\\
\|v\|_{L^2} &:= \langle v,v\rangle^{\frac{1}{2}},\\
\|\nabla v\|_{L^1} &:= \sum_{i=1}^N |v_{i+1}-v_i| = \int_{(0,1)} |v_x|\,dx.
\end{aligned}$$

Definition 4.1. *We define the linear map $D^+\colon V_N \to W_N$ as the restriction of the weak derivative taking $H^1(0,1)$ in $L^2(0,1)$. In coordinates,*

$$(D^+v)_i = N(v_{i+1}-v_i), \qquad i \in \{1,\dots,N\}.$$

We let $D^-\colon W_N \to V_N$ be the adjoint operator of $-D^+$.

The operator D^- satisfies $\langle D^-w, v\rangle = -\langle w, D^+v\rangle$ for all $v \in V_N$ and $w \in W_N$. In coordinates,

$$(D^-w)_i = N(w_i - w_{i-1}), \qquad i \in \{1,\dots,N\}.$$

Definition 4.2. *Given $v \in V_N$ we define $\Psi_v \in W_N$ in coordinates by*

$$(\Psi_v)_i = \begin{cases} 1 & \text{if } |(D^+v)_i| \le 1,\\ 0 & \text{otherwise,}\end{cases} \qquad i \in \{1,\dots,N\}.$$

If v is ϕ-admissible, the function $\Psi_v\colon (0,1) \to \mathbb{R}$ is the characteristic function of the set $\sigma_G^\phi(v)$.

Note that the restriction of F_ϕ to V_N reads as follows: given $v \in V_N$,

$$
\begin{aligned}
F_\phi(v) &= \frac{1}{2N} \sum_{i=1}^{N} \min\big(((D^+v)_i)^2, 1\big) \\
&= \frac{1}{2}\langle \Psi_v D^+v, D^+v\rangle + \frac{1}{2}\int_{(0,1)} (1-\Psi_v)\,dx,
\end{aligned}
\tag{4.1}
$$

where

$$
\langle \Psi_v D^+v, D^+v\rangle = \sum_{i=1}^{N} (\Psi_v)_i (D^+v)_i (D^+v)_i.
$$

Remark 4.3. The function $F_{\phi|V_N}$ is Lipschitz in V_N and is of class $\mathcal{C}^\infty$ out of the polyhedral hypersurface $H := \bigcup_{i=1}^N H_i$, where $H_i := \{v \in V_N : |(D^+v)_i| = 1\}$. Assume that $v \in V_N \setminus H$. Then, for any $\overline{v} \in V_N$, we have

$$
\lim_{\lambda\to 0} \frac{\Psi_{v+\lambda\overline{v}} - \Psi_v}{\lambda} = 0 \in V_N.
$$

Therefore, using also (4.1), we get

$$
\begin{aligned}
\lim_{\lambda\to 0} \frac{F_\phi(v+\lambda\overline{v}) - F_\phi(v)}{\lambda} &= \frac{1}{2}\langle \Psi_v D^+\overline{v}, D^+v\rangle + \frac{1}{2}\langle \Psi_v D^+v, D^+\overline{v}\rangle \\
&= -\big\langle D^-\big(\Psi_v D^+v\big), \overline{v}\big\rangle.
\end{aligned}
\tag{4.2}
$$

More generally, for $v \in V_N$ there exists the limit

$$
\begin{aligned}
&\lim_{\lambda\to 0^+} \frac{F_\phi(v+\lambda\overline{v}) - F_\phi(v)}{\lambda} \\
&= \sum_{i\colon |(D^+v)_i|<1} (D^+v)_i(D^+\overline{v})_i + \sum_{i\colon |(D^+v)_i|=1} \min\big(\big((D^+v)_i(D^+\overline{v})_i\big), 0\big) \\
&= -\big\langle D^-\big(\Psi_v D^+v\big), \overline{v}\big\rangle - \sum_{i\colon |(D^+v)_i|=1} \max\big(\big((D^+v)_i(D^+\overline{v})_i\big), 0\big) \\
&\le -\big\langle D^-\big(\Psi_v D^+v\big), \overline{v}\big\rangle.
\end{aligned}
\tag{4.3}
$$

Note that both the limits in (4.2) and (4.3) attain their minimum on $\{v \in V_N : \|v\|_{L^2} = 1\}$ at

$$
\overline{v} = \frac{D^-\big(\Psi_v D^+v\big)}{\big\|D^-\big(\Psi_v D^+v\big)\big\|_{L^2}}.
$$

We are now in a position to define the discretized operator.

Definition 4.4. *Given any $v \in V_N$ we define the linear operator $A_v\colon V_N \to V_N$ as follows: for any $\overline{v} \in V_N$ we let*

$$
A_v\overline{v} := D^-\big(\Psi_v D^+\overline{v}\big).
$$

In coordinates, we have

$$(A_v\overline{v})_i = \frac{(\Psi_v)_i[\overline{v}_{i+1} - \overline{v}_i] - (\Psi_v)_{i-1}[\overline{v}_i - \overline{v}_{i-1}]}{1/N^2}.$$

Remark 4.5. By Remark 4.3, if $v \in V_N \setminus H$, then $A_v = -\nabla(F_{\phi_{|V_N}})(v)$, where ∇ indicates the gradient of the function $F_{\phi_{|V_N}}$ defined in the finite-dimensional space V_N. Note also that the equality holds in the last line of (4.3) if we take $v \in V_N$ and $\overline{v} = A_v v$.

Remark 4.6. If $v, \overline{v} \in V_N$ are such that $\Psi_v = \Psi_{\overline{v}}$, then $A_v = A_{\overline{v}}$.

5. Discretized evolution. Maximum principles. The aim of this section is to prove Theorem 5.4, which is a key step in the proof of Theorem 2.4. We begin with some elementary lemmata.

LEMMA 5.1. *Let $u_1, \dots, u_n$ be real continuous right-differentiable functions in an interval $[0, t_1)$. Define $M(t) := \max_{i=1,\dots,n} u(t)_i$. Then $M(t)$ is continuous and right-differentiable in $[0, t_1)$ and*

$$\frac{d}{dt^+}M(t) = \max_{i=1,\dots,n}\left\{\frac{d}{dt^+}u(t)_i \colon u(t)_i = M(t)\right\}, \qquad t \in [0, t_1).$$

Proof. It is enough to prove the lemma when $n = 2$. Set $f := u_1$, $g := u_2$, and let $t \in [0, t_1)$. If $f(t) \neq g(t)$, the claim is trivial since $M(t)$ equals one of the two functions in a neighborhood of t. Suppose $f(t) = g(t) = M(t)$. If $\frac{d}{dt^+}f(t) > \frac{d}{dt^+}g(t)$, then for all $h > 0$ sufficiently small $M(t+h) = f(t+h)$; hence $\frac{d}{dt^+}M(t) = \frac{d}{dt^+}f(t)$. If $\frac{d}{dt^+}f(t) = \frac{d}{dt^+}g(t)$, then $M(t+h) - M(t)$ belongs to $[f(t+h) - f(t),\ g(t+h) - g(t)]$ if $f(t+h) \leq g(t+h)$ or to $[g(t+h) - g(t),\ f(t+h) - f(t)]$ if $f(t+h) \geq g(t+h)$. Hence $\frac{d}{dt^+}M(t) = \frac{d}{dt^+}f(t) = \frac{d}{dt^+}g(t)$. $\square$

LEMMA 5.2. *Let u be a real continuous right-differentiable function in an interval $[0, t_1)$. If $\frac{d}{dt^+}u \leq 0$ on $[0, t_1)$, then u is nonincreasing.*

Proof. See, for instance, [25, p. 298]. $\square$

LEMMA 5.3. *Let u be a real continuous right-differentiable function in an interval $[0, t_1)$, and let $g = |u|$. Then g is right-differentiable on $[0, t_1)$ and*

$$\frac{d}{dt^+}g(t) = \begin{cases} (\operatorname{sign} u(t))\frac{d}{dt^+}u(t) & \text{if } u(t) \neq 0, \\ |\frac{d}{dt^+}u(t)| & \text{if } u(t) = 0, \end{cases} \qquad t \in [0, t_1).$$

Proof. If $u(t) \neq 0$, the assertion is trivial, since g is right-differentiable at t. Suppose $u(t) = 0$. Given $h > 0$ we have $\frac{g(t+h)-g(t)}{h} = |\frac{u(t+h)}{h}|$. Being u right-differentiable at t we find that $\frac{d}{dt^+}g(t) = |\frac{d}{dt^+}u(t)|$. $\square$

THEOREM 5.4. *Let $N \in \mathbb{N}$ and $u_0 \in V_N$. Then there exists a unique function u_N such that*

(a) *$u_N : [0, +\infty) \to V_N$ is continuous and right-differentiable, and satisfies*

$$(5.1) \qquad \begin{cases} \frac{d}{dt^+}u_N(t) = A_{u_N(t)}u_N(t), \qquad t \in [0, +\infty), \\ u_N(0) = u_0. \end{cases}$$

In addition, u_N satisfies the following properties:

(b) *The set-valued map* $t \in [0,+\infty) \to \{\Psi_{u_N(t)} = 1\} \subseteq (0,1)$ *is nondecreasing, and the set-valued map* $t \in [0,+\infty) \to \#\partial\{\Psi_{u_N(t)} = 1\}$ *is nonincreasing. Moreover, for any* $t \geq 0$ *there exists* $\varepsilon > 0$ *such that* $\Psi_{u_N(\tau)}$ *is constant for any* $\tau \in [t, t+\varepsilon]$*. In particular,* $\frac{d}{dt^+}\Psi_{u_N(t)} = 0$ *for any* $t \geq 0$*.*

(c) *The function* $t \in [0,+\infty) \mapsto \sup_{x\in(0,1)} u_N(x,t)$ *is nonincreasing, and the function* $t \in [0,+\infty) \mapsto \inf_{x\in(0,1)} u_N(x,t)$ *is nondecreasing.*

(d) *The function* $t \in [0,+\infty) \mapsto \|\nabla u_N(t)\|_{L^1}$ *is nonincreasing.*

(e) *The function* $t \in [0,+\infty) \mapsto F_\phi(u_N(t))$ *is continuous and right-differentiable, and*

$$\frac{d}{dt^+} F_\phi(u_N(t)) = -\left\| \frac{d}{dt^+} u_N(t) \right\|_{L^2}^2 \leq 0. \tag{5.2}$$

(f) *There exist* $M \in \mathbb{N}$*,* $M \leq N$*, and positive times* $t_1, \dots, t_M$ *such that* u_N *is analytic on each interval of* $(0,+\infty)\setminus\{t_1,\dots,t_M\}$*, and* $\{t_1,\dots,t_M\}$ *coincides with the jump set of the function* $t \in [0,+\infty) \to \Psi_{u_N(t)}$*.*

Proof. Let $t_0 := 0$, and consider the function $u : [t_0,+\infty) \to V_N$,

$$u(t) = u_0 \exp((t - t_0)A_{u_0}), \qquad t \geq t_0, \tag{5.3}$$

i.e., the solution of

$$\begin{cases} \frac{d}{dt^+} u(t) = A_{u_0}\, u(t), & t \in (t_0, +\infty), \\ u(t_0) = u_0, \end{cases}$$

where we view the operator A_{u_0} as an $(N \times N)$-matrix.

For any $t \geq t_0$ let

$$\begin{aligned} \widetilde{M}(t) &:= \max\left(0, \max_{i=1,\dots,N} \{(D^+u(t))_i \colon (\Psi_{u_0})_i = 1\}\right), \\ \widetilde{m}(t) &:= \min\left(0, \min_{i=1,\dots,N} \{(D^+u(t))_i \colon (\Psi_{u_0})_i = 1\}\right). \end{aligned}$$

Observe that

$$-1 \leq \widetilde{m}(t_0) \leq \widetilde{M}(t_0) \leq 1. \tag{5.4}$$

In addition, the maps $t \in [t_0,+\infty) \to (D^+u(t))_i$ are continuously differentiable for any $i \in \{1,\dots,N\}$; hence, by Lemma 5.1, $\widetilde{M}(t)$ and $\widetilde{m}(t)$ are right-differentiable for any $t \geq t_0$.

Claim 1. For any $t \geq t_0$ we have

$$\frac{d}{dt^+}\widetilde{M}(t) \leq 0, \qquad \frac{d}{dt^+}\widetilde{m}(t) \geq 0. \tag{5.5}$$

Since D^+ is a linear operator, for all $t \geq t_0$ we have

$$\frac{d}{dt^+} D^+u(t) = D^+ \frac{d}{dt^+} u(t) = D^+ A_{u_0} u(t) = D^+ D^- \left(\Psi_{u_0} D^+ u(t)\right).$$

Therefore, if $i \in \{1, \dots, N\}$ is such that $(D^+u(t))_i = \widetilde{M}(t)$, we have

$$
\begin{aligned}
\frac{d}{dt^+}(D^+u)_i &= N\left[(D^-(\Psi_{u_0}D^+u))_{i+1} - (D^-(\Psi_{u_0}D^+u))_i\right] \\
&= N^2\left[(\Psi_{u_0})_{i+1}(D^+u)_{i+1} - (\Psi_{u_0})_i(D^+u)_i\right] \\
&\quad - N^2\left[(\Psi_{u_0})_i(D^+u)_i - (\Psi_{u_0})_{i-1}(D^+u)_{i-1}\right] \\
&= N^2\left[(\Psi_{u_0})_{i+1}(D^+u)_{i+1} - \widetilde{M}(t)\right. \\
&\quad \left. + (\Psi_{u_0})_{i-1}(D^+u)_{i-1} - \widetilde{M}(t)\right],
\end{aligned}
\tag{5.6}
$$

where both sides are evaluated at $t \geq t_0$. Since $(\Psi_{u_0}D^+u)_j \leq \widetilde{M}(t)$ for all $j \in \{1, \dots, N\}$, from the previous equation we obtain $\frac{d}{dt^+}(D^+u(t))_i \leq 0$ for all $i \in \{1, \dots, N\}$ such that $(\Psi_{u_0})_i = 1$ and $(D^+u(t))_i = \widetilde{M}(t)$. As a consequence we get

$$
0 \geq \max_{i=1,\dots,N}\left\{\frac{d}{dt^+}\left(D^+u(t)\right)_i : (\Psi_{u_0})_i = 1,\ (D^+u(t))_i = \widetilde{M}(t)\right\} = \frac{d}{dt^+}\widetilde{M}(t),
$$

where the last equality follows from Lemma 5.1. In a similar way we can prove that if $i \in \{1, \dots, N\}$ is such that $(\Psi_{u_0})_i = 1$ and $(D^+u(t))_i = \widetilde{m}(t)$, we have $\frac{d}{dt^+}(D^+u(t))_i \geq 0$; hence $\frac{d}{dt^+}\widetilde{m}(t) \geq 0$. This concludes the proof of Claim 1.

Claim 1 and Lemma 5.2 imply that $t \to \widetilde{M}(t)$ is nonincreasing and that $t \to \widetilde{m}(t)$ is nondecreasing. Recalling (5.4) we conclude that $-1 \leq \widetilde{m}(t) \leq \widetilde{M}(t) \leq 1$ for any $t \geq t_0$. Hence

$$
\Psi_{u(t)} = 1 \quad \text{at those nodes where} \quad \Psi_{u_0} = 1.
$$

It follows that the set-valued map $t \in [t_0, +\infty) \to \{|D^+u(t)| \leq 1\} = \sigma_G^\phi(u(t)) \subseteq (0,1)$ is nondecreasing.

Let us define

$$
t_1 := \sup\{t \geq t_0 \colon A_{u(s)} = A_{u_0} \quad \forall\, s \in [t_0, t)\}. \tag{5.7}
$$

We want to show that $t_1 > t_0$.

For all $i \in \{1, \dots, N\}$ such that $|(D^+u_0)_i| \leq 1$ we have $|(D^+u(t))_i| \leq 1$ for all $t \geq t_0$. In addition, $t \to D^+u(t)$ being a continuous function, if $|(D^+u_0)_i| > 1$, then there exists $\varepsilon > 0$ independent of i such that $|(D^+u(t))_i| > 1$ for any $t \in [t_0, t_0 + \varepsilon)$. Hence $\Psi_{u(t)} = \Psi_{u_0}$ for any $t \in [t_0, t_0 + \varepsilon]$. From Remark 4.6 it follows that $A_{u(t)} = A_{u_0}$ for any $t \in [t_0, t_0 + \varepsilon)$, which gives $t_1 \geq t_0 + \varepsilon > t_0$.

We have proven that the function $u(t)$ in (5.3) satisfies (5.1) for $t \in [t_0, t_1)$. We have also proven that either $t_1 = +\infty$ or $\Psi_{u(t_1)} \geq \Psi_{u(t_0)}$ and $(\Psi_{u(t_1)})_i > \Psi_{u(t_0)}$ for some $i \in \{1, \dots, N\}$.

If $t_1 < +\infty$, repeating the previous construction with t_1 in place of t_0 and $u(t_1)$ in place of u_0, we find a time $t_2 > t_1$ and a solution u of (5.1) defined in $[t_1, t_2)$ which satisfies (5.1). Repeating this argument, we can construct an increasing sequence (t_k) of times. Since at step k the number of nodes where $\Psi_{u(t)} = 1$ is nondecreasing, we can only have a finite number $M \leq N$ of steps, and in the last step we find that $t_M = +\infty$. Gluing together the solutions defined in the intervals $[t_k, t_{k+1})$ we find a function u_N defined for all $t \geq 0$ such that (a), (b), and (f) hold.

Let us prove (c), (d), and (e). Write for notational simplicity u in place of u_N. Let $t \in [0, +\infty)$. We say that $i \in \{1, \dots, N\}$ is a relative maximum (resp., minimum) for $u(t)$ if

$$
u(t)_i \geq \max\{u(t)_{i-1}, u(t)_{i+1}\} \qquad (\text{resp., } u(t)_i \leq \min\{u(t)_{i-1}, u(t)_{i+1}\}).
$$

Claim 2. Let $t \in [0, +\infty)$. If i is a relative maximum (resp., minimum) for $u(t)$, then $\frac{d}{dt^+} u(t)_i \le 0$ (resp., ≥ 0).

By (5.1),

$$
\begin{aligned}
\frac{d}{dt^+} u(t)_i &= N\Big[(\Psi_{u(t)})_i (D^+ u(t))_i - (\Psi_{u(t)})_{i-1} (D^+ u(t))_{i-1}\Big] \\
&= N^2\Big[(\Psi_{u(t)})_i (u(t)_{i+1} - u(t)_i) - (\Psi_{u(t)})_{i-1} (u(t)_i - u(t)_{i-1})\Big].
\end{aligned} \tag{5.8}
$$

Hence, if i is a relative maximum, we have $\frac{d}{dt^+} u(t)_i \le 0$ since $u(t)_{i+1} - u(t)_i \le 0$ and $u(t)_i - u(t)_{i-1} \ge 0$. Similarly, we can reason when i is a relative minimum, and Claim 2 follows.

Assertion (c) then follows from Claim 2.

Consider now the function

$$
S_i(t) := \begin{cases} \operatorname{sign}(u(t)_{i+1} - u(t)_i) & \text{if } u(t)_{i+1} \ne u(t)_i, \\ |\frac{d}{dt^+}(u(t)_{i+1} - u(t)_i| & \text{if } u(t)_{i+1} = u(t)_i. \end{cases}
$$

By Lemma 5.3 we have

$$
\begin{aligned}
\frac{d}{dt^+} \|\nabla u(t)\|_{L^1} &= \sum_{i=1}^N \frac{d}{dt^+} |u(t)_{i+1} - u(t)_i| = \sum_{i=1}^N S_i(t) \left(\frac{d}{dt^+} u(t)_{i+1} - \frac{d}{dt^+} u(t)_i \right) \\
&= \sum_{i=1}^N (S_{i-1}(t) - S_i(t)) \frac{d}{dt^+} u(t)_i.
\end{aligned}
$$

In order to prove that $\frac{d}{dt^+} \|\nabla u(t)\|_{L^1} \le 0$, it is enough to show that

$$
(S_{i-1}(t) - S_i(t)) \frac{d}{dt^+} u(t)_i \le 0 \qquad \forall\, i \in \{1, \dots, N\}. \tag{5.9}
$$

We divide the proof into four cases. We write for simplicity u in place of $u(t)$ and S in place of $S(t)$.

Case 1: the point i is simultaneously a relative maximum and a relative minimum, i.e., $u_{i-1} = u_i = u_{i+1}$. From (5.8) we deduce that $\frac{d}{dt^+} u_i = 0$, and (5.9) is satisfied.

Case 2: the point i is a relative maximum but not a relative minimum. Then either $u_i > u_{i-1}$ or $u_i > u_{i+1}$. So either $S_{i-1} = 1$ or $S_i = -1$, and in both cases $S_{i-1} - S_i \ge 0$. With $(D^+ u)_i \le 0$ and $(D^+ u)_{i-1} \ge 0$, from (5.8) we find that $\frac{d}{dt^+} u \le 0$, and (5.9) follows.

Case 3: the point i is a relative minimum but not a relative maximum. Then either $S_{i-1} = -1$ or $S_i = 1$, while $\frac{d}{dt^+} u \ge 0$.

Case 4: the point i is neither a relative maximum nor a relative minimum. Then either $u_{i-1} < u_i < u_{i+1}$ or $u_{i-1} > u_i > u_{i+1}$. In both cases we have $S_{i-1} = S_i$, and hence (5.9) holds.

Then (d) follows from Claim 1 and Lemma 5.2.

Let us now prove (e). Recalling that $\frac{d}{dt^+} \Psi_u = 0$ and using the expression of $F_\phi(u)$ as

$$
F_\phi(u) = \frac{1}{2} \int_{(0,1)} [\Psi_u (D^+ u)^2 + (1 - \Psi_u)]\, dx, \tag{5.10}
$$

we have

$$\begin{aligned}\frac{d}{dt^+}F_\phi(u) &= \frac{1}{2}\int_{(0,1)}\Psi_u\frac{d}{dt^+}(D^+u)^2\,dx = \int_{(0,1)}\Psi_u D^+u\frac{d}{dt^+}D^+u\,dx\\ &= \left\langle D^+\frac{d}{dt^+}u, \Psi_u D^+u\right\rangle = -\left\langle \frac{d}{dt^+}u, D^-\left(\Psi_u D^+u\right)\right\rangle\\ &= -\left\langle \frac{d}{dt^+}u, A_u u\right\rangle = -\int_{(0,1)}\left(\frac{d}{dt^+}u\right)^2 dx = -\left\|\frac{d}{dt^+}u\right\|_{L^2}^2 \le 0,\end{aligned}$$

which proves (5.2).

For all $t \geq 0$ for which $\Psi_{u(\cdot)}$ is continuous at t, the continuity of $F_\phi(u(\cdot))$ at t is a consequence of (5.10). On the other hand, if $(\Psi_{u(\cdot)})_i$ has a discontinuity at $\bar{t} \geq 0$, we know that there exists $\sigma > 0$ such that $(\Psi_u)_i = 0$ in $(\bar{t} - \sigma, \bar{t})$ and $(\Psi_u)_i = 1$ in $[\bar{t}, \bar{t} + \sigma)$. This implies that $|(D^+u)_i| > 1$ in $(\bar{t} - \sigma, \bar{t})$ and $|(D^+u)_i| \leq 1$ in $[\bar{t}, \bar{t} + \sigma)$. Since $(D^+u(\cdot))_i$ is continuous, we deduce that $(D^+u(\bar{t}))_i^2 = 1$. As a result,

$$\lim_{t \to \bar{t}^\pm} \Psi_{u(t)}(D^+u(t))^2 + (1 - \Psi_{u(t)}) = 1.$$

This implies the continuity of the map $t \mapsto F_\phi(u(t))$ at $\bar{t}$.

To conclude the proof of the theorem, we need to show that the function u_N is unique. The proof is divided into two steps.

Step 1. Let $\underline{u}_N : [0, +\infty) \to V_N$ be a continuous right-differentiable function satisfying (5.1). Assume, in addition, that for any $t \geq 0$ there exists $\varepsilon > 0$ such that $\Psi_{\underline{u}_N(\tau)}$ is constant for any $\tau \in [t, t+\varepsilon]$. Then $\underline{u}_N = u_N$.

Let $\varepsilon > 0$ be such that $\Psi_{\underline{u}_N}$ is constant on $[0, \varepsilon]$. It follows that $\underline{u}_N = u_N$ in $[0, \varepsilon]$, since the solution of (5.1), in $[0, \varepsilon]$, is uniquely given by (5.3). Without loss of generality, we can assume that

$$\varepsilon < t_1, \tag{5.11}$$

where t_1 is defined in (5.7) and is the first time for which Ψ_{u_N} is discontinuous. Recall that, by definition, $\{\Psi_{u_N(\varepsilon)} = 0\} = \{|D^+u_N(\varepsilon)| > 1\}$.

We claim that

$$\{\Psi_{u_N(\varepsilon)} = 1\} = \{|D^+u_N(\varepsilon)| < 1\}. \tag{5.12}$$

Indeed, denote by $I_j = (j/N, (j+1)/N)$ the generic interval of the grid and by $\sigma_j(t)$ the slope of $u_N(t)$ in I_j. A closer look at the last term in (5.6) reveals that for any $t \in [0, t_1)$, if

$$\widetilde{M}(t) = (D^+u(t))_i = 1, \text{ and either } \sigma_{i-1}(t) \neq 1 \text{ or } \sigma_{i+1}(t) \neq 1, \tag{5.13}$$

then

$$\frac{d}{dt^+}(D^+u(t))_i < 0, \tag{5.14}$$

where we recall that u stands for u_N. Similarly, if

$$\widetilde{m}(t) = (D^+u(t))_i = -1, \text{ and either } \sigma_{i-1}(t) \neq -1 \text{ or } \sigma_{i+1}(t) \neq -1, \tag{5.15}$$

then

$$\frac{d}{dt^+}(D^+u(t))_i > 0. \tag{5.16}$$

Observe that from (5.13), (5.14), (5.15), and (5.16), we already deduce that if $|\sigma_i(t)| = 1$ and if either $|\sigma_{i-1}(t)| \neq 1$ or $|\sigma_{i+1}(t)| \neq 1$, then $|\sigma_i(t+\tau)| < 1$ for any $\tau > 0$ small enough. What remains is the most delicate case; namely, we have to consider those intervals I_i of the grid where $|\sigma_i(t)| = 1$ and also $|\sigma_{i-1}(t)| = |\sigma_{i+1}(t)| = 1$. The following observation again follows from the expression on the right-hand side of (5.6). For any $t \in [0, t_1)$, if

$$\widetilde{M}(t) = (D^+u(t))_i = 1, \text{ and } \sigma_{i-1}(t) = 1 = \sigma_{i+1}(t), \tag{5.17}$$

then

$$\frac{d}{dt^+}(D^+u(t))_i = 0. \tag{5.18}$$

Similarly, if

$$\widetilde{m}(t) = (D^+u(t))_i = -1, \text{ and } \sigma_{i-1}(t) = -1 = \sigma_{i+1}(t), \tag{5.19}$$

then

$$\frac{d}{dt^+}(D^+u(t))_i = 0. \tag{5.20}$$

Hence (5.18) and (5.20) do not allow us to conclude that if $|\sigma_i(t)| = 1$ and if $|\sigma_{i-1}(t)| = 1 = |\sigma_{i+1}(t)|$, then $|\sigma_i(t+\tau)| < 1$ for any $\tau > 0$ small enough. However, such an inequality is valid and can be proved as follows. Let us denote by C the connected component of $\{\Psi_{u(t)} = 1\}$ containing I_i and by I_{i^-} (resp., I_{i^+}) the extremal left (resp., right) interval of the grid belonging to C (note that thanks to the boundary conditions, 0 is not a boundary point of I_{i^-} and 1 is not a boundary point of I_{i^+}). By (5.14) and (5.16) it follows that $|\sigma_{I_{i^-}}(t+\tau)| < 1$ and $|\sigma_{I_{i^+}}(t+\tau)| < 1$ for any $\tau > 0$. Using the previous arguments, we deduce that $|\sigma_{I_{i^-+1}}(t+\tau)| < 1$ and $|\sigma_{I_{i^+-1}}(t+\tau)| < 1$ for any $\tau > 0$ small enough. After a finite number of iterations, we deduce that $|\sigma_i(t+\tau)| < 1$ for any $\tau > 0$ small enough. This concludes the proof of the claim.

We can now repeat the reasoning taking ε as initial time, and we conclude that $\underline{u}_N = u_N$ in $[0, t_1]$. Iterating the argument for any $i = 1, \ldots, M$ we obtain that $\underline{u}_N = u_N$ in $[0, +\infty)$.

Step 2. Let $\underline{u}_N : [0, +\infty) \to V_N$ be a continuous right-differentiable function satisfying (5.1). Then for any $t \geq 0$ there exists $\varepsilon > 0$ such that $\Psi_{\underline{u}_N(\tau)}$ is constant for any $\tau \in [t, t+\varepsilon]$.

Let us consider an interval I_i where the slope $\underline{\sigma}_i(t)$ of $\underline{u}_N(t)$ satisfies $|\underline{\sigma}_i(t)| = 1$. Arguing as in Step 1, independently of the values of $|\underline{\sigma}_{i-1}(t)|$ and $|\underline{\sigma}_{i+1}(t)|$, we deduce that $|\underline{\sigma}_i(t+\tau)| < 1$ for any $\tau > 0$ sufficiently small. This implies that $\Psi_{\underline{u}_N(t)}$ is right continuous and proves Step 2.

Steps 1 and 2 conclude the proof of uniqueness, and hence the proof of the theorem. $\square$

Remark 5.5. We have already observed in the introduction that the right-hand side of the ODE's system $\dot{u} = -\nabla(F_{\phi_{|V_N}})$ (see (5.1)) is only a bounded function, since $F_{\phi_{|V_N}}$ is Lipschitz. Nevertheless, due to the special form of F_ϕ the solution in the sense of Theorem 5.4 is unique. This is not the case if we change the notion of solution to (5.1), for instance if we consider solutions to the system (5.1) only for almost all times. This is shown in the following example, which is related to the nonuniqueness example (Example 1 of section 2) and also shows another interesting phenomenon: the solution considered in Theorem 5.4 does not depend continuously on the initial datum.

Example 2. Assume that the initial datum $u_0 = u_{0N} \in V_N$ (with N even, in such a way that $1/2$ is a point of the mesh) is as follows:

$u_0 = 0$ in $(0, 1/2)$;

u_0 is increasing in $(1/2, 1/2 + 1/N)$ with slope exactly 1;

u_0 is piecewise linear, with slopes (in modulus) strictly less than 1 in $(1/2 + 1/N, 1)$.

Note that such an initial datum can be obtained from the discretization of solutions considered in the nonuniqueness example (Example 1 of section 2) at a time slightly smaller than t_* (and converging to t_* as $N \to +\infty$). The (unique) solution u_N of Theorem 5.4 is such that the linear part in the interval $(1/2, 1/2 + 1/N)$ for small positive times decreases its slope to a value less than 1. This solution, in the limit $N \to +\infty$, produces the solution $u_1(\cdot + t_*)$ of Example 1 of section 2.

Given $\varepsilon \in (0, 1)$ let us consider the functions $u_0^{\varepsilon\pm} = u_{0N}^{\varepsilon\pm} \in V_N$ defined as follows: $u_0^{\varepsilon\pm} := u_0$ in $(0, 1/2)$, $u_0^{\varepsilon\pm} := u_0 \pm \frac{\varepsilon}{N}$ in $(1/2 + 1/N, 1)$, and $u_0^{\varepsilon\pm}$ is increasing in $(1/2, 1/2 + 1/N)$ with slope $1 \pm \varepsilon$. Then, if $u_N^{\varepsilon\pm}$ denotes the solution of Theorem 5.4 having $u_0^{\varepsilon\pm}$ as initial datum, we have

$$\lim_{\varepsilon \to 0^+} u_N^{\varepsilon -} = u_N,$$

while

$$\lim_{\varepsilon \to 0^+} u_N^{\varepsilon +} = \widetilde{u}_N,$$

where $\widetilde{u}_N \in V_N$ satisfies (a) of Theorem 5.4 for any $t > 0$ but *not* for $t = 0$, and

$$\lim_{N \to +\infty} \widetilde{u}_N(\cdot) = u_2(\cdot + t_*),$$

where u_2 is as in Example 1 of section 2. Hence the solution u_N of Theorem 5.4 is not continuous with respect to initial data. We can summarize the above discussion, coupled with the remarks of section 2, with the following conclusion: solutions to (iv), (v), and (vi) of Theorem 2.4 are not unique thanks to Example 1 of section 2 (which, however, we believe to be nongeneric). On the other hand, solutions of Theorem 5.4 are unique; however, they do not depend in a continuous way on the initial data. It is such an instability at the discrete level (i.e., for fixed N) which seems to produce nonuniqueness in the limit $N \to +\infty$.

6. Convergence of the approximating schemes. In this section we prove Theorem 2.4. We begin with the following elementary lemma.

LEMMA 6.1. *Let* $u_0 \in \mathcal{A}_\phi(0, 1)$. *Then there exists a sequence* $(u_0^N) \subset V_N$ *of functions satisfying assertion* (i) *of Theorem* 2.4.

Proof. Define $u_0^N \in V_N$ as $(u_0^N)_i := u_0(i/N)$. Then $\|u_0^N\|_{BV(0,1)} \le \|u_0\|_{BV(0,1)}$ for any $N \in \mathbb{N}$, (u_0^N) converges to u_0 weakly* in $BV(0, 1)$ and strongly in $L^2(0, 1)$, and $\lim_{N \to +\infty} \|u_0^N\|_{BV(0,1)} = \|u_0\|_{BV(0,1)}$. Note that for any $x \in [0, 1]$ such that $\mathrm{dist}(x, \sigma_B^\phi(u_0)) > 1/N$ (resp., $\mathrm{dist}(x, \sigma_G^\phi(u_0)) > 1/N$), then $x \in \sigma_G^\phi(u_0^N)$ (resp., $x \in$

$\sigma_B^\phi(u_0^N))$. It follows that $\lim_{N\to+\infty} d_{\mathcal{H}}(\sigma_G^\phi(u_0^N), \sigma_G^\phi(u_0)) = 0$. Since any isolated point in $\sigma_B^\phi(u)$ belongs to $\sigma_B^\phi(u_0^N)$ for N large enough, we also have $d_{\mathcal{H}}(\sigma_B^\phi(u_0^N), \sigma_B^\phi(u_0)) \to 0$ as $N \to +\infty$.

Now let $K \subset \sigma_G^\phi(u_0)$ be an interval with $\overline{K} \subset (0,1)$. Then $\|u_0^N\|_{L^2(K)} \leq \|u_0\|_{L^2(K_N)}$, where $K_N := \{x \in \mathbb{R} : \operatorname{dist}(x,K) < 1/N\}$ and N is large enough in such a way that $K_N \subset (0,1)$. Hence $\|u_0^N\|_{L^2(K)} \leq \|u_0\|_{L^2(K)} + \frac{2}{N}$, (u_0^N) weakly converges to u_0 in $H^1(K)$, and $\|u_0^N{}_x\|_{L^2(K)}$ converges to $\|u_{0x}\|_{L^2(K)}$. Therefore $\lim_{N\to+\infty} F_\phi(u_0^N) = F_\phi(u_0)$, and this concludes the proof. ∎

By construction, $u_0^N \in V_N \subset \mathcal{A}_\phi(0,1)$; moreover, we can assume that if N is large enough, the number of connected components of $\sigma_B^\phi(u_0^N)$ equals m, the number of connected components of $\sigma_B^\phi(u_0)$, and we can uniquely write $\sigma_B^\phi(u_0^N)$ as in (2.2).

DEFINITION 6.2. *Let $u_0 \in \mathcal{A}_\phi(0,1)$, and let (u_0^N) be as in Lemma* 6.1. *We denote by $u^N\colon [0,+\infty) \to V_N$ the solution of*

$$\begin{cases} \dfrac{d}{dt^+}u(t) = A_{u(t)}u(t), & t \in (0,+\infty), \\ u^N(0) = u_0^N \end{cases} \tag{6.1}$$

given by Theorem 5.4 *(with u_0 in* (5.1) *replaced by u_0^N).*

Note that all assertions in Theorem 2.4(ii) are satisfied.

Remark 6.3.

(a) For any $j \in \{1,\dots,m\}$ we define

$$T_j^N := \sup\left\{t \geq 0\colon \sigma_B^\phi(u^N(t)) \cap [a_j^N(0), b_j^N(0)] \neq \emptyset\right\} > 0,$$
$$[a_j^N(t), b_j^N(t)] := \sigma_B^\phi(u^N(t)) \cap [a_j^N(0), b_j^N(0)], \qquad t \in [0, T_j^N).$$

Then $a_j^N(0) = a_j^0$, $b_j^N(0) = b_j^0$, and

$$\sigma_B^\phi(u^N(t)) = \bigcup_{j=1}^m [a_j^N(t), b_j^N(t)], \qquad t \in [0,+\infty),$$

where we have set

$$[a_j^N(t), b_j^N(t)] := \emptyset \qquad \text{if } t \geq T_j^N.$$

(b) The map $t \in [0,T_j^N) \mapsto a_j^N(t)$ is continuous and nondecreasing, and the map $t \in [0,T_j^N) \mapsto b_j^N(t)$ is continuous and nonincreasing.

(c) Since $u^N(\cdot,t) = u_0^N(\cdot)$ on $\sigma_B^\phi(u^N(t))$, for any $j \in \{1,\dots,m\}$ we have that either $u_0^N{}_x(x,t) > 1$ for a.e. $x \in [a_j^N(t), b_j^N(t)]$ or $u_0^N{}_x(x,t) < -1$ for a.e. $x \in [a_j^N(t), b_j^N(t)]$.

LEMMA 6.4. *There exists a constant $C > 0$ depending only on u_0 such that*

$$\sup_{t>0}\sup_{N\in\mathbb{N}} F_\phi(u^N(t)) \leq C,$$
$$\sup_{N\in\mathbb{N}} \left\|\frac{d}{dt^+}u^N\right\|_{L^2((0,+\infty);L^2(0,1))} \leq C,$$
$$\sup_{N\in\mathbb{N}} \|u^N\|_{L^\infty((0,+\infty);BV(0,1))} \leq C.$$

Proof. The first two inequalities follow from (2.3) and (5.2). The last one follows from Theorem 5.4 (c) and (d) and (2.3). ∎

Remark 6.5. Thanks to Lemma 6.4 (and extracting if necessary a not relabelled subsequence) the sequence (u^N) converges weakly in $H^1_{\mathrm{loc}}((0,+\infty);L^2(0,1))$ and weakly* in $L^\infty((0,+\infty);BV(0,1))$ to a function u as $N\to+\infty$, and this gives assertion (iii) of Theorem 2.4. In particular, for almost every $x\in(0,1)$ the function $t\in[0,+\infty)\to u(x,t)$ is continuous, and $u^N(x,\cdot)\to u(x,\cdot)$ uniformly on $[0,+\infty)$. As a consequence the function $u(\cdot,t)$ is well defined for all $t\in[0,+\infty)$ and $\|Du(\cdot,t)\|\le\|Du_0\|$. It also follows that $u^N(t)\to u(t)$ weakly* in $BV(0,1)$ for almost every $t\ge 0$.

Remark 6.6. Possibly extracting a further subsequence, we can assume that for any $j\in\{1,\dots,m\}$, $T_j^N\to T_j$ as $N\to+\infty$ for some $T_j\in[0,+\infty]$. If $T_j>0$, since the functions $a_j^N(\cdot)$ (resp., $b_j^N(\cdot)$) are nondecreasing (resp., nonincreasing), there exist nondecreasing functions $a_j:[0,T_j)\to[0,1]$ (resp., nonincreasing functions $b_j:[0,T_j)\to[0,1]$) such that $a_j^N\to a_j$ (resp., $b_j^N\to b_j$) weakly* in $BV(0,T_j-\varepsilon)$ as $N\to+\infty$ for all $\varepsilon>0$ small enough. Since $a_j^N(t)<b_j^N(t)$ for all $t\in[0,T_j^N)$, passing to the limit we obtain that $a_j(t)\le b_j(t)$ for all $t\in[0,T_j)$. Recall that $a_j(0)=a_j^0$ and $b_j(0)=b_j^0$ for any $j\in\{1,\dots,m\}$.

In the following, set $\mathcal{J}(0):=\{1,\dots,m\}$.

DEFINITION 6.7. *For any $t\in[0,+\infty)$ we define*

$$
\begin{aligned}
\mathcal{J}(t) &:= \{j\in\{1,\dots,m\}: t<T_j\},\\
B(t) &:= \bigcup_{j\in\mathcal{J}(t)}[a_j(t),b_j(t)],\\
G(t) &:= [0,1]\setminus B(t),\\
\widetilde{B}(t) &:= \bigcup_{j\in\mathcal{J}(t):\,a_j(t)<b_j(t)}[a_j(t),b_j(t)]\cup\bigcup_{j\in\mathcal{J}(t):\,a_j(t)=b_j(t)\in J_{u(t)}}\{a_j(t)\},\\
\widetilde{G}(t) &:= [0,1]\setminus\widetilde{B}(t).
\end{aligned}
$$

Note that

$$
\overline{\mathrm{int}(B(t))}\subseteq\widetilde{B}(t)\subseteq B(t). \tag{6.2}
$$

LEMMA 6.8. *For any $j\in\{1,\dots,m\}$ we have $T_j>0$, and the functions a_j and b_j are continuous on $[0,T_j)$.*

Proof. Assume by contradiction that there exists $j\in\{1,\dots,m\}$ such that $T_j=0$. Then $[a_j^0,b_j^0]\in\sigma_G^\phi(u(s))$ for any $s>0$. Hence $u(s)$ is one-Lipschitz in $[a_j^0,b_j^0]$ for any $s>0$.

Case 1. Assume that $a_j^0<b_j^0$. Using the triangular property and $u(0)=u_0$, for any $x,x'\in[a_j^0,b_j^0]$, $x\neq x'$, we have

$$
\begin{aligned}
|u(x,s)-u(x,0)|+|u(x',s)-u(x',0)| &\ge |u_0(x)-u_0(x')|-|u(x,s)-u(x',s)|\\
&\ge |u_0(x)-u_0(x')|-|x-x'|>0.
\end{aligned}
$$

This means that $s\mapsto u(x,s)$ has a discontinuity at $s=0$ for a.e. $x\in[a_j^0,b_j^0]$, and this is in contradiction with $u\in AC^2([0,+\infty);L^2(0,1))$.

Case 2. Assume that $a_j^0=b_j^0$. Let $L:=u_0(a^0_{j+})$ and $l:=u_0(a^0_{j-})$. We can assume $l<L$. Let $\delta:=\min(\frac{(L-l)}{4},a_1^0,(1-b_m^0),\min_{j=1,\dots,m-1}(a_{j+1}^0-b_j^0))>0$, and define $x^{\pm} := a_j^0 \pm \delta$. Note that $u(s)$ is one-Lipschitz in (x^-, x^+) for any $s > 0$. For any $x, x' \in (x^-, x^+)$, $x \neq x'$, we have

$$\begin{aligned}|u(x,s) - u(x,0)| + |u(x',s) - u(x',0)| &\geq |u_0(x) - u_0(x')| - |x - x'| \\ &\geq |u_0(a_{j\,-}^0) - u_0(a_{j\,+}^0)| - |u_0(x) - u_0(a_{j\,-}^0)| \\ &\quad - |u_0(x') - u_0(a_{j\,+}^0)| - |x - x'| \\ &\geq L - l - 4\delta > 0.\end{aligned}$$

As above, this is in contradiction with $u \in AC^2([0, +\infty); L^2(0, 1))$.

Let us now prove that a_j and b_j are continuous. Assume by contradiction that a_j has a discontinuity at $t = \bar{t} \in [0, T_j)$. Since a_j is nondecreasing, $\bar{t}$ is a jump point of a_j. If $\bar{t} = 0$, we can argue in analogy to Case 1. Assume $\bar{t} > 0$, and let $x^- := \lim_{t \to \bar{t}^-} a_j(t) < x^+ := \lim_{t \to \bar{t}^+} a_j(t)$. Since $u^N(\cdot, t)$ coincides with $u_0^N(\cdot)$ in $\sigma_B^\phi(u^N(t))$, it follows that $u(\cdot, t)$ coincides with $u_0(\cdot)$ in each connected component of $\mathrm{int}(B(t))$. In particular, the function $u(t)$ coincides with u_0 in (x^-, x^+) for all $t \in [0, \bar{t})$. We then obtain

$$|u(x,t) - u(x',t)| = |u_0(x) - u_0(x')| > |x - x'| \qquad \forall\ x, x' \in (x^-, x^+).$$

On the other hand, $u(s)$ is one-Lipschitz in (x^-, x^+) for any $s > \bar{t}$. It follows that, for any $x, x' \in (x^-, x^+)$,

$$|u(x,t) - u(x,s)| + |u(x',t) - u(x',s)| \geq |u_0(x) - u_0(x')| - |x - x'| > 0,$$

which contradicts $u \in AC^2([0, +\infty); L^2(0, 1))$. This proves the continuity of a_j. The continuity of b_j follows using a similar argument. ∎

Remark 6.9. Whenever $T_j < +\infty$, arguing as in Lemma 6.8 with $\bar{t} = T_j$, we get $\lim_{t \to T_j^-} a_j(t) = \lim_{t \to T_j^-} b_j(t)$.

Remark 6.10.

(a) Since $u^N(\cdot, t)$ is one-Lipschitz in each connected component of $\sigma_G^\phi(u^N(t))$, it follows that $u(\cdot, t)$ is one-Lipschitz in each connected component of $\widetilde{G}(t)$.

(b) The function $u(\cdot, t)$ coincides with $u_0(\cdot)$ in each connected component of $\mathrm{int}(B(t))$.

Remark 6.11. As a consequence of Lemma 6.8 the sequence (a_j^N) (resp., (b_j^N)) converges to a_j (resp., to b_j) uniformly in $[0, T_j - \varepsilon)$ as $N \to +\infty$ for any $\varepsilon > 0$ small enough. In particular, for any connected component I of $B(t)$ there exists a connected component I_N of $\sigma_B^\phi(u^N(t))$ such that

$$\lim_{N \to +\infty} d_{\mathcal{H}}(I_N, I) = 0.$$

LEMMA 6.12. *The function $u(t)$ is ϕ-admissible for any $t \geq 0$ and*

$$\overline{\mathrm{int}(B(t))} \subseteq \sigma_B^\phi(u(t)) \subseteq B(t) \qquad \forall\ t \in [0, +\infty). \tag{6.3}$$

Proof. Recalling Remark 6.5, let us fix $t \geq 0$ such that $u^N(t) \to u(t)$ weakly* in $BV(0, 1)$. From Remark 6.10(a) it follows that $u(t)$ is one-Lipschitz in each connected component of $\widetilde{G}(t)$; hence

$$\widetilde{G}(t) \subseteq \sigma_G^\phi(u(t)).$$

Moreover, from Remarks 6.3(c) and 6.11 it follows that the assertion in Remark 2.2(c) holds with u replaced by $u(t)$ for any connected component I of $\widetilde{B}(t)$ and any Borel set $A \subseteq I$. Indeed, if A is compactly contained in I, then $Du^N(A) = Du_0^N(A)$ for $N \in \mathbb{N}$ large enough, and by construction (see Lemma 6.1) in the first case $|A| < \lim_{N\to+\infty} Du_0^N(A) = Du(A)$ or in the second case $-|A| > \lim_{N\to+\infty} Du_0^N(A) = Du(A)$. If A is a boundary point of I, then (using Remarks 6.11 and 6.10(a)) in the first case $0 \le \lim_{N\to+\infty} Du^N(A) = Du(A)$ or in the second case $0 \le \lim_{N\to+\infty} Du^N(A) = Du(A)$. To obtain the desired inequalities when A is a generic Borel set in I, it is enough to write $A = (A \cap \operatorname{int}(I)) \cup (A \cap \partial I)$, to approximate $A \cap \operatorname{int}(I)$ with a sequence of subsets of A compactly contained in I, and to use the previous arguments. It follows that

$$\widetilde{B}(t) \subseteq \sigma_B^\phi(u(t)).$$

In particular, for almost every $t \ge 0$, $u(t)$ is ϕ-admissible, $\sigma_B^\phi(u(t)) = \widetilde{B}(t)$, and (6.3) follows from (6.2).

Assume now that $t \ge 0$ is generic, and pick a sequence $(t_n) \subset (0, +\infty)$ converging to t as $n \to +\infty$ such that $u(t_n) \in \mathcal{A}_\phi(0,1)$ and for which (6.3) holds with t_n in place of t. Since $u \in AC^2([0,+\infty); L^2(0,1))$ and $u(t) \in BV(0,1)$, we have $u(t_n) \to u(t)$ weakly* in $BV(0,1)$ as $n \to +\infty$. It is then enough to repeat the previous arguments, and the assertion follows. ◻

Remark 6.13.

(a) We have $\lim_{N\to+\infty} d_{\mathcal{H}}(\Gamma_{u^N}, \Gamma_u) = 0$, where $\Gamma_{u^N} := \bigcup_{t\in(0,+\infty)} (\sigma_G^\phi(u^N(t)) \times \{t\})$. In particular, by Lemma 6.8, for all $t \in [0, +\infty)$ we have

$$\begin{aligned} &\lim_{N\to+\infty} d_{\mathcal{H}}\big(\sigma_G^\phi(u^N(t)), \sigma_G^\phi(u(t))\big) = 0, \\ &\lim_{N\to+\infty} d_{\mathcal{H}}\big(\operatorname{int}(\sigma_B^\phi(u^N(t))), \operatorname{int}(\sigma_B^\phi(u(t)))\big) = 0. \end{aligned} \tag{6.4}$$

(b) Since $u^N \rightharpoonup u$ weakly* in $L^\infty([0,+\infty); BV(0,1))$ and $u^N \equiv u_0^N$ in $[0,1] \times [0,+\infty) \setminus \overline{\Gamma_{u^N}}$ by Remark 6.3(c), we have $u \equiv u_0$ in $[0,1] \times [0,+\infty) \setminus \overline{\Gamma_u}$.

THEOREM 6.14. *The function u satisfies $u_{xx} \in L^2(\Gamma_u)$ and is a solution of*

$$\begin{cases} u_t = u_{xx}, & x \in \sigma_G^\phi(u(t)),\ t \in (0,+\infty), \\ u_t = 0, & x \in \operatorname{int}(\sigma_B^\phi(u(t))),\ t \in (0,+\infty), \\ \displaystyle\lim_{y\to x,\, y\in\sigma_G^\phi(u(t))} u_x(y,t) = 0, & x \in \partial\sigma_G^\phi(u(t)) \setminus \{0,1\},\ t \in (0,+\infty), \\ u(x,0) = u_0(x), & x \in (0,1), \\ u(0,t) = u(1,t),\ u_x(0,t) = u_x(1,t), & t \in (0,+\infty). \end{cases} \tag{6.5}$$

Proof. Let $\psi \in \mathcal{C}_c^1([0,+\infty) \times [0,1])$, and let $\psi^N : [0,+\infty) \to V_N$, $\psi^N \in \operatorname{Lip}_c([0,+\infty) \times [0,1])$, be such that $\psi^N(t) \to \psi(t)$ in $H^1(0,1)$ for any $t \ge 0$. We have

$$
\begin{aligned}
\mathrm{I}_N(t) := \int_{\sigma_G^\phi(u^N(t))} u_x^N(t)\psi_x^N(t)\,dx &= \sum_{i:\,(\Psi_{u^N(t)})_i=1} \frac{D^+u_i^N(t)\,D^+\psi_i^N(t)}{N} \\
&= -\int_{(0,1)} D^-(\Psi_{u^N(t)}D^+u^N(t))\psi^N(t)\,dx \\
&= -\int_{(0,1)} A_{u^N(t)}u^N(t)\psi^N(t)\,dx \\
&= -\int_{(0,1)} \frac{d}{dt^+}u^N(t)\psi^N(t)\,dx =: \mathrm{II}_N(t).
\end{aligned} \tag{6.6}
$$

From (6.4) (which is valid for any $t \geq 0$ thanks to Lemma 6.8) and from the weak $H^1_{\mathrm{loc}}(\mathrm{int}(\Gamma_u))$-convergence of (u^N) to u, using (6.3) it follows that

$$
\lim_{N\to+\infty} \mathrm{I}_N(t) = \int_{\sigma_G^\phi(u(t))} u_x(t)\psi_x(t)\,dx \qquad \text{for a.e. } t \geq 0. \tag{6.7}
$$

On the other hand, $\frac{d}{dt^+}u^N \rightharpoonup \frac{d}{dt^+}u$ in $L^2((0,1)\times(0,+\infty))$ as $N \to +\infty$; hence

$$
\lim_{N\to+\infty} \mathrm{II}_N(t) = \int_{(0,1)} \frac{d}{dt^+}u(t)\psi(t)\,dx \qquad \text{for a.e. } t \geq 0. \tag{6.8}
$$

Recalling also Remark 6.13(b), equalities (6.7), (6.8) coupled with (6.6) imply that u solves the problem

$$
\begin{cases}
u_t = u_{xx} & \text{in } \mathrm{int}(\Gamma_u),\\
u_t = 0 & \text{in } [0,1]\times[0,+\infty)\setminus\overline{\Gamma_u},\\
u(0) = u_0 & \text{in } [0,1]\times\{0\}.
\end{cases} \tag{6.9}
$$

In particular, we have $u \in \mathcal{C}^\infty(\mathrm{int}(\Gamma_u))$. Moreover, since $u_t \in L^2((0,1)\times(0,+\infty))$, we also get $u_{xx} \in L^2(\Gamma_u)$. It then follows that there exists the limit

$$
\lim_{x\to\bar{x},\, x\in\sigma_G^\phi(u(t))} u_x(x,t) = 0 \qquad \text{for a.e. } t \geq 0,\ \bar{x} \in \partial\sigma_G^\phi(u(t)); \tag{6.10}
$$

i.e., $u|_{\mathrm{int}(\Gamma_u)}$ satisfies zero Neumann boundary conditions on $\partial\Gamma_u$. Problem (6.9), together with the boundary condition (6.10), is equivalent to problem (6.5).

The periodic boundary conditions are a consequence of u being ϕ-admissible. ◻

Remark 6.15. The same results of Theorem 2.4 hold if we replace in the definition (1.1) of ϕ the function ξ^2 with a function $f \in \mathcal{C}^\infty(\mathbb{R})$ which satisfies $f(0) = 0$, $f(1) = 1$, $f(\xi) = f(-\xi)$, and $f''(\xi) > 0$ for all $\xi \in (-1,1)$. It is clear that the equation $u_t = u_{xx}$ in (2.5) is replaced by $u_t = \frac{1}{2}f''(u_x)u_{xx}$.

Remark 6.16. Let $N \in \mathbb{N}$, and set $\phi^N(\xi) := \min(\xi^2, N)$ for any $\xi \in \mathbb{R}$. Define the functional $F_{\phi^N,N} : L^1(0,1) \to [0,+\infty]$ as

$$
F_{\phi^N,N}(v) := \frac{1}{2N}\sum_{i=1}^N \min\left(((D^+v)_i)^2, N\right), \qquad v \in V_N
$$

(and extended to $+\infty$ elsewhere). In [14] it is proved that the sequence $(F_{\phi^N,N})$ Γ-converges, as $N \to +\infty$, to the Mumford–Shah functional. Let $\overline{u} \in BV(0,1)$, with

$\overline{u}(0) = \overline{u}(1)$, having a finite set $\overline{x}_1, \ldots, \overline{x}_n$ of jump points in $(0,1)$, and of class $\mathcal{C}^1(\overline{I})$, for any interval $I \subset (0,1) \setminus \{\overline{x}_1, \ldots, \overline{x}_n\}$. Then $\overline{u}$ is ϕ^N-admissible for N large enough; i.e., $\overline{u}$ satisfies Definition 2.1, where (1) is replaced by $|\overline{u}(x) - \overline{u}(y)| \leq \sqrt{N}|x-y|$ whenever $[x,y] \subset \sigma_G^{\phi^N}(\overline{u})$, and where the inequality involving u in (3) is replaced by $|\overline{u}(x) - \overline{u}(y)| > \sqrt{N}|x-y|$. Let us consider the solutions ω^N to the rescaled gradient flow system of ODEs

$$\begin{cases} \omega_t^N = -N\nabla(F_{\phi^N, N|V_N})(\omega^N), \\ \omega^N(0) = \overline{u}^N, \end{cases} \tag{6.11}$$

$\overline{u}^N$ as in Lemma 6.1. Reasoning as in Theorem 2.4 we get that, as $N \to +\infty$, the sequence (ω^N) converges, up to a subsequence, to a function ω which satisfies the heat equation with zero Neumann interior conditions on each interval of $(0,1) \setminus \{\overline{x}_1, \ldots, \overline{x}_n\}$ (except in $\{0,1\}$), has periodic conditions in $\{0,1\}$, and keeps the points $\overline{x}_1, \ldots, \overline{x}_n$ fixed in time ($\overline{x}_j$ may disappear at time $\overline{t}_j < +\infty$ if $\lim_{x \to \overline{x}_{j_-}} \omega(x, \overline{t}_j) = \lim_{x \to \overline{x}_{j_+}} \omega(x, \overline{t}_j)$). Therefore ω can be considered as a reasonable global solution to the gradient flow of the Mumford–Shah functional in one dimension starting from $\overline{u}$ (compare [22], [20]).

7. Numerical simulations. In this section we show a numerical simulation which confirms the behaviors predicted by Theorem 2.4. Let $u_0 \in \mathcal{A}_\phi(0,1)$ be the upper graph in Figure 7.2; see also Figure 7.1. We have

$$\sigma_B^\phi(u_0) = [a_1^0, b_1^0] \cup [a_2^0, b_2^0] \cup [a_3^0, b_3^0],$$

where $a_1^0 = 0.05$, $b_1^0 = 0.2$, $a_2^0 = b_2^0 = 0.6$, $a_3^0 = 0.9$, and $b_3^0 = 0.99$. Note that $J_{u_0} = \{a_2^0, a_3^0\}$.

The sequence of graphs displayed in Figures 7.1 and 7.2 presents the solution u starting from u_0 at subsequent times. The computation solves the discrete evolution presented in section 5 with space discretization $\Delta x = 1/N$ with $N = 500$. The algorithm used is a forward Euler scheme with time step $\Delta t = (\Delta x)^2/10$. Let us list the main features of the computed evolution u, all of which are in accordance with Theorem 2.4.

(1) We have $a_1(t) \equiv a_1^0$ for all $t > 0$, and on the interval $(0, a_1(t))$ the solution u evolves according to the heat equation with zero Neumann boundary condition at $a_1(t)$. In addition,

$$a_1(t) \in J_{u(t)} \qquad \forall\ t > 0.$$

Since $a_1^0 \notin J_{u_0}$, $a_1(t)$ "instantly" becomes a discontinuity point of the solution; see also Figure 7.1.

(2) The function $t \to b_1(t)$ is decreasing for positive times. The interval $[a_1(t), b_1(t)]$ is gradually eroded, from the right, by the interval $[b_1(t), a_2(t)]$, where the solution evolves according to the heat equation, with zero Neumann boundary conditions.

(3) There exists $T_2 > 0$ such that $a_2(t) \equiv a_2^0$ and $a_2^0 \in J_{u(t)}$ for $t \in [0, T_2)$, and then a_2^0 becomes a continuity point of $u(t)$ for $t \geq T_2$. In the region $[b_2^0, a_3(t)]$, for all times $t \in (0, T_2)$, the solution evolves according to the heat equation with zero Neumann boundary conditions at b_2^0 and $a_3(t)$. Note that

$$\sigma_B^\phi(u(t)) = [a_1(t), b_1(t)] \cup [a_3(t), b_3(t)], \qquad t \geq T_2,$$

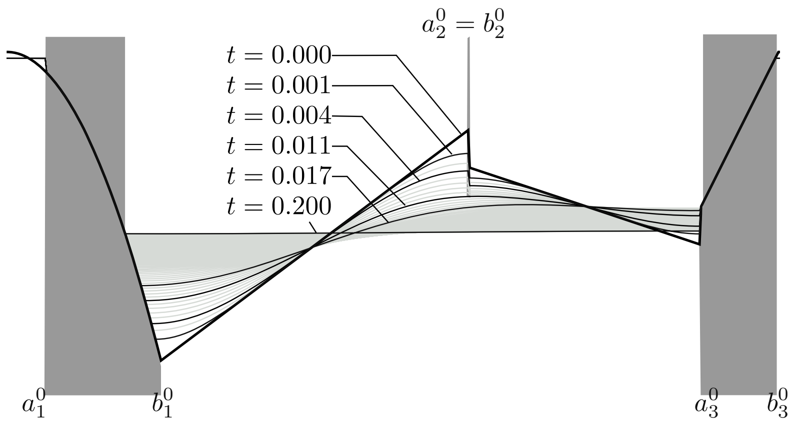


FIG. 7.1. *A simulation of the discretized evolution. The function is plotted in black for some relevant time values. The initial datum u_0 is plotted thick. The gray regions represent the intervals $[a_j(t), b_j(t)]$.*

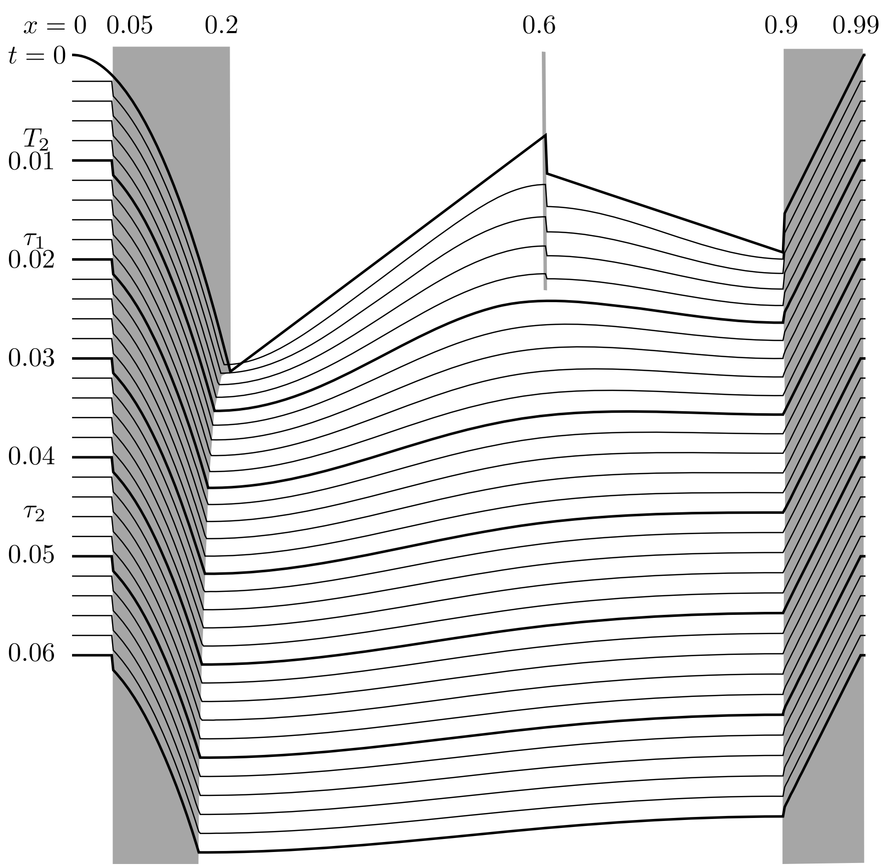


FIG. 7.2. *A vertical translation has been added to the evolution to distinguish the functions.*

and u evolves accordingly to the heat equation in the interval $(b_1(t), a_3(t))$ with zero Neumann boundary conditions.

(4) There exist two positive times $0 < \tau_1 < \tau_2$ such that $a_3(t) \equiv a_3^0 \in J_{u(t)}$ for $t \in [0, \tau_1)$, the point $a_3(t)$ becomes a continuity point of u for $t \in [\tau_1, \tau_2]$, and the function $t \to a_3(t)$ is strictly increasing in that interval, $a_3(t) \equiv a_3(\tau_2) \in J_{u(t)}$ for all $t > \tau_2$. The function $t \to b_3(t)$ is strictly decreasing.

(5) On the interval $(b_3(t), 1)$ the solution u evolves according to the heat equation with zero Neumann boundary conditions for all $t > 0$.

Remark 7.1. We conclude the paper by observing that, for energy densities different from (1.1), in particular for the function ϕ_1 considered in Figure 3.1 (the nonconvex region of which is bounded), the discrete approximation scheme discussed in sections 5 and 6, which keeps *fixed in time* the nodes of the mesh in $(0, 1)$, could converge to functions $\widetilde{u}$ which are not solutions to (3.6). In particular, the functions $\widetilde{u}$ might not satisfy the condition $\widetilde{u}_t = 0$ in $\mathrm{int}(\sigma_B^{\phi_1}(\widetilde{u}(t)))$; see also the comments in [21, p. 590]. This behavior of $\widetilde{u}$, which is related to the interactions of the nonconvex region of ϕ_1 with the numerical scheme with fixed nodes, deserves further investigation.

Acknowledgments. The authors gratefully acknowledge the hospitality and the support of the Centro De Giorgi of the Scuola Normale Superiore di Pisa, where this paper was partially written, and of the Max Planck Institute in the Sciences in Leipzig, where this paper was concluded. The authors are also grateful to the referees for several interesting comments.

REFERENCES

[1] G. Alberti and S. Müller, *A new approach to variational problems with multiple scales*, Comm. Pure Appl. Math., 54 (2001), pp. 761–825.

[2] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Lectures Math. ETH Zurich, Birkhäuser Verlag, Basel, 2005.

[3] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Clarendon Press, Oxford, UK, 2000.

[4] F. Andreu-Vaillo, V. Caselles, and J. M. Mazón, *Parabolic Quasilinear Equations Minimizing Linear Growth Functionals*, Birkhäuser Verlag, Basel, 2004.

[5] G. I. Barenblatt, M. Bertsch, R. Dal Passo, and M. Ughi, *A degenerate pseudoparabolic regularization of a nonlinear forward-backward heat equation arising in the theory of heat and mass exchange in stably stratified turbulent shear flow*, SIAM J. Math. Anal., 24 (1993), pp. 1414–1439.

[6] G. I. Barenblatt, M. Bertsch, R. Dal Passo, V. M. Prostokishin, and M. Ughi, *A mathematical model of turbulent heat and mass transfer in stably stratified shear flow*, J. Fluid Mech., 253 (1993), pp. 341–358.

[7] G. Bellettini and G. Fusco, *A regularized Perona-Malik functional: Some aspects of the gradient dynamics*, in Proceedings of the International Conference on Differential Equations, Hasselt, Belgium, 2003, pp. 639–644.

[8] G. Bellettini and G. Fusco, *The Γ-Limit and the Related Gradient Flow for Singular Perturbation Functionals of Perona-Malik Type*, Preprint Centro De Giorgi, Scuola Normale Superiore di Pisa, Pisa, Italy, 2004, submitted.

[9] G. Bellettini, G. Fusco, and N. Guglielmi, *A Concept of Solution for Forward-Backward Equations of the Form $u_t = \frac{1}{2}(\phi'(u_x))_x$ and Numerical Experiments for the Singular Perturbation $u_t = -\varepsilon^2 u_{xxxx} + \frac{1}{2}(\phi'(u_x))_x$*, Preprint Centro De Giorgi, Scuola Normale Superiore di Pisa, Pisa, Italy, 2005, submitted.

[10] G. Bellettini and M. Novaga, *Minimal barriers for geometric evolutions*, J. Differential Equations, 139 (1997), pp. 76–103.

[11] A. Blake and A. Zisserman, *Visual Reconstruction*, MIT Press, Cambridge, MA, 1987.

[12] H. Brezis, *Operateurs Maximaux Monotones*, North–Holland, Amsterdam, Elsevier, New York, 1973.

[13] F. Catté, P.-L. Lions, J.-M. Morel, and T. Coll, *Image selective smoothing and edge detection by nonlinear diffusion*, SIAM J. Numer. Anal., 29 (1992), pp. 182–193.
[14] A. Chambolle, *Image segmentation by variational methods: Mumford–Shah functional and the discrete approximations*, SIAM J. Appl. Math., 55 (1995), pp. 827–863.
[15] M. G. Crandall, H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
[16] E. De Giorgi, *New problems on minimizing movements*, in Boundary Value Problems for Partial Differential Equations and Applications, RMA Res. Notes Appl. Math. 29, C. Baiocchi and J. L. Lions, eds., Masson, Paris, 1993, pp. 81–98.
[17] E. De Giorgi, *Su alcuni problemi instabili legati alla teoria della visione*, in Atti del Convegno in onore di Carlo Ciliberto (Napoli, 1995), T. Bruno, P. Buonocore, L. Carbone, and V. Esposito, eds., La Città del Sole, Napoli, Italy, 1997, pp. 91–98.
[18] E. De Giorgi, *Congetture riguardanti alcuni problemi di evoluzione*, Duke Math. J., 81 (1995), pp. 255–268.
[19] S. Demoulini, *Young measure solutions for a nonlinear parabolic equation of forward-backward type*, SIAM J. Math. Anal., 27 (1996), pp. 376–403.
[20] S. Esedoglu, *An analysis of the Perona-Malik scheme*, Comm. Pure Appl. Math., 54 (2001), pp. 1442–1487.
[21] F. Fierro, R. Goglione, and M. Paolini, *Numerical simulations of mean curvature flow in presence of a nonconvex anisotropy*, Math. Models Methods Appl. Sci., 8 (1998), pp. 573–601.
[22] M. Gobbino, *Gradient flow for the one-dimensional Mumford-Shah functional*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 27 (1998), pp. 145–193.
[23] M. Gobbino, *Entire Solutions of the One-Dimensional Perona-Malik Equation*, preprint, 2003.
[24] M. E. Gurtin, H.-M. Soner, and P. E. Souganidis, *Anisotropic motion of an interface relaxed by the formation of infinitesimal wrinkles*, J. Differential Equations, 119 (1995), pp. 54–108.
[25] E. Hewitt and K. Stromberg, *Real and Abstract Analysis. A Modern Treatment of the Theory of Functions of a Real Variable*, Springer-Verlag, New York, Berlin, 1969.
[26] K. Höllig, *Existence of infinitely many solutions for a forward-backward heat equation*, Trans. Amer. Math. Soc., 278 (1983), pp. 299–316.
[27] K. Höllig and J. A. Nohel, *A diffusion equation with a nonmonotone constitutive function*, in Systems of Nonlinear Differential Equations, J. Ball, ed., D. Reidel, Boston, MA, 1983, pp. 409–422.
[28] B. Kawohl and N. Kutev, *Maximum and comparison principle for one-dimensional anisotropic diffusion*, Math. Ann., 311 (1998), pp. 107–123.
[29] S. Kichenassamy, *The Perona–Malik paradox*, SIAM J. Appl. Math., 57 (1997), pp. 1328–1342.
[30] M. M. Lavrent'ev, V G. Romanov, and S. P. Shishat·skiĭ, *Ill-Posed Problems of Mathematical Physics and Analysis*, Transl. Math. Monogr. 64, AMS, Providence, RI, 1980.
[31] G. M. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, Singapore, 1996.
[32] N. E. Muler, *An L^∞ bound for a time-discrete regularization of a forward-backward parabolic equation*, Indiana Univ. Math. J., 45 (1996), pp. 683–693.
[33] M. Nitzberg, D. Mumford, and T. Shiota, *Filtering, Segmentation and Depth*, Lecture Notes in Comput. Sci. 662, Springer-Verlag, Berlin, 1993.
[34] P. Perona and J. Malik, *Scale space and edge detection using anisotropic diffusion*, IEEE Trans. Pattern Anal. Mach. Intell., 12 (1990), pp. 629–639.
[35] P. I. Plotnikov, *Passage to the limit with respect to viscosity in an equation with variable parabolicity direction*, Differential Equations, 30 (1994), pp. 614–622.
[36] M. Slemrod, *Dynamics of measured valued solutions to a backward-forward heat equation*, J. Dynam. Differential Equations, 3 (1991), pp. 614–622.

SIAM J. MATH. ANAL.
Vol. 37, No. 5, pp. 1688–1708

DYNAMICS OF A STAGE-STRUCTURED POPULATION MODEL ON AN ISOLATED FINITE LATTICE*

Y. KYRYCHKO†, S. A. GOURLEY†, AND M. V. BARTUCCELLI†

Abstract. In this paper we derive a stage-structured model for a single species on a finite one-dimensional lattice. There is no migration into or from the lattice. The resulting system of equations, to be solved for the total adult population on each patch, is a system of delay equations involving the maturation delay for the species, and the delay term is nonlocal involving the population on all patches. We prove that the model has a positivity preserving property. The main theorems of the paper are comparison principles for the cases when the birth function is increasing and when the birth function is a nonmonotone function. Using these theorems we prove results on the global stability of a positive equilibrium.

Key words. nonlocal, stage structure, lattice, time delay, comparison

AMS subject classifications. 34K25, 34K12, 92D25

DOI. 10.1137/S003614100444441X

1. Background.

There has recently been some interest in the study of stage-structured population models on lattices. The following model, which is of particular relevance to the present paper, was derived and studied by Weng, Huang, and Wu [8]:

$$\begin{aligned}\frac{dw_j(t)}{dt} &= \frac{\mu}{2\pi}\sum_{k=-\infty}^{\infty}\mathcal{B}_\alpha(j-k)b(w_k(t-r)) + D_m\left[w_{j+1}(t)+w_{j-1}(t)-2w_j(t)\right]\\ &\quad - d_m w_j(t), \qquad t>0, \qquad j\in\mathbf{Z},\end{aligned}\tag{1.1}$$

where $j \in \mathbf{Z} := \{0, \pm 1, \pm 2, \dots\}$ are the integer nodes of an infinite one-dimensional lattice. The r-dependent parameters μ and α are given by

$$\mu = \exp\left(-\int_0^r d(a)\,da\right), \qquad \alpha = \int_0^r D(a)\,da\tag{1.2}$$

with $d(a)$, $D(a)$, and r defined below, and the function $\mathcal{B}_\alpha(l)$ in (1.1), which we shall sometimes refer to as the *kernel*, is given by

$$\mathcal{B}_\alpha(l) = 2e^{-2\alpha}\int_0^\pi \cos(l\omega)e^{2\alpha\cos\omega}\,d\omega.\tag{1.3}$$

In their paper, Weng, Huang, and Wu [8] actually used the symbol β for their kernel. However, it is important that their kernel should not be confused with the corresponding kernel for a finite lattice (which we are calling β in this paper); thus we shall refer to the infinite lattice kernel of Weng, Huang, and Wu as $\mathcal{B}$.

In (1.1) the parameter r measures the time from birth until reaching maturity and $w_j(t)$ denotes the total number of adults (i.e., the total number of age at least r) in the jth patch. The function $b(\cdot)$, which always satisfies $b(0) = 0$, is the birth function

*Received by the editors June 9, 2004; accepted for publication (in revised form) June 1, 2005; published electronically January 27, 2006.
http://www.siam.org/journals/sima/37-5/44441.html
†Department of Mathematics and Statistics, University of Surrey, Guildford, Surrey GU2 7XH, U.K. (s.gourley@surrey.ac.uk).

and the constants D_m and d_m are, respectively, the diffusion coefficient and death rate for the mature population.

In (1.2), the functions $D(a)$ and $d(a)$ are the diffusion coefficient and death rate for the immature population. For the immature population these rates can depend on age a, but for the mature population the diffusion coefficient and death rate must be independent of age and they are taken as D_m and d_m.

Weng, Huang, and Wu [8] derived their model from the following von Foerster type of equation:

$$\frac{\partial u_j}{\partial t} + \frac{\partial u_j}{\partial a} = D(a)\left[u_{j+1}(t,a) + u_{j-1}(t,a) - 2u_j(t,a)\right] - d(a)u_j(t,a) \tag{1.4}$$

with $D(a) = D_m$ and $d(a) = d_m$ for $a \geq r$. Equation (1.4) incorporates a discrete representation of diffusion. Von Foerster equations for the case of continuous space have been considered also, in which case the Laplacian operator can be used to model Fickian diffusion (see [4]). In (1.4), $u_j(t,a)$ is the density of age a at time t in the jth patch. Furthermore

$$w_j(t) = \int_r^\infty u_j(t,a)\,da.$$

In [8] the interest is mainly in the existence of travelling front solutions connecting two distinct equilibria. The highly nontrivial matter of the stability of these fronts is also investigated. Gourley and Wu [3] continued the study in [8] by providing conditions under which the population will go extinct, and conditions for the existence of periodic travelling waves.

The aim of the present paper is to derive and study an equation analogous to (1.1) for the case when the lattice is finite, with the nodes being given by $j = 1, 2, \ldots, N$. As we shall see, the model changes in two main respects. The first is that the discrete representation of diffusion will only be $w_{j+1}(t)+w_{j-1}(t)-2w_j(t)$ at "interior" points of the lattice (i.e., the nodes $j = 2, 3, \ldots, N-1$) with a different expression for the nodes $j = 1$ and $j = N$. We shall use the expression appropriate for an isolated lattice which individuals cannot escape from or enter into; this is the analogy of the homogeneous Neumann problem (i.e., no flux at boundaries) in the case of continuous space. The second and more complicated difference between the model of the present paper and (1.1) is that the term with the time delay assumes a rather different appearance. The function $\mathcal{B}_\alpha(l)$ given by (1.3) is completely inappropriate for the case of a finite lattice. Additionally, as we shall see, in the case of a finite lattice the time delay term no longer assumes a "convolution" structure (i.e., depending on the lattice index through the variables $j-k$ and k with summation over k). This convolution formulation cannot allow for interactions with the endpoints of a finite lattice and therefore is strictly for infinite lattices only.

The derivation in [8] relies heavily on the fact that their lattice was infinite (their derivation utilizes a discrete Fourier transform technique). For a finite lattice a different model derivation is required and this will be the subject of the next section.

A model on a finite lattice similar to the one we propose in this paper was considered in Smith and Thieme [7]. Their model has n patches each of which offers a different quality of life (the per capita reproduction rates, mortality rates, and maturation delays can vary from patch to patch). Additionally their model permits individuals to migrate from any one patch directly to any other. Their model assumes the form

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial a}\right)u_j(t,a)=\sum_{k=1}^{n}\gamma_{jk}u_k(t,a)-\left(\sum_{k=1}^{n}\gamma_{kj}+\mu_j\right)u_j(t,a)$$

with birth law

$$u_j(t,0)=g_j\left(\int_{\tau_j}^{\infty}u_j(t,a)\,da\right),$$

where each g_j is a bounded function. In some respects this model is more general than the one we propose in this paper, particularly with regard to the migration terms (γ_{jk} is the per capita migration rate from node k to node j). Their key assumption on the migration terms is that the matrix (γ_{jk}) be irreducible. Our model in this paper considers migration only on a nearest neighbor basis, but we study our model in somewhat more detail.

Our approach is to start with an age-structured model (the "original problem") given by system (2.1) below, and derive from it a delay differential equation system (the "reduced problem") for $w_j(t)$, the total mature population at the jth patch. As we shall discover, the reduced system (system (2.10) below) is valid only after a transient period of length r. For $0<t<r$ the variable $w_j(t)$ is governed by some different nonautonomous equations that involve the initial data $\mathbf{u}_0(a)$ for the original problem (2.1). We shall not be concerned in this paper with these other nonautonomous equations but will effectively neglect the transient phase and study the reduced problem independently (i.e., we study system (2.10) below for $t>0$, subject to (2.11)). However, there are delicate issues regarding initial data and positivity which will be discussed later in this paper and which are treated in more detail in Bocharov and Hadeler [2] for problems without diffusion. For example, only certain initial data for the reduced problem with delay are related to the original problem. Also, while positive solutions of the original problem lead to positive solutions of the reduced problem, the cone of positive solutions of the reduced problem is larger in general, since we study the reduced problem for arbitrary nonnegative initial data, and not just those initial data that are related to the original problem. These issues will be discussed further as they arise.

2. Finite lattice: Model derivation. Let $u_j(t,a)$ denote the density of the population of the species at the jth patch at time $t\geq 0$ and age $a\geq 0$. Let $D(a)$ and $d(a)$ denote the diffusion and death rates of the population at age a. Assume that the patches are located at the integer nodes $j=1,2,\ldots,N$ of a one-dimensional lattice and that spatial diffusion occurs only at the nearest neighborhood and is proportional to the difference of the densities of the population at adjacent patches. These assumptions lead to the model

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial a}\right)\mathbf{u}(t,a)=D(a)A\mathbf{u}(t,a)-d(a)\mathbf{u}(t,a) \tag{2.1}$$

for $t>0$, where

$$\mathbf{u}(t,a)=(u_1(t,a),\ldots,u_N(t,a))^T$$

and

$$A = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix}, \tag{2.2}$$

subject to

$$u_j(t,0) = b(w_j(t)), \tag{2.3}$$

where $w_j(t)$ is the total mature population at the jth patch, given by

$$w_j(t) = \int_r^\infty u_j(t,a)\,da \tag{2.4}$$

and $b(\cdot)$ is the birth function, which satisfies $b(0) = 0$. Furthermore, it is natural to assume that

$$u_j(t,\infty) = 0, \qquad t \geq 0, \qquad j = 1, \dots, N.$$

The initial data for (2.1) has the form

$$\mathbf{u}(0,a) = \mathbf{u}_0(a), \qquad 0 \leq a < \infty \tag{2.5}$$

with $\mathbf{u}_0(a)$ prescribed.

Note that, at the node $j = 1$, the diffusion term is $D(a)(u_2 - u_1)$ with a similar expression for the other "end" node $j = N$. In this way the model (2.1) has been set up so as to be the discrete analogue of what is commonly called the homogeneous Neumann problem in the continuous case, in which no-flux boundary conditions are applied. For the heat equation $u_t = \Delta u$ on a finite domain Ω with homogeneous Neumann boundary conditions $\partial u/\partial n = 0$ on $\partial\Omega$, it is well known that $\int_\Omega u(t,x)\,dx$ is constant. An analogous result holds for (2.1) in the case when there are no births or deaths (i.e., $b(\cdot) = d(\cdot) = 0$). Indeed, in this case,

$$\begin{aligned} \underbrace{\frac{d}{dt}\sum_{j=1}^N \int_0^\infty u_j(t,a)\,da}_{\text{total population}} &= \sum_{j=1}^N \int_0^\infty \frac{\partial u_j(t,a)}{\partial t}\,da = -\sum_{j=1}^N \int_0^\infty \frac{\partial u_j(t,a)}{\partial a}\,da \\ &+ \sum_{j=2}^{N-1} \int_0^\infty D(a)(u_{j-1}(t,a) - 2u_j(t,a) + u_{j+1}(t,a))\,da \\ &+ \int_0^\infty D(a)(-u_1(t,a) + u_2(t,a))\,da + \int_0^\infty D(a)(u_{N-1}(t,a) - u_N(t,a))\,da \\ &= -\sum_{j=1}^N (\underbrace{u_j(t,\infty)}_{=0} - \underbrace{b(w_j(t))}_{=0 \text{ if no births}}) = 0. \end{aligned}$$

Our intention is to derive from (2.1), (2.3) a system of equations satisfied by the total matured population $w_j(t)$, $j = 1, 2, \dots, N$. Before doing so, let us introduce the function $\beta(t,k,j)$ defined by

$$\beta(t,k,j) = \frac{1}{N} + \frac{2}{N}\sum_{l=1}^{N} e^{-4\sin^2\left(\frac{l\pi}{2N}\right)t}\cos\left[(2j-1)\frac{l\pi}{2N}\right]\cos\left[(2k-1)\frac{l\pi}{2N}\right]. \tag{2.6}$$

We will prove the following result which is useful for later calculations.

PROPOSITION 2.1. *The function $\beta(t,k,j)$ defined by* (2.6) *has the following properties:*

(i) *it satisfies*

$$\frac{d}{dt}\begin{pmatrix}\beta(t,k,1)\\ \vdots\\ \beta(t,k,N)\end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 & 0\\ 1 & -2 & 1 & \cdots & 0 & 0 & 0\\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 1 & -2 & 1\\ 0 & 0 & 0 & \cdots & 0 & 1 & -1\end{pmatrix}\begin{pmatrix}\beta(t,k,1)\\ \vdots\\ \beta(t,k,N)\end{pmatrix}, \tag{2.7}$$

(ii) *$\beta(0,k,j)$ is the Kronecker delta:*

$$\beta(0,k,j) = \begin{cases} 1 & \text{if } j = k,\\ 0 & \text{if } j \neq k;\end{cases} \tag{2.8}$$

(iii) *$\sum_{k=1}^{N}\beta(t,k,j) = 1$ for each $j = 1,2,\ldots,N$ and all $t \geq 0$;*
(iv) *$\beta(t,k,j) > 0$ for all $t > 0$ and all $1 \leq k,j \leq N$.*

Proof. Property (i) is straightforward. Property (ii) is fairly easily seen. Indeed,

$$\begin{aligned}\beta(0,k,k) &= \frac{1}{N} + \frac{2}{N}\sum_{l=1}^{N}\cos^2(2k-1)\frac{l\pi}{2N} = \frac{1}{N} + \frac{1}{N}\sum_{l=1}^{N}\left(1+\cos(2k-1)\frac{l\pi}{N}\right)\\ &= \frac{1}{N} + 1 + \frac{1}{N}\mathrm{Re}\left(\sum_{l=1}^{N} e^{i(2k-1)\frac{l\pi}{N}}\right)\\ &= \frac{1}{N} + 1 + \frac{2}{N}\mathrm{Re}\left(\frac{e^{i(2k-1)\frac{\pi}{N}}}{1-e^{i(2k-1)\frac{\pi}{N}}}\right)\\ &= 1\end{aligned}$$

after some algebra. Similarly, $\beta(0,k,j) = 0$ for $j \neq k$.

To show statement (iii) it is clearly sufficient to prove that

$$\sum_{k=1}^{N}\cos(2k-1)\frac{l\pi}{2N} = 0$$

and this is easily shown.

Finally we prove (iv), that $\beta(t,k,j) > 0$ for all $t > 0$. We have already noted that the function $\beta(t,k,j)$ defined by (2.6) satisfies the system of differential equations (2.7) with the initial condition (2.8). Here, j is thought of as the spatial coordinate and $k \in \{1,2,\ldots,N\}$ as fixed. Certain theorems in the theory of matrices (Berman and Plemmons [1]) are useful here. The matrix $\beta(t) = [\beta(t,k,j)]_{N\times N}$ is the solution of the linear system $\dot{\beta} = A\beta$, where A is the matrix defined by (2.2), subject to $\beta(0) = I$. Thus

$$\beta(t) = e^{At} = e^{-3t}e^{(3I+A)t} = e^{-3t}\sum_{i=0}^{\infty}\frac{t^i(3I+A)^i}{i!}.$$

It is easily checked that $3I + A$ is a positive and irreducible matrix. Therefore, it follows [1] that $(3I + A)^i$ is a strictly positive matrix for all $i \geq N + 1$. Therefore, for any $t > 0$ the infinite sum above furnishes a matrix all of whose elements are strictly positive. Therefore, $\beta(t, k, j) > 0$ for all $t > 0$. The proof of Proposition 2.1 is complete. $\square$

We will now prove the following theorem.

THEOREM 2.2. *Assume that the diffusion and death rates of the mature population are age-independent, i.e.,*

$$D(a) = D_m, \qquad d(a) = d_m \quad \textit{for } a \in [r, \infty),$$

where $D_m > 0$ and $d_m > 0$ are constants. Let the function $\beta(t, k, j)$ be defined by (2.6) *and let*

$$\mu = e^{-\int_0^r d(z)\,dz}, \qquad \alpha = \int_0^r D(z)\,dz. \tag{2.9}$$

Then for $t \geq r$ the total matured population $w_j(t)$ defined by (2.4) *satisfies*

$$\frac{d}{dt}\begin{pmatrix} w_1(t) \\ w_2(t) \\ \vdots \\ w_{N-1}(t) \\ w_N(t) \end{pmatrix} = \mu \sum_{k=1}^{N} b(w_k(t-r)) \begin{pmatrix} \beta(\alpha, k, 1) \\ \beta(\alpha, k, 2) \\ \vdots \\ \beta(\alpha, k, N-1) \\ \beta(\alpha, k, N) \end{pmatrix}$$

$$+D_m \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} w_1(t) \\ w_2(t) \\ \vdots \\ w_{N-1}(t) \\ w_N(t) \end{pmatrix} - d_m \begin{pmatrix} w_1(t) \\ w_2(t) \\ \vdots \\ w_{N-1}(t) \\ w_N(t) \end{pmatrix}. \tag{2.10}$$

Remark 1. For $0 < t < r$ the variable $w_j(t)$ does not obey system (2.10) but is governed instead by some different (nonautonomous) equations that still contain information about the initial data of the system (2.1). This issue is discussed in detail in Bocharov and Hadeler [2] for systems without diffusion. However, in this paper we shall concentrate on the problem consisting of system (2.10) for $t > 0$, subject to the initial conditions

$$w_j(s) = w_j^0(s) \geq 0, \qquad j = 1, 2, \ldots, N, \qquad s \in [-r, 0] \tag{2.11}$$

with $w_j^0(s)$ prescribed.

Remark 2. Related to the above point, is the issue of initial data. If we solve (2.10) for $t > 0$ an important question arises: does an arbitrary nonnegative initial function $w_j^0(s)$, $s \in [-r, 0]$, necessarily result, in any sense, from an initial datum $\mathbf{u}_0(a)$ for system (2.1)? The answer is, not necessarily. Given an initial datum $\mathbf{u}_0(a)$ for system (2.1), one should first evolve the variable $w_j(t)$ until time r according to the nonautonomous equations which govern $w_j(t)$ for $0 < t < r$, and then evolve $w_j(t)$ for $t > r$ according to (2.10) with initial time r, using as initial data the values of $w_j(t)$, $t \in [0, r]$, found by solving the nonautonomous equations. Those functions $w_j(t)$, $t \in [0, r]$, which arise in this way are, once translated r units back in time, those functions which are admissible in (2.11) as initial data for system (2.10) with initial time 0. For a class of systems without diffusion, Bocharov and Hadeler [2] characterize completely those initial data for their delay equation starting at time r, which result from a positive initial datum of their analogy to our system (2.1). In this paper we shall not develop this issue further but will consider (2.10) for $t > 0$ subject only to (2.11) (i.e., as though the system had not been derived from a structured population model).

Proof of Theorem 2.2. Letting

$$\mathbf{w}(t) = (w_1(t), \ldots, w_N(t))^T,$$

we have

$$\frac{d\mathbf{w}(t)}{dt} = \int_r^\infty \frac{\partial}{\partial t}\mathbf{u}(t,a)\,da = \int_r^\infty \left[-\frac{\partial}{\partial a}\mathbf{u}(t,a) + D(a)A\mathbf{u}(t,a) - d(a)\mathbf{u}(t,a)\right] da. \tag{2.12}$$

We obtain from (2.1) and (2.12) that

$$\frac{d\mathbf{w}(t)}{dt} = \mathbf{u}(t,r) + D_m A\mathbf{w}(t) - d_m\mathbf{w}(t) \qquad \text{for } t > 0. \tag{2.13}$$

In order to have a complete system for $w_j(t)$ we need to calculate $u_j(t,r)$, $j = 1, 2, \ldots, N$. For fixed $s \geq 0$ let $\mathbf{V}^s(t) = (V_1^s(t), \ldots, V_N^s(t))^T$, where

$$V_j^s(t) = u_j(t, t-s) \quad \text{for } s \leq t \leq s + r. \tag{2.14}$$

Since only the mature population can reproduce, we have

$$V_j^s(s) = u_j(s, 0) = b(w_j(s)), \tag{2.15}$$

where $b : \mathbb{R}_+ \to \mathbb{R}_+$ is the birth function. From (2.1), we have

$$\frac{d}{dt}\mathbf{V}^s(t) = \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial a}\right)\mathbf{u}(t,a)\bigg|_{a=t-s} = D(t-s)A\mathbf{V}^s(t) - d(t-s)\mathbf{V}^s(t). \tag{2.16}$$

We want to solve (2.16) subject to (2.15). First, let

$$V_j^s(t) = \tilde{V}_j^s(t) e^{-\int_0^{t-s} d(z)\,dz}$$

and $\tilde{\mathbf{V}}^s(t) = (\tilde{V}_1^s(t), \ldots, \tilde{V}_N^s(t))^T$; then system (2.16) becomes

$$\frac{d}{dt}\tilde{\mathbf{V}}^s(t) = D(t-s)A\tilde{\mathbf{V}}^s(t).$$

Making a further transformation of time t as

$$\tilde{t} = \int_0^{t-s} D(z)\,dz$$

so that $d\tilde{t}/dt = D(t-s)$, we obtain

$$\frac{d}{d\tilde{t}}\tilde{\mathbf{V}}^s(\tilde{t}) = A\tilde{\mathbf{V}}^s(\tilde{t}) \tag{2.17}$$

which has to be solved subject to the initial condition (2.15) in the form

$$\tilde{V}_j^s(0) = b(w_j(s)), \quad j = 1, 2, \ldots, N. \tag{2.18}$$

We look for a solution of (2.17) in the form

$$\tilde{V}_j^s(\tilde{t}) = \sum_{k=1}^{N} c_k \beta(\tilde{t}, k, j),$$

where the function $\beta(t, k, j)$ is defined in (2.6) and c_k are unknown constants to be found using the initial condition (2.18). That is, with $\tilde{t} = 0$ one gets

$$\tilde{V}_j^s(0) = \sum_{k=1}^{N} c_k \beta(0, k, j) = c_j,$$

by (ii) of Proposition 2.1. Consequently,

$$c_j = b(w_j(s)),$$

and the solution of the system (2.17) is given as follows:

$$\tilde{V}_j^s(\tilde{t}) = \sum_{k=1}^{N} b(w_k(s)) \beta(\tilde{t}, k, j),$$

or, in terms of t,

$$\tilde{V}_j^s(t) = \sum_{k=1}^{N} b(w_k(s)) \beta\left(\int_0^{t-s} D(z)dz, k, j\right).$$

Hence,

$$u_j(t, t-s) = V_j^s(t) = e^{-\int_0^{t-s} d(z)dz} \sum_{k=1}^{N} b(w_k(s)) \beta\left(\int_0^{t-s} D(z)dz, k, j\right).$$

Recalling that $s \geq 0$ we deduce that, for $t \geq r$,

$$u_j(t, r) = \mu \sum_{k=1}^{N} b(w_k(t-r)) \beta(\alpha, k, j), \tag{2.19}$$

where $b(w_k(t))$ is the birth function introduced above, and μ and α are given by (2.9). Thus, the system for w_j has the form (2.10) for $t \geq r$.

3. Positivity of solutions. In the previous sections we derived the reduced model for $w_j(t)$ and proved various properties of $\beta(t, k, j)$. Using the fact that $\beta(t, k, j) > 0$ for all $t > 0$ we shall now prove a positivity-preserving property for system (2.10).

The result which we shall prove in this section is for the system consisting of the differential equations (2.10) for $t > 0$, supplemented by the initial data (2.11). As noted previously, the reduced system (2.10) is actually only valid for $t \geq r$. For $0 < t < r$ another nonautonomous system applies and governs $w_j(t)$. Thus there is a subtle relationship between the original problem (2.1) and the reduced problem (2.10) as regards positivity. This issue is discussed in detail in [2] for systems without diffusion, and the issues concerning positivity are basically the same for systems with and without diffusion.

First, we consider the initial value problem

$$\frac{d\mathbf{v}(t)}{dt} = DA\mathbf{v}(t) - d\mathbf{v}(t) + \mathbf{h}(t), \tag{3.1}$$

where $\mathbf{v}(t) = (v_1(t), \ldots, v_N(t))^T$ and $\mathbf{h}(t) = (h_1(t), \ldots, h_N(t))^T$, subject to

$$v_j(0) = c_j,\ c_j \in \mathbb{R},\ j = 1, \ldots, N.$$

From the definition of $\beta(t, k, j)$ it is easy to check that the solution to this problem is

$$v_j(t) = e^{-dt}\sum_{k=1}^{N}\beta(Dt, k, j)c_k + \sum_{k=1}^{N}\int_0^t e^{-d(t-s)}\beta(D(t-s), k, j)h_k(s)\,ds,\ j = 1, \ldots, N. \tag{3.2}$$

Therefore if $c_j \geq 0$ and $h_j(t) \geq 0$ for all $t \geq 0$ and $j = 1, \ldots, N$, then, since $\beta(t, k, j) \geq 0$, we have $v_j(t) \geq 0$ for all $j = 1, \ldots, N$ and $t \geq 0$. We shall now prove that the solutions of (2.10) enjoy positivity-preserving properties analogous to results that can be proved using the strong maximum principle in the case of continuous space.

THEOREM 3.1. *Let $b(0) = 0$ and $b(w) > 0$ when $w > 0$, and let $\{w_j\}$ be the solution of system* (2.10), *$t > 0$, corresponding to the initial data $w_j(s) = w_j^0(s)$, $s \in [-r, 0]$. If $w_j^0(s) \geq 0$ for all $j = 1, \ldots, N$ and $s \in [-r, 0]$, then $w_j(t) \geq 0$ for all $j = 1, \ldots, N$ and $t \geq 0$.*

Also, if $w_j^0(s) \not\equiv 0$ on $(s, j) \in [-r, 0] \times \{1, 2, \ldots, N\}$, then $w_j(t) > 0$ for all $j = 1, \ldots, N$ and $t \geq r$.

Proof. We first prove the nonnegativity property $w_j(t) \geq 0$. This is achieved in steps; we first prove the result for $t \in [0, r]$. Applying (3.2) to system (2.10) with

$$h_j(t) = \mu\sum_{k=1}^{N}\beta(\alpha, k, j)b(w_k(t-r))$$

gives, for $t \in [0, r]$,

$$\begin{aligned} w_j(t) = {} & e^{-d_m t}\sum_{k=1}^{N}\beta(D_m t, k, j)w_k(0) \\ & + \mu\sum_{k=1}^{N}\int_0^t e^{-d_m(t-s)}\beta(D_m(t-s), k, j)\sum_{l=1}^{N}\beta(\alpha, l, k)b(w_l(s-r))\,ds. \end{aligned} \tag{3.3}$$

Now $t \in [0, r]$ so the right-hand side of the above expression refers only to the initial data for $w_j(t)$, which is nonnegative by hypothesis. Since β is nonnegative also, we conclude that $w_j(t) \geq 0$ for $t \in [0, r]$. On the interval $t \in [r, 2r]$ an expression analogous to (3.3) can easily be found, referring in its right-hand side to $w_j(\cdot)$ only at times between 0 and r, on which interval we have just shown nonnegativity. Therefore, $w_j(t) \geq 0$ for $t \in [r, 2r]$. This argument can be continued indefinitely, and so we have shown $w_j(t) \geq 0$ for all $t \geq 0$.

We now prove the second part of the theorem: strict positivity of $w_j(t)$ for all $t \geq r$, provided the initial data is not identically zero. From (3.3) we can infer that

$$w_j(r) > 0 \qquad \text{for all } j = 1, \ldots, N.$$

Indeed, if we had $w_j(r) = 0$, then, since $\beta(t, k, j) > 0$ for $t > 0$ and since $b(\cdot)$ is positive definite, it would follow that $w_j(s) \equiv 0$ on $(s, j) \in [-r, 0] \times \{1, 2, \ldots, N\}$, contrary to hypothesis. For $t \geq r$ an expression for $w_j(t)$ similar to (3.3) can be found, and from this expression we can infer that, for $t \geq r$,

$$w_j(t) \geq e^{-d_m(t-r)} \sum_{k=1}^{N} \beta(D_m(t-r), k, j) w_k(r) > 0$$

for all $j = 1, \ldots, N$. The proof of the theorem is complete. $\square$

4. Comparison principle: Monotone birth functions. In this section we shall prove a comparison theorem for system (2.10) for the case when the birth function is increasing, and we shall use it to prove that in this case if a positive uniform equilibrium solution exists, then it is globally stable. Theorem 4.1 below can also be established using results in Chapter 5 of Smith [6] (indeed, for increasing birth functions system (2.10) satisfies the quasimonotone condition on page 78 of [6], and (2.10) constitutes a cooperative and irreducible system of functional differential equations). However, we will include a self-contained proof of our Theorem 4.1 here because, as will become clear later, a detailed understanding of the monotone case throws much light on how one can approach the case of a nonmonotone birth function which we do later in Theorem 5.1.

THEOREM 4.1. *Let the birth function $b(w)$ be increasing and differentiable for all $w \geq 0$ and let $\beta(t, k, j)$ be given by (2.6). Let $\bar{\mathbf{w}}(t) = (\bar{w}_1(t), \ldots, \bar{w}_N(t))^T$ and $\hat{\mathbf{w}}(t) = (\hat{w}_1(t), \ldots, \hat{w}_N(t))^T$ be such that*

$$\frac{d\bar{\mathbf{w}}(t)}{dt} - D_m A \bar{\mathbf{w}}(t) + d_m \bar{\mathbf{w}}(t) - \mu \sum_{k=1}^{N} b(\bar{w}_k(t-r)) \begin{pmatrix} \beta(\alpha, k, 1) \\ \beta(\alpha, k, 2) \\ \vdots \\ \beta(\alpha, k, N-1) \\ \beta(\alpha, k, N) \end{pmatrix}$$

$$\geq \frac{d\hat{\mathbf{w}}(t)}{dt} - D_m A \hat{\mathbf{w}}(t) + d_m \hat{\mathbf{w}}(t) - \mu \sum_{k=1}^{N} b(\hat{w}_k(t-r)) \begin{pmatrix} \beta(\alpha, k, 1) \\ \beta(\alpha, k, 2) \\ \vdots \\ \beta(\alpha, k, N-1) \\ \beta(\alpha, k, N) \end{pmatrix} \tag{4.1}$$

for $t > 0$ and

$$\bar{w}_j(s) \geq \hat{w}_j(s), \qquad j = 1, \ldots, N, \qquad s \in [-r, 0].$$

Then $\bar{w}_j(t) \geq \hat{w}_j(t)$ for all $t > 0$ and $j = 1, \ldots, N$.

Proof. Let $\mathbf{W}(t)$ be the vector with components $W_j(t) := \bar{w}_j(t) - \hat{w}_j(t)$. Then (4.1) can be rewritten as

$$\frac{d\mathbf{W}(t)}{dt} - D_m A\mathbf{W}(t) + d_m\mathbf{W}(t) - \mu\sum_{k=1}^{N}[b(\bar{w}_k(t-r)) - b(\hat{w}_k(t-r))]\begin{pmatrix}\beta(\alpha,k,1)\\ \beta(\alpha,k,2)\\ \vdots\\ \beta(\alpha,k,N-1)\\ \beta(\alpha,k,N)\end{pmatrix} \geq 0.$$

Applying the mean value theorem to the last term in the left-hand side of the above inequality we obtain

$$\frac{d\mathbf{W}(t)}{dt} - D_m A\mathbf{W}(t) + d_m\mathbf{W}(t) - \mu\sum_{k=1}^{N} b'(\theta_k(t-r))W_k(t-r)\begin{pmatrix}\beta(\alpha,k,1)\\ \beta(\alpha,k,2)\\ \vdots\\ \beta(\alpha,k,N-1)\\ \beta(\alpha,k,N)\end{pmatrix} \geq 0, \tag{4.2}$$

where $\theta_k(t)$ is between $\bar{w}_k(t)$ and $\hat{w}_k(t)$, $k = 1, \ldots, N$. By hypothesis $\bar{w}_j(s) \geq \hat{w}_j(s)$ for $s \in [-r, 0]$, so $W_j(s) \geq 0$ for $s \in [-r, 0]$. To prove the theorem we need to show that $W_j(t) \geq 0$ for all $t > 0$, and as a first step we shall prove this fact for $t \in (0, r]$. For $t \in (0, r]$,

$$f_j(t) := \mu\sum_{k=1}^{N} b'(\theta_k(t-r))W_k(t-r)\beta(\alpha,k,j) \geq 0.$$

Inequality (4.2) becomes

$$\frac{d\mathbf{W}(t)}{dt} - D_m A\mathbf{W}(t) + d_m\mathbf{W}(t) \geq \mathbf{f}(t), \tag{4.3}$$

where $\mathbf{f}(t) = (f_1(t), \ldots, f_N(t))^T$. We claim that $W_j(t) \geq 0$ for all $j = 1, \ldots, N$ and $t \in [0, r]$. Suppose this is false, i.e., that $W_j(t)$ goes negative. Then $W_j(t)$ must attain a negative minimum on the set $(t, j) \in [0, r] \times \{1, 2, \ldots, N\}$. Let this happen at time t^* and at the node j^*. Since $W_j(0) \geq 0$ we must have $t^* > 0$, but it is possible that $t^* = r$. In any case,

$$\frac{dW_{j^*}(t^*)}{dt} \leq 0,$$

and, of course, $W_{j^*}(t^*) < 0$. Also, if j^* is an "interior" node, then

$$W_{j^*-1}(t^*) - 2W_{j^*}(t^*) + W_{j^*+1}(t^*) \geq 0$$

while if $j^* = 1$, then $-W_1(t^*) + W_2(t^*) \geq 0$ and similarly if $j^* = N$. Extracting the j^*th component of (4.3) and evaluating it at time t^* gives

$$\underbrace{\frac{dW_{j^*}(t^*)}{dt} - (j^*\text{th component of } D_m \text{ term})}_{\leq 0} + \underbrace{d_m W_{j^*}(t^*)}_{<0} \geq \underbrace{f_{j^*}(t^*)}_{\geq 0}$$

which is a contradiction. Thus $W_j(t) \geq 0$ for all $j = 1, \ldots, N$ and $t \in [0, r]$. Repeating this argument establishes that $W_j(t) \geq 0$ for $t \in [r, 2r]$ and the argument can be continued to include all positive times. The proof of the theorem is complete. ◻

4.1. Convergence to equilibrium. In this section we will prove that if the birth function $b(w)$ is increasing and is such that there exists a uniform equilibrium solution w^* (independent of both j and t) to system (2.10) and is biologically realistic, then solutions of (2.10) approach the equilibrium w^*. Note first that a uniform equilibrium state w^* must necessarily satisfy

$$\mu b(w^*) = d_m w^*. \tag{4.4}$$

We will prove the following theorem.

THEOREM 4.2. *In system* (2.10) *let the birth function* $b(w)$ *satisfy* $b(0) = 0$ *and be an increasing differentiable function for all* $w \geq 0$. *Assume there exists* $w^* > 0$ *such that* $\mu b(w) > d_m w$ *when* $0 < w < w^*$ *and* $\mu b(w) < d_m w$ *when* $w > w^*$. *Assume further that, in* (2.11), $w_j^0(s) \not\equiv 0$ *on* $(s, j) \in [-r, 0] \times \{1, 2, \ldots, N\}$. *Then the solution* $w_j(t)$ *of* (2.10) *for* $t > 0$, *subject to* (2.11), *satisfies* $w_j(t) \to w^*$ *as* $t \to \infty$, *for each* $j = 1, 2, \ldots, N$.

Proof. To prove this theorem we shall use Theorem 4.1. More specifically, we shall show using Theorem 4.1 that the solution $w_j(t)$ of (2.10) subject to (2.11) can be bounded above and below by solutions of (2.10) that are functions of t only.

Indeed, if we denote by $w(t)$ any solution of the scalar equation

$$\frac{dw(t)}{dt} = \mu b(w(t - r)) - d_m w(t), \tag{4.5}$$

then the function

$$(w_1(t), w_2(t), \ldots, w_N(t)) := (w(t), w(t), \ldots, w(t))$$

satisfies (2.10). Two applications of Theorem 4.1 are required. In the first, we choose $\hat{w}_j(t)$ to be the solution $w_j(t)$ of (2.10) subject to (2.11) and $\bar{w}_j(t) = \bar{w}(t)$ for each j, where $\bar{w}(t)$ satisfies

$$\begin{aligned} &\frac{d\bar{w}(t)}{dt} = \mu b(\bar{w}(t - r)) - d_m \bar{w}(t), \\ &\bar{w}(s) = \max\{w_j^0(s),\ j = 1, 2, \ldots, N\} \quad \text{for } s \in [-r, 0]. \end{aligned} \tag{4.6}$$

Then Theorem 4.1 yields

$$w_j(t) \leq \bar{w}(t), \qquad t > 0, \qquad j = 1, 2, \ldots, N.$$

For the second application of Theorem 4.1 the most obvious choices are to take $\bar{w}_j(t)$ as the solution $w_j(t)$ of (2.10) subject to (2.11) and, for each j, $\hat{w}_j(t) = \hat{w}(t)$ where $\hat{w}(t)$ satisfies

$$\begin{aligned} &\frac{d\hat{w}(t)}{dt} = \mu b(\hat{w}(t - r)) - d_m \hat{w}(t), \\ &\hat{w}(s) = \min\{w_j^0(s),\ j = 1, 2, \ldots, N\} \quad \text{for } s \in [-r, 0] \end{aligned} \tag{4.7}$$

so that

$$w_j(t) \geq \hat{w}(t), \qquad t > 0, \qquad j = 1, 2, \ldots, N,$$

but this presents a possible problem in that $\hat{w}(s)$ could be zero on all of $s \in [-r, 0]$ without violating the assumption $w_j^0(s) \not\equiv 0$ on $(s, j) \in [-r, 0] \times \{1, 2, \ldots, N\}$ (e.g., if $w_j^0(s)$ were zero for all s on one particular node) in which case $\hat{w}(t)$ would be zero for all $t > 0$, which is not helpful for us. The way round this difficulty is to remember that we showed earlier (Theorem 3.1) that $w_j(t) > 0$ for all $t \geq r$. Consider the initial value problem starting at time $t = 2r$ and consisting of equation (2.10) for $t > 2r$, with initial data taken to be the solution $w_j(t)$, $t \in [r, 2r]$, of the original problem. The solution of this new initial value problem for $t > 2r$ is clearly the same as the solution $w_j(t)$ of the original problem, but the new problem has strictly positive initial data. This means that, without loss of generality, the minimum in (4.7) can be assumed to be strictly positive for all $s \in [-r, 0]$.

To complete the proof of Theorem 4.2 it therefore suffices to prove that every solution of the scalar ODE (4.5) such that $w(s) > 0$ for all $s \in [-r, 0]$ will satisfy $w(t) \to w^*$, if the hypotheses on the parameters and the function $b(w)$ are satisfied. This follows immediately from Theorem 9.1 on page 159 of the book by Kuang [9]. Therefore, the proof of Theorem 4.2 is complete. $\square$

5. Nonmonotone birth functions. Of considerable interest to ecologists, is the case of a birth function $b(w)$ which is increasing up to a certain value of w and decreasing thereafter (for example, a function qualitatively resembling $b(w) = Pwe^{-Aw}$). Such birth functions are important in modelling certain insect populations in which the birth rate is observed to be roughly proportional to the number of adults if the number of adults is small, but effectively zero if the number of adults is large, since competition for resources then becomes so intense that the adults require all their resources for their own maintenance. The aims of this section are to establish a comparison principle that works for very general birth functions, and then to use the comparison principle to prove convergence theorems in the case when the birth function qualitatively resembles $b(w) = Pwe^{-Aw}$.

THEOREM 5.1. *Let the birth function $b(w)$ be a differentiable function for all $w \geq 0$ and satisfy $b(0) = 0$, $b(w) > 0$ when $w > 0$. Let $\hat{w}$ and $\bar{w}$ be a pair of sub- and supersolutions for* (2.10), (2.11)*, i.e., a pair of functions satisfying*

(i) *$\hat{w}_j(t) \leq \bar{w}_j(t)$ for all $t \in [-r, \infty)$, $j = 1, 2, \ldots, N$;*

(ii) *letting $\bar{\mathbf{w}}(t) = (\bar{w}_1(t), \ldots, \bar{w}_N(t))^T$ and $\hat{\mathbf{w}}(t) = (\hat{w}_1(t), \ldots, \hat{w}_N(t))^T$ for $t > 0$ and $j = 1, 2, \ldots, N$,*

$$\frac{d\hat{\mathbf{w}}(t)}{dt} \leq D_m A\hat{\mathbf{w}}(t) - d_m\hat{\mathbf{w}}(t) + \mu \sum_{k=1}^{N} b(\varphi_k(t-r)) \begin{pmatrix} \beta(\alpha, k, 1) \\ \beta(\alpha, k, 2) \\ \vdots \\ \beta(\alpha, k, N-1) \\ \beta(\alpha, k, N) \end{pmatrix}, \tag{5.1}$$

and

$$\frac{d\bar{\mathbf{w}}(t)}{dt} \geq D_m A\bar{\mathbf{w}}(t) - d_m\bar{\mathbf{w}}(t) + \mu \sum_{k=1}^{N} b(\varphi_k(t-r)) \begin{pmatrix} \beta(\alpha, k, 1) \\ \beta(\alpha, k, 2) \\ \vdots \\ \beta(\alpha, k, N-1) \\ \beta(\alpha, k, N) \end{pmatrix} \tag{5.2}$$

for all functions $\varphi_j(t)$ *such that* $\hat{w}_j(t) \le \varphi_j(t) \le \bar{w}_j(t)$, $t \in [-r, \infty)$, $j = 1, 2, \ldots, N$;

(iii) $\hat{w}_j(s) \le w_j^0(s) \le \bar{w}_j(s)$, $s \in [-r, 0]$, $j = 1, 2, \ldots, N$, *where* $w_j^0(s)$ *is the initial data for* (2.10).

Then the solution $w_j(t)$ *of* (2.10), (2.11) *satisfies*

$$\hat{w}_j(t) \le w_j(t) \le \bar{w}_j(t) \quad \textit{for all } t > 0,\ j = 1, 2, \ldots, N.$$

Proof. Using (2.10), inequality (5.1) can be rewritten as

$$\frac{d\hat{\mathbf{w}}(t)}{dt} - D_m A\hat{\mathbf{w}}(t) + d_m\hat{\mathbf{w}}(t) - \mu\sum_{k=1}^{N} b(\varphi_k(t-r))\begin{pmatrix}\beta(\alpha,k,1)\\ \beta(\alpha,k,2)\\ \vdots\\ \beta(\alpha,k,N-1)\\ \beta(\alpha,k,N)\end{pmatrix}$$

$$\le \frac{d\mathbf{w}(t)}{dt} - D_m A\mathbf{w}(t) + d_m\mathbf{w}(t) - \mu\sum_{k=1}^{N} b(w_k(t-r))\begin{pmatrix}\beta(\alpha,k,1)\\ \beta(\alpha,k,2)\\ \vdots\\ \beta(\alpha,k,N-1)\\ \beta(\alpha,k,N)\end{pmatrix}. \tag{5.3}$$

Define $W_j(t) = w_j(t) - \hat{w}_j(t)$ and $\mathbf{W}(t)$ to be the vector with components $W_j(t)$. Then (5.3) becomes

$$\begin{aligned}\frac{d\mathbf{W}(t)}{dt} &- D_m A\mathbf{W}(t) + d_m\mathbf{W}(t) - \mu\sum_{k=1}^{N}[b(w_k(t-r))\\ &- b(\varphi_k(t-r))]\begin{pmatrix}\beta(\alpha,k,1)\\ \beta(\alpha,k,2)\\ \vdots\\ \beta(\alpha,k,N-1)\\ \beta(\alpha,k,N)\end{pmatrix} \ge 0\end{aligned} \tag{5.4}$$

and by hypothesis this holds for all $\varphi_j(t)$ such that $\hat{w}_j(t) \le \varphi_j(t) \le \bar{w}_j(t)$, $t \in [-r, \infty)$, $j = 1, 2, \ldots, N$.

We need to prove that $W_j(t) \ge 0$ for all $t > 0$ and all j, and we shall first prove this conclusion for $t \in (0, r]$. In inequality (5.4), for each j choose $\varphi_j(t)$ to be any function between $\hat{w}_j(t)$ and $\bar{w}_j(t)$, $t \in [-r, \infty)$, which is such that $\varphi_j(s) = w_j^0(s)$ when $s \in [-r, 0]$. From this choice for $\varphi_j(t)$ we infer that, for $t \in (0, r]$ only, the last term in the left-hand side of inequality (5.4) is zero, so that the inequality holds for $t \in (0, r]$ with just the first three terms in the left-hand side. The proof that $W_j(t) \ge 0$ for $t \in (0, r]$ then proceeds the same way as in the proof of Theorem 4.1 because our inequality is the same as (4.3) in the case when the functions $f_i(t)$ of the latter are zero.

Proving that $w_j(t) \le \bar{w}_j(t)$ for $t \in (0, r]$ and all j is similar. Thus

$$\hat{w}_j(t) \le w_j(t) \le \bar{w}_j(t) \quad \text{for } t \in (0, r],\ j = 1, 2, \ldots, N. \tag{5.5}$$

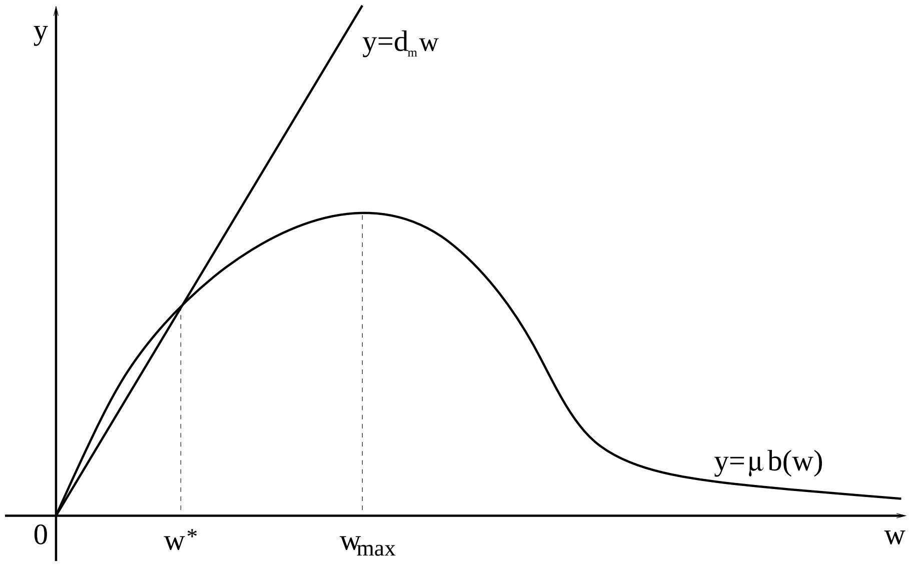


FIG. 1. *Graphical depiction of the situation in which Theorem* 5.2 *holds. In particular, the equilibrium* w^* *has to satisfy* $0 < w^* < w_{\max}$.

Proving that $W_j(t) = w_j(t) - \hat{w}_j(t) \geq 0$ for $t \in (r, 2r]$ is similar. Inequality (5.4) still holds for all $t > 0$ and in particular for $t \in (r, 2r]$. This time, we choose $\varphi_j(t)$ to be any function between $\hat{w}_j(t)$ and $\bar{w}_j(t)$, $t \in [-r, \infty)$, which is such that $\varphi_j(t) = w_j(t)$ when $t \in [0, r]$. This choice furnishes for us inequality (5.4), on $t \in (r, 2r]$ only, but without the term involving summation. We thus conclude that $W_j(t) \geq 0$ for $t \in (r, 2r]$ and it is clear how to continue the proof. $\square$

Remark. A comparison theorem similar to Theorem 5.1 was proved for the case of continuous space by Redlinger [5].

5.1. Convergence to equilibrium when * $_{\max}$. We will use Theorem 5.1 to establish, essentially, that if $b(w)$ qualitatively resembles Pwe^{-Aw} and if a nonzero equilibrium of (2.10) exists, is unique, and is in the interval of w for which $b(w)$ is increasing, then the equilibrium is globally stable as a solution of (2.10).

THEOREM 5.2. *In system* (2.10) *let the birth function* $b(w)$ *satisfy* $b(0) = 0$ *and* $b(w) > 0$ *when* $w > 0$. *Also, let* $b(w)$ *be increasing for* $0 < w < w_{\max}$, *with* $b'(w_{\max}) = 0$, *and decreasing for* $w > w_{\max}$. *Assume further that there exists* $w^* > 0$ *such that* $\mu b(w) > d_m w$ *when* $0 < w < w^*$ *and* $\mu b(w) < d_m w$ *when* $w > w^*$, *and assume that* $w^* < w_{\max}$.

Then, if $w_j^0(s) \not\equiv 0$ *on* $(s, j) \in [-r, 0] \times \{1, 2, \ldots, N\}$, *the solution* $w_j(t)$ *of* (2.10) *for* $t > 0$, *subject to* (2.11), *satisfies* $w_j(t) \to w^*$ *as* $t \to \infty$, *for each* $j = 1, 2, \ldots, N$.

Remark. The situation we have in mind is shown in Figure 1.

Proof of Theorem 5.2. Let $\hat{w}_j(t) = 0$ and $\bar{w}_j(t) = v(t)$ for each $j = 1, 2, \ldots, N$, where $v(t)$ is the solution of

$$\frac{dv(t)}{dt} = \mu b(w_{\max}) - d_m v(t),$$

$$v(s) = \max\{w_j^0(s),\ j = 1, 2, \ldots, N\}, \qquad s \in [-r, 0].$$

It is easily seen that $\hat{w}_j(t)$ and $\bar{w}_j(t)$ are a pair of sub- and supersolutions. Thus, by Theorem 5.1,

$$0 \le w_j(t) \le v(t) \qquad \text{for all } t > 0,\ \ j = 1, \dots, N.$$

Thus

$$\limsup_{t\to\infty} \max_{j\in\{1,2,\dots,N\}} w_j(t) \le \lim_{t\to\infty} v(t) = \frac{\mu b(w_{\max})}{d_m}.$$

Under the hypotheses it can be shown that

$$w^* < \frac{\mu b(w_{\max})}{d_m} < w_{\max}.$$

Choose $\epsilon > 0$ sufficiently small such that

$$\frac{\mu b(w_{\max})}{d_m} + \epsilon < w_{\max}.$$

There exists a time $T > 0$ such that, for all $t > T$ and all j,

$$w_j(t) \le \frac{\mu b(w_{\max})}{d_m} + \epsilon < w_{\max}.$$

Then as soon as t exceeds $T + r$ there is effectively no record, as far as system (2.10) is concerned, of the solution $w_j(t)$ ever having taken values outside the interval $[0, w_{\max}]$. From this point on the analysis proceeds as if the birth function were increasing for all w, and therefore it follows from Theorem 4.2 that the solution converges to w^*. The proof is complete. ◻

5.2. Convergence to equilibrium when * $_{\max}$. This section will show that if $w^* > w_{\max}$, then solutions of (2.10), (2.11) will still converge to w^* if additional conditions hold. These additional conditions will hold if w^* is not too much larger than $w_{\max}$.

THEOREM 5.3. *In system* (2.10) *let the birth function $b(w)$ satisfy $b(0) = 0$ and $b(w) > 0$ when $w > 0$. Also, let $b(w)$ be increasing for $0 < w < w_{\max}$, with $b'(w_{\max}) = 0$, and decreasing for $w > w_{\max}$. Assume further that there exists $w^* > 0$ such that $\mu b(w) > d_m w$ when $0 < w < w^*$ and $\mu b(w) < d_m w$ when $w > w^*$. Assume that $w^* > w_{\max}$ and that*

$$\frac{1}{d_m}\mu b\left(\frac{\mu b(w_{\max})}{d_m}\right) > w_{\max}. \tag{5.6}$$

Furthermore, we assume that

$$(d_m + \bar{f})r < 1, \tag{5.7}$$

where

$$\bar{f} = \mu \max\{|b'(w)|,\ \ w \in [\underline{w}, \overline{w}]\}$$

with $\overline{w} = (\mu/d_m)b(w_{\max})$ and $\underline{w} = b^{-1}(b(w^*))|_{[0,w_{\max}]}$. Assume further that

$$\frac{\mu}{d_m} b'(w^*) > -1. \tag{5.8}$$

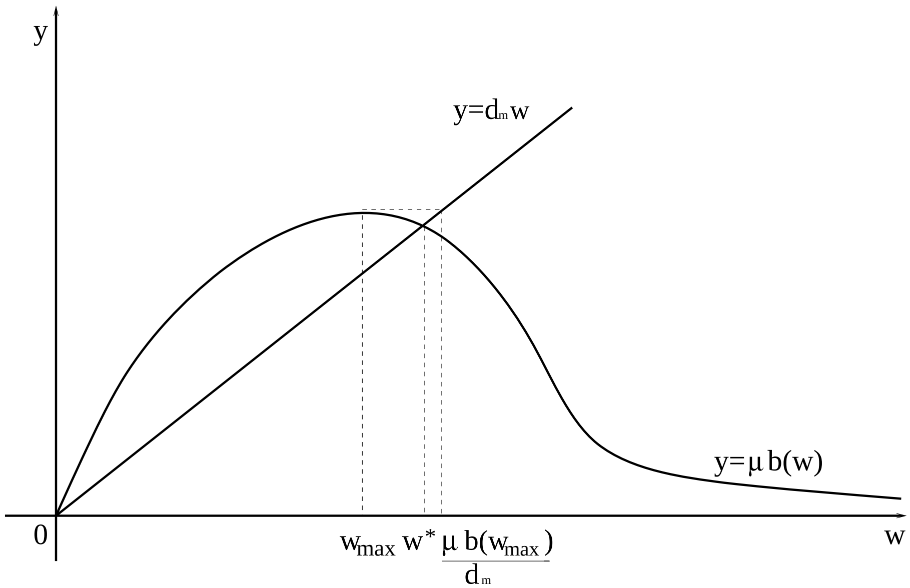


FIG. 2. *Graphical depiction of the situation in which Theorem* 5.3 *holds. In particular, the equilibrium w^* has to satisfy $w^* > w_{\max}$.*

Then, if $w_j^0(s) \not\equiv 0$ on $(s, j) \in [-r, 0] \times \{1, 2, \ldots, N\}$, the solution $w_j(t)$ of (2.10) *for $t > 0$, subject to* (2.11) *satisfies $w_j(t) \to w^*$ as $t \to \infty$, for each $j = 1, 2, \ldots, N$.*

Remarks. It is natural to question whether these hypotheses can be satisfied. The graph shown in Figure 2 shows that they can, and also suggests that the hypotheses are likely to be satisfied only when w^* is not too much greater than $w_{\max}$.

The notation $b^{-1}(b(w^*))|_{[0,w_{\max}]}$ requires some explaining. Under the assumptions on $b(w)$, $b^{-1}(w)$ will, if defined, have in general two values. Thus $b^{-1}(b(w^*))$ is either w^* or a value in $[0, w_{\max}]$ and $b^{-1}(b(w^*))|_{[0,w_{\max}]}$ means the latter value.

Proof of Theorem 5.3. Let us define

$$\begin{aligned} w^0_{\max} &= \max\{w^*, \max\{w_j^0(s),\ j = 1, 2, \ldots, N,\ \ s \in [-r, 0]\}\}, \\ w^0_{\min} &= \min\{w^*, \min\{w_j^0(s),\ j = 1, 2, \ldots, N,\ \ s \in [-r, 0]\}\}. \end{aligned}$$

We can assume without loss of generality that $w^0_{\min} > 0$ (this can be justified similarly to the proof of Theorem 4.2).

The proof begins with the observation that $(\hat{w}_j(t), \bar{w}_j(t)) = (0, V_1(t))$ is a sub/supersolution pair for (2.10), (2.11), where $V_1(t)$ satisfies

$$\frac{dV_1(t)}{dt} = \mu b(w_{\max}) - d_m V_1(t),$$

$$V_1(s) = w^0_{\max},\ s \in [-r, 0].$$

Therefore,

$$\limsup_{t\to\infty} \max_{j\in\{1,2,\ldots,N\}} w_j(t) \le \lim_{t\to\infty} V_1(t) = \frac{\mu b(w_{\max})}{d_m}.$$

Next it is easy to see that (v_1, V_1) is a sub/supersolution pair, where V_1 is the above function and v_1 satisfies

$$\frac{dv_1}{dt} = \mu \min\{b(v_1(t-r)), b(V_1(t-r))\} - d_m v_1(t), \tag{5.9}$$
$$v_1(s) = w^0_{\min},\ s \in [-r, 0].$$

Inequality (5.7) assures us that the solution of

$$\frac{dv}{dt} = \mu b(v(t-r)) - d_m v(t)$$

with positive initial data satisfies $\lim_{t\to\infty} v(t) = w^*$. This follows from Theorem 9.5 in Kuang (see [9, page 165]). On the other hand, any solution of

$$\frac{dv}{dt} = \mu b(V_1(t-r)) - d_m v(t)$$

satisfies

$$\lim_{t\to\infty} v(t) = \frac{\mu}{d_m} b\left(\frac{\mu b(w_{\max})}{d_m}\right).$$

Therefore the solution $v_1(t)$ of (5.9) tends, as $t \to \infty$, to either w^* or $\frac{\mu}{d_m} b(\frac{\mu b(w_{\max})}{d_m})$ and so our proof proceeds by considering two cases.

Case 1. $\lim_{t\to\infty} v_1(t) = w^*$. In this case we obtain immediately that

$$\liminf_{t\to\infty} \min_{j\in\{1,2,\dots,N\}} w_j(t) \geq w^*.$$

It can be shown similarly that

$$\limsup_{t\to\infty} \max_{j\in\{1,2,\dots,N\}} w_j(t) \leq w^*,$$

completing the proof of the theorem for Case 1.

Case 2. $\lim_{t\to\infty} v_1(t) = \frac{\mu}{d_m} b(\frac{\mu b(w_{\max})}{d_m})$. What we initially get in this case is

$$\liminf_{t\to\infty} \min_{j\in\{1,2,\dots,N\}} w_j(t) \geq \frac{\mu}{d_m} b\left(\frac{\mu b(w_{\max})}{d_m}\right),$$

which, by (5.6), strictly exceeds $w_{\max}$. Therefore, for t sufficiently large, $w_j(t) > w_{\max}$. Since the problem has finite delay, this further means that for t sufficiently large there is no history of $w_j(t)$ ever having assumed values below $w_{\max}$, so that for the remainder of the proof $b(w)$ can be treated as decreasing in w. In fact we may shift the origin of time such as to assume, without loss of generality, that $w^0_{\min} > w_{\max}$. With this fact in mind, our proof now proceeds by successive refinement of pairs of sub/supersolutions.

In general, for each $n = 2, 3, \dots$, let (v_n, V_n) be defined by

$$\frac{dv_n}{dt} = \mu b(V_{n-1}(t-r)) - d_m v_n, \tag{5.10}$$
$$\frac{dV_n}{dt} = \mu b(v_{n-1}(t-r)) - d_m V_n$$

with the initial conditions $V_n(s) = w^0_{\max},\ s \in [-r, 0]$, and $v_n(s) = w^0_{\min},\ s \in [-r, 0]$. We will show that (v_n, V_n) is a sub/supersolution pair for each $n = 2, 3, \ldots$. According to Theorem 5.1, what we need to show is that

$$b(V_{n-1}(t-r)) \le \sum_{k=1}^{N} b(\varphi_k(t-r))\beta(\alpha, k, j)$$

whenever $\varphi_j(t)$ is such that $v_n(t) \le \varphi_j(t) \le V_n(t)$ for $j = 1, 2, \ldots, N$ and $t \in [-r, \infty)$. Since we are now working in an interval of w in which $b(w)$ is decreasing, it is enough to establish that

$$b(V_{n-1}(t-r)) \le \sum_{k=1}^{N} b(V_n(t-r))\beta(\alpha, k, j) = b(V_n(t-r))$$

and the latter is true if $V_n(t) \le V_{n-1}(t)$ for each $n = 2, 3, \ldots$. Similarly, we need to show that $v_n(t) \ge v_{n-1}(t)$ for each n. In fact, we shall show by induction that

$$v_1(t) \le \cdots \le v_{n-1}(t) \le v_n(t) < w^* < V_n(t) \le V_{n-1}(t) \le \cdots \le V_1(t). \tag{5.11}$$

To achieve this we assume (5.11) and prove that

$$w^* < V_{n+1}(t) \le V_n(t) \tag{5.12}$$

and

$$w^* > v_{n+1}(t) \ge v_n(t). \tag{5.13}$$

We will prove only (5.12). Now

$$\begin{aligned}\frac{dV_{n+1}(t)}{dt} &= \mu b(v_n(t-r)) - d_m V_{n+1}(t)\\ &\le \mu b(v_{n-1}(t-r)) - d_m V_{n+1}(t).\end{aligned}$$

Therefore

$$\frac{dV_{n+1}(t)}{dt} + d_m V_{n+1}(t) \le \frac{dV_n(t)}{dt} + d_m V_n(t),$$

or, equivalently,

$$\frac{d}{dt}(V_{n+1} - V_n)(t) + d_m(V_{n+1} - V_n)(t) \le 0.$$

Therefore

$$V_{n+1}(t) - V_n(t) \le \underbrace{(V_{n+1}(0) - V_n(0))}_{=0} e^{-d_m t}$$

and so

$$V_{n+1}(t) \le V_n(t).$$

Also

$$\begin{aligned}\frac{dV_{n+1}(t)}{dt} &= \mu b(v_n(t-r)) - d_m V_{n+1}(t)\\ &\ge \mu b(w^*) - d_m V_{n+1}(t)\\ &= d_m(w^* - V_{n+1})\end{aligned}$$

so that

$$\frac{d}{dt}(V_{n+1} - w^*) \geq -d_m(V_{n+1} - w^*).$$

Therefore,

$$V_{n+1}(t) - w^* \geq \underbrace{(V_{n+1}(0) - w^*)}_{\geq 0} e^{-d_m t} \geq 0$$

and so

$$V_{n+1}(t) \geq w^*$$

which establishes (5.12).

Denote $V_n^* = \lim_{t\to\infty} V_n(t)$ and $v_n^* = \lim_{t\to\infty} v_n(t)$. That these limits exist follows from (5.10) and an inductive argument. We know $v_1(t)$ and $V_1(t)$ approach limits as $t \to \infty$. There are theories of asymptotically autonomous differential equations which allow us to let $t \to \infty$ in system (5.10), with $n = 2$, giving an autonomous system of differential equations from which it becomes clear that $v_2(t)$ and $V_2(t)$ approach limits as $t \to \infty$. The argument can be continued and we thus conclude the existence of the limits as $t \to \infty$ for each n. From (5.10) these limits satisfy

$$\begin{aligned} \mu b(V_{n-1}^*) &= d_m v_n^*, \\ \mu b(v_{n-1}^*) &= d_m V_n^*. \end{aligned} \tag{5.14}$$

We define $V^* = \lim_{n\to\infty} V_n^*$ and $v^* = \lim_{n\to\infty} v_n^*$; then (5.14) reduces to a limiting system

$$\begin{aligned} \mu b(V^*) &= d_m v^*, \\ \mu b(v^*) &= d_m V^*. \end{aligned} \tag{5.15}$$

These equations imply that $V^* = v^* = w^*$ (condition (5.8) assures us that they have no other solutions). Therefore, $\lim_{t\to\infty} w_j(t) = w^*$ and the proof of the theorem is complete. ∎

6. Discussion. The main results in this paper are, first, the derivation of the model (2.10) itself, which is nontrivial and substantially different from the case of an infinite lattice considered previously in [8]. Second, the positivity-preservation properties of solutions which turn out to be analogous to results that can be proved using the strong maximum principle in the case of continuous space. Third, the comparison principle of section 4 for the case of increasing birth functions. Fourth, the use of this comparison principle to prove that the positive equilibrium of system (2.10), if feasible, is globally asymptotically stable for increasing birth functions. Fifth, in section 5, the comparison principle for the case of nonmonotone birth functions is proved. With the help of this principle we have shown that the nontrivial equilibrium is globally asymptotically stable if it is in the interval of w for which the birth function $b(w)$ is increasing, and that it can remain globally stable if it is larger, but not too much larger, than the value of w at which $b(w)$ attains its maximum, assuming additionally the delay is not too large.

Acknowledgment. We would like to thank the referees for their very valuable comments.

REFERENCES

[1] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Classics Appl. Math. 9, SIAM, Philadelphia, 1994.

[2] G. Bocharov and K. P. Hadeler, *Structured population models, conservation laws, and delay equations*, J. Differential Equations, 168 (2000), pp. 212–237.

[3] S. A. Gourley and J. Wu, *Extinction and periodic oscillations in an age-structured population model in a patchy environment*, J. Math. Anal. Appl., 289 (2004), pp. 431–445.

[4] J. A. J. Metz and O. Diekmann, eds., *The Dynamics of Physiologically Structured Populations*, Springer-Verlag, New York, 1986.

[5] R. Redlinger, *Existence theorems for semilinear parabolic systems with functionals*, Nonlinear Anal., 8 (1984), pp. 667–682.

[6] H. L. Smith, *Monotone Dynamical Systems. An Introduction to the Theory of Competitive and Cooperative Systems*, AMS, Providence, RI, 1995.

[7] H. L. Smith and H. R. Thieme, *Strongly order preserving semiflows generated by functional-differential equations*, J. Differential Equations, 93 (1991), pp. 332–363.

[8] P. Weng, H. Huang, and J. Wu, *Asymptotic speed of propagation of wave fronts in a lattice delay differential equation with global interaction*, IMA J. Appl. Math., 68 (2003), pp. 409–439.

[9] Y. Kuang, *Delay Differential Equations with Applications in Population Dynamics*, Academic Press, Boston, MA, 1993.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1709–1731

NEW LYAPUNOV FUNCTIONALS OF THE VLASOV–POISSON SYSTEM*

MYEONGJU CHAE† AND SEUNG-YEAL HA‡

Abstract. We present new Lyapunov functionals and L^1-stability for the Vlasov–Poisson system with a self-consistent electrostatic force in a small initial data case. Lyapunov functionals measure possible future crossings of projected particle trajectories in the physical space, and they are nonincreasing along smooth solutions when the initial datum is smooth and decays fast enough at infinity in the phase space. For sufficiently large physical dimensions, we also show that smooth solutions are uniformly L^1-stable.

Key words. Vlasov–Poisson system, Lyapunov functional, L^1-stability

AMS subject classifications. 35A80, 35L50

DOI. 10.1137/040621338

1. Introduction.

Consider a collisionless plasma consisting of M distinct species under the effect of an electromagnetic field. The issue can be understood by the Vlasov–Maxwell system. When the speed of light is taken to be infinity and a magnetic field is ignored, the Vlasov–Poisson system with a self-consistent electrostatic field is addressed. Suppose there are M species of particles with mass m_α, charge q_α, and density in the phase space $f_\alpha(x, v, t)$ for $\alpha = 1, \dots, M$. Here $(x, v) \in \mathbb{R}^N \times \mathbb{R}^N$ denotes position and velocity, respectively. The self-consistent electric field is denoted by $E = -\nabla_x \phi$ and the total charge density ρ is

$$\rho(x,t) = \sum_{\alpha=1}^{M} q_\alpha \int_{\mathbb{R}^N} f_\alpha(x,v,t)\, dv.$$

In this case, the Vlasov–Poisson system reads

$$\begin{cases} \partial_t f_\alpha + v \cdot \nabla_x f_\alpha + \dfrac{q_\alpha}{m_\alpha} \nabla_x \phi \cdot \nabla_v f_\alpha = 0 \quad \text{in} \quad \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}_+, \\ -\Delta_x \phi = N(N-2)\omega_N \rho, \end{cases} \tag{1.1}$$

with the prescribed initial datum

$$f_\alpha(x,v,0) = f_{\alpha,0}(x,v) \quad \text{in} \quad \mathbb{R}^N \times \mathbb{R}^N, \tag{1.2}$$

where ω_N is the volume of the unit ball in $\mathbb{R}^N$.

The Vlasov–Poisson system has many applications in the modeling of an electron gun, plasma sheath, and galaxies as a large ensemble of stars in plasma physics and astrophysics, respectively. The global existence of a smooth solution in one space variable has been proved by Iordanskii [25] and that of two space variables by Ukai

*Received by the editors December 23, 2004; accepted for publication (in revised form) May 3, 2005; published electronically February 15, 2006.
http://www.siam.org/journals/sima/37-6/62133.html

†School of Mathematics, KIAS 207-43, Cheongnyangni 2-dong, Dongdaemun-gu, Seoul 130-722, Korea (mchae@kias.re.kr). This author's research was supported by a research grant from KIAS.

‡Department of Mathematics, Seoul National University, Seoul 151-747, Korea (syha@math.snu.ac.kr). This author's research was partially done when he visited Prof. Tai-Ping Liu at the Institute of Mathematics, Academia Sinica in Taiwan during the summer of 2004.

and Okabe [36]. On the other hand, for the three-dimensional case, Batt [3] and Horst [21, 22, 23] established the global existence for spherical and cylindrically symmetric data, respectively. For general but small data, Bardos and Degond [2] obtained global existence, and finally Pfaffelmoser [30] proved the global existence of a smooth solution with large data, and simpler proofs were provided by Schaeffer [33] and Lions and Perthame [27]. For other issues such as weak solutions, relativistic effects, stability, and dispersion estimates, we refer the reader to [1, 4, 12, 23, 28, 29, 31, 34, 35, 38, 40].

The aim of this paper is twofold. First, we present Lyapunov functionals for the repulsive Vlasov–Poisson system when the dimension of the physical space $N \neq 2$. These functionals measure the future collisions between charged particles, although the Vlasov–Poisson system does not register them. As in the Boltzmann equation [19, 20], these functionals might be useful to establish L^1-scattering-type results for the Vlasov–Poisson–Boltzmann equation near vacuum [18]. Second, we study the L^1-stability of smooth C^2-solutions to (1.1) using Gronwall-type estimates when the dimension of the physical space is sufficiently large, and the initial data are smooth and decay fast enough in the phase space. Throughout this paper, we denote by C a universal positive constant independent of time t.

The above two main results of this paper are subject to the following estimates on the time-decay of the electric field and the velocity variation of the density function f_α (see section 2).

(E1) The electric field $E(x,t)$ decays, and it is integrable in t:

$$\int_0^\infty \|E(t)\|_{L^\infty(\mathbb{R}^N_x)}\, dt < \infty.$$

(E2) For fixed $(x,t) \in \mathbb{R}^N \times \mathbb{R}_+$ and $\alpha \in \{1,\dots,M\}$,

$$\Big\| \int_{\mathbb{R}^N} |\nabla_v f_\alpha(t)|\, dv \Big\|_{L^\infty(\mathbb{R}^N_x)} \le \frac{C}{(1+t)^{N-1}},$$
$$\Big\| \int_{\mathbb{R}^N} |\nabla_v f_\alpha(t)|\, dv \Big\|_{L^1(\mathbb{R}^N_x)} \le C(1+t).$$

Based on the time-integrable decay of the electric field E, we first construct an interaction potential $\mathcal{D}(f_\alpha)$ which is nonincreasing along smooth solutions to (1.1) with $N \ge 3$, and which satisfies a Lyapunov estimate: for $\alpha \in \{1,\dots,M\}$,

$$\mathcal{D}(f_\alpha(t)) + \int_0^t \Lambda(f_\alpha(s))\, ds = \mathcal{D}(f_{\alpha,0}), \qquad t \ge 0,$$

where $\Lambda(f_\alpha(s)) \ge 0$ is an interaction production functional

$$\Lambda(f_\alpha(t)) := \iiint_{\mathbb{R}^N\times\mathbb{R}^N\times\mathbb{R}^N} f_\alpha(x,v,t) f_\alpha(x,v_*,t)\, dv_*\, dv\, dx.$$

The above functional measures the possible crossings of the projected particle trajectories in the physical space, and hence it can be regarded as a kinetic counterpart of Glimm's interaction potential [17] in hyperbolic conservation laws in one-space dimension. Due to the strong dispersive effect of the free transport part, the crossings of projected trajectories in the physical space will tend to be nonincreasing in time t. In fact, in the absence of external forces, such interaction potentials have been constructed for the full Boltzmann equation in [6, 19, 20] based on the free transport equation $\partial_t f + v \cdot \nabla_x f = 0$ with small initial data. Unlike the Boltzmann equation, where the interaction potential is constructed based on the free transport part, the external force term $\nabla_x \phi \cdot \nabla_v f_\alpha$ produces error terms which cannot be controlled by the interaction production functional $\Lambda(f_\alpha(t))$. We overcome this difficulty by devising a new functional capturing the nonlinear feature of the system, and by essentially using time-integrable decay of the electric field; thus small initial data are assumed.

Second, we establish the uniform L^1-stability for (1.1) with $N \geq 4$:

$$\sup_{0 \leq t < \infty} \|f(t) - \bar{f}(t)\|_{L^1} \leq G \|f_0 - \bar{f}_0\|_{L^1}, \tag{1.3}$$

where G is a positive constant independent of t, and a simplified notation for L^1-norm is used:

$$\|f(t) - \bar{f}(t)\|_{L^1} := \sum_{\alpha=1}^{M} \|f_\alpha(\cdot, \cdot, t) - \bar{f}_\alpha(\cdot, \cdot, t)\|_{L^1(\mathbb{R}^N \times \mathbb{R}^N)}.$$

Unlike the Boltzmann equation [19, 20], the nonlinear functional approach incorporating the Lyapunov functional $\mathcal{D}(f_\alpha)$ cannot be applied to L^1-stability estimates for (1.1). By direct calculation, $|f_\alpha - \bar{f}_\alpha|$ satisfies

$$\partial_t |f_\alpha - \bar{f}_\alpha| + v \cdot \nabla_x |f_\alpha - \bar{f}_\alpha| + E(f) \cdot \nabla_v |f_\alpha - \bar{f}_\alpha| \leq |E(\bar{f}) - E(f)| |\nabla_v \bar{f}_\alpha|, \tag{1.4}$$

which yields

$$\frac{d}{dt} \|f_\alpha(t) - \bar{f}_\alpha(t)\|_{L^1} \leq \iint_{\mathbb{R}^N \times \mathbb{R}^N} |E(\bar{f}) - E(f)| |\nabla_v \bar{f}_\alpha| \, dv \, dx. \tag{1.5}$$

In contrast, the time-derivative of $\mathcal{D}(|f_\alpha - \bar{f}_\alpha|)$ based on the left-hand sides of (1.1) and (1.4) yields good decay terms:

$$-\iiint_{\mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N} |f_\alpha - \bar{f}_\alpha|(x, v, t) f_\alpha(x, v_*, t) \, dv_* \, dv \, dx;$$

however, these terms cannot control the right-hand side of (1.5). In fact these good terms can be used to control the error terms due to the difference $Q(f, f) - Q(\bar{f}, \bar{f})$ in the L^1-stability estimates of the Vlasov–Poisson–Boltzmann system (see [19, 20] for the related issue of the Boltzmann equation); hence instead of using the nonlinear functional approach for (1.3), we employ Gronwall-type estimates incorporating (E2) in Theorem 1.2 (see section 4):

$$\|f(t) - \bar{f}(t)\|_{L^1} \leq \|f_0 - \bar{f}_0\|_{L^1} + C \int_0^t (1 + s)^{-(N-2)} \|f(s) - \bar{f}(s)\|_{L^1} \, ds.$$

Since

$$\int_0^t (1 + s)^{-(N-2)} < C \quad \text{for} \quad N \geq 4,$$

we can establish the uniform L^1-stability (1.3) for the physical space dimension $N \geq 4$.

The main hypotheses (H) employed in this paper are as follows:

(H1) The initial data $f_{\alpha,0}$ $(\alpha = 1, \ldots, M)$ are twice continuously differentiable:

$$f_{\alpha,0} \in C^2(\mathbb{R}^N \times \mathbb{R}^N).$$

(H2) The initial data are small and decay at infinity in the phase space:

$$\max_{\alpha=1}^{M} \sum_{0\le i,j\le 2} \sup_{x,v} (1+|x|^2)^{\frac{\mu_1}{2}}(1+|v|^2)^{\frac{\mu_2}{2}} |\nabla_x^i \nabla_v^j f_{\alpha,0}(x,v)| \le \varepsilon_0,$$

where $\mu_1 > N+2$, $\mu_2 > N+1$, and $0 < \varepsilon_0 \ll 1$.

Our first main result shows that the system (1.1) with $N \ge 3$ admits an interaction potential.

THEOREM 1.1. *Suppose the main hypotheses* (H) *with* $N \ge 3$ *hold, and let* f_α $(1 \le \alpha \le M)$ *be a smooth solution to* (1.1) *corresponding to the initial datum* $f_{\alpha,0}$. *Then there exists an interaction potential* $\mathcal{D}(f_\alpha)$ *satisfying a Lyapunov estimate:*

$$\mathcal{D}(f_\alpha(t)) + \int_0^t \Lambda(f_\alpha(s))\, ds = \mathcal{D}(f_{\alpha,0}), \quad t \ge 0.$$

REMARK 1.1. (1) *Following the arguments in* [2]*, the global existence of smooth* C^2*-solutions satisfying the main estimates* (E1)–(E2) *can be shown.*

(2) *For* $N = 1$*, we can construct a similar Lyapunov functional under the monotonicity assumption of the electric field (see section* 3.1*).*

Our second main theorem is concerned with the uniform L^1-stability.

THEOREM 1.2. *Suppose the main hypotheses* (H) *with* $N \ge 4$ *hold, and let* f_α *and* $\bar{f}_\alpha$ *be smooth solutions to* (1.1) *corresponding to the initial data* $f_{\alpha,0}$ *and* $\bar{f}_{\alpha,0}$, *respectively. Then smooth solutions are uniformly* L^1*-stable with respect to the initial data:*

$$\sup_{0\le t<\infty} \|f(t) - \bar{f}(t)\|_{L^1} \le G\|f_0 - \bar{f}_0\|_{L^1},$$

where G *is a positive constant independent of time* t.

The rest of this paper is organized as follows. In section 2, we review basic a priori estimates for the Vlasov–Poisson system, and in section 3 we construct kinetic Glimm-type interaction potentials for the one-dimensional (1D) and multidimensional cases, respectively. In section 4, we study the uniform L^1-stability estimate. Finally, in the appendix, we briefly discuss an interaction potential for the relativistic Vlasov–Maxwell system with small data.

2. Preliminaries. We briefly review dispersion estimates for the local charge density, the electric force, and the velocity variation of the density function, and we study the strict separation property of backward particle trajectories. The details can be found in [2, 7].

2.1. Time-decay estimate of the electric field. In this part, we review the dispersion estimates studied in [2]. Consider C^2-initial data for the system (1.1) satisfying

$$\max_{\alpha=1}^{M} \sum_{0\le i,j\le 2} \sup_{x,v} (1+|x|^2)^{\frac{\mu_1}{2}}(1+|v|^2)^{\frac{\mu_2}{2}} |\nabla_x^i \nabla_v^j f_{\alpha,0}(x,v)| \le \varepsilon_0,$$

where $\mu_1 > N+2$, $\mu_2 > N+1$, and $0 < \varepsilon_0 \ll 1$.

Let $h \in C(\mathbb{R}_+; (L^1 \cap W^{1,\infty})(\mathbb{R}^N))$ be given, and define a Newtonian potential $\mathcal{P}$ by the representation formula

$$\mathcal{P}(h)(x,t) := \int_{\mathbb{R}^N} \frac{h(y,t)}{|x-y|^{N-2}}\, dy. \tag{2.1}$$

The following lemmas can be found in [2] for $N = 3$, and their proofs are exactly the same as those for $N > 3$; hence we omit its proof.

LEMMA 2.1. *Suppose that a function* $h(\cdot, t) \in (L^1 \cap W^{1,\infty})(\mathbb{R}^N)$ *for all* $t \in \mathbb{R}^+$. *Then we have*

(1) $\|\mathcal{P}(h)(t)\|_{L^\infty} \le C(N)\|h(t)\|_{L^1}^{\frac{2}{N}}\|h(t)\|_{L^\infty}^{\frac{N-2}{N}}$.

(2) $\|\nabla_x\mathcal{P}(h)(t)\|_{L^\infty} \le C(N)\|h(t)\|_{L^1}^{\frac{1}{N}}\|h(t)\|_{L^\infty}^{\frac{N-1}{N}}$.

(3) $\|\nabla_x^2\mathcal{P}(h)(t)\|_{L^\infty} \le C(N,\kappa)\|h(t)\|_{L^1}^{\frac{\kappa}{\kappa+N}}\|h(t)\|_{L^\infty}^{\frac{N(1-\kappa)}{N+\kappa}}\|\nabla_x h(t)\|_{L^\infty}^{\frac{N\kappa}{N+\kappa}}$ *for* $\kappa \in (0, 1)$.

A direct modification of the proof given in [2] for $N = 3$ and Lemma 2.1 give the global existence of smooth C^2-solutions under the main hypotheses (H).

THEOREM 2.2 (see [2]). *Suppose the main hypotheses* (H) *hold. Then the Vlasov–Poisson system* (1.1) *has a unique global in time solution such that* $f_\alpha \in C^1(\mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}_+)$ *satisfies the following uniform estimate:*

$$\|f(t)\|_{L^\infty(\mathbb{R}^N\times\mathbb{R}^N)} \le \varepsilon_0, \quad \|f(t)\|_{L^1(\mathbb{R}^N\times\mathbb{R}^N)} \le C\varepsilon_0,$$
$$\Big\| \int f(t)\, dv \Big\|_{L^\infty(\mathbb{R}^N_x)} \le \frac{C\varepsilon_0}{(1+t)^N}.$$

Moreover, we have for any $\kappa \in (0, 1)$,

$$\max_{1\le k\le 2} \|\nabla_x^k E(\cdot, t)\|_{L^\infty} \le \frac{C\varepsilon_0}{(1+t)^{\frac{N^2}{N+\kappa}}}.$$

REMARK 2.1. *The time-decay rate* $\frac{N^2}{N+\kappa}$ *has a lower bound which is strictly greater than* 2 (*see* [2]).

$$e(N,\kappa) := \frac{N^2}{N+\kappa} \ge \frac{N^2}{N+1} > 2 \quad \textit{for} \;\; N \ge 3. \tag{2.2}$$

The decay estimates of the electric field and the initial data yield the following estimate. The detailed proof for $N = 3$ can be found in [2], and it can be directly applied for the general case $N \ge 4$.

LEMMA 2.3 (see [2]). *Suppose the main hypotheses* (H) *hold, and let* f *be the smooth solution to* (1.1). *Then we have*

$$\Big\| \int_{\mathbb{R}^N} |\nabla_v f(t)|\, dv \Big\|_{L^\infty(\mathbb{R}^N_x)} \le \frac{C\varepsilon_0}{(1+t)^{N-1}} \quad \textit{and} \quad \Big\| \int_{\mathbb{R}^N} |\nabla_v f(t)|\, dv \Big\|_{L^1(\mathbb{R}^N_x)} \le C\varepsilon_0(1+t).$$

The above estimate will be used in the L^1-stability analysis in section 4.

2.2. Particle trajectories. In this part, we study the "*strict separation property*" of particle trajectories when an external force decays sufficiently fast in time t. Consider a Vlasov equation with a sufficiently smooth electric field E:

$$\partial_t f + v \cdot \nabla_x f + E \cdot \nabla_v f = 0. \tag{2.3}$$

We define a particle trajectory as the characteristic curve related to (2.3), i.e., for a fixed point $(x, v) \in \mathbb{R}^N \times \mathbb{R}^N$ at time t, the particle trajectory $[X(s; t, x, v), V(s; t, x, v)]$ passing through (x, v) at time t is defined as the unique solution of the ODE system:

$$\frac{d}{ds}X(s; t, x, v) = V(s; t, x, v) \quad \text{and} \quad \frac{d}{ds}V(s; t, x, v) = E(X(s; t, x, v), s) \tag{2.4}$$

with an initial datum

$$X(t;t,x,v) = x \quad \text{and} \quad V(t;t,x,v) = v.$$

We integrate (2.4) along the particle trajectory to see that Liouville's principle holds:

$$f(x,v,t) = f_0(X(0;t,x,v), V(0;t,x,v)), \qquad t \geq 0;$$

therefore, the L^p-norm of f is conserved for every p, $1 \leq p \leq \infty$,

$$\|f(t)\|_{L^p} = \|f_0\|_{L^p}, \quad t \geq 0,$$

since the mapping $(x,v) \to (X,V)$ is measure preserving.

Note that in the absence of an external field ($F \equiv 0$), the particle trajectory is simply a straight line:

$$X(s;t,x,v) = x - v(t-s) \quad \text{and} \quad V(s;t,x,v) = v.$$

Hence particle trajectories with different initial velocities are well separated in the sense that

$$|X(s;t,x,v) - X(s;t,x,w)| = (t-s)|v-w|.$$

While external forces such as an electromagnetic force act on particles, trajectories are no longer straight lines; they, nevertheless, behave almost like straight lines as long as the external force is sufficiently small and decays fast enough in time. The following lemma shows this strict separation property of particle trajectories.

LEMMA 2.4 (see [7]). *Suppose the main hypotheses* (H) *in section* 1 *hold and* $N \geq 3$. *Then particle trajectories are strictly separated: for fixed* x, v, t, *there exists a positive constant* C_1 *independent of* t *and* s *such that*

$$\frac{1}{C_1}|t-s||v-w| \leq |X(s;t,x,v) - X(s;t,x,w)| \leq C_1|t-s||v-w|, \qquad 0 \leq s \leq t.$$

Proof. Let x, v, w, and t be given, and for the simplicity of notation set

$$\begin{aligned} &X_1(s) := X(s;t,x,v), \quad V_1(s) := V(s;t,x,v) \quad \text{and} \\ &X_2(s) := X(s;t,x,w), \quad V_2(s) := V(s;t,x,w). \end{aligned}$$

Then it follows from (2.4) that

$$X_1(s) = x - \int_s^t V_1(\tau)\,d\tau \quad \text{and} \quad X_2(s) = x - \int_s^t V_2(\tau)\,d\tau.$$

We use the above two equations to find

$$\begin{aligned} X_1(s) - X_2(s) &= -\int_s^t (V_1(\tau) - V_2(\tau))\,d\tau \\ &= \int_s^t \left[\int_\tau^t (\partial_\theta V_1(\theta) - \partial_\theta V_2(\theta))\,d\theta - (v-w)\right] d\tau \\ &= \int_s^t \int_\tau^t (E(X_1(\theta),\theta) - E(X_2(\theta),\theta))d\theta\,d\tau - (v-w)(t-s). \end{aligned} \tag{2.5}$$

We set

$$A(s) := |X_1(s) - X_2(s) + (t-s)(v-w)|.$$

Then (2.5) implies

$$\begin{aligned}
A(s) &\le \int_s^t \int_\tau^t \|\nabla_x E(\cdot,\theta)\|_{L^\infty} |X_2(\theta) - X_1(\theta)|\, d\theta\, d\tau \\
&\le \int_s^t \int_\tau^t \|\nabla_x E(\cdot,\theta)\|_{L^\infty} (|X_2(\theta) - X_1(\theta) + (t-\theta)(v-w)| + (t-\theta)|v-w|)\, d\theta\, d\tau \\
&\le \mathcal{O}(\varepsilon_0) \int_s^t \int_\tau^t \left(\frac{A(\theta)}{(1+\theta)^{e(N,\kappa)}} + \frac{(t-\theta)|v-w|}{(1+\theta)^{e(N,\kappa)}} \right) d\theta\, d\tau \\
&\le \mathcal{O}(\varepsilon_0) \int_s^t \int_s^\theta \left(\frac{A(\theta)}{(1+\theta)^{e(N,\kappa)}} + \frac{t-\theta}{(1+\theta)^{e(N,\kappa)}} |v-w| \right) d\tau\, d\theta \\
&\le \mathcal{O}(\varepsilon_0) \int_s^t \left(\frac{(\theta-s)A(\theta)}{(1+\theta)^{e(N,\kappa)}} + \frac{(t-\theta)(\theta-s)}{(1+\theta)^{e(N,\kappa)}} |v-w| \right) d\theta \\
&\le \mathcal{O}(\varepsilon_0) \left(\int_s^t \frac{A(\theta) d\theta}{(1+\theta)^{e(N,\kappa)-1}} + (t-s)|v-w| \right),
\end{aligned}$$

where we used Theorem 2.2 and

$$e(N,\kappa) > 2, \quad \theta - s \le \theta + 1, \quad \text{and} \quad \int_s^t \frac{(t-\theta)(\theta-s)}{(1+\theta)^{e(N,\kappa)}}\, d\theta \le \mathcal{O}(1)|t-s|.$$

Since

$$\exp\left(\int_0^\infty \frac{d\theta}{(1+\theta)^{e(N,\kappa)-1}} \right) = \mathcal{O}(1),$$

Gronwall's inequality implies

$$A(s) \le \mathcal{O}(\varepsilon_0)(t-s)|v-w|.$$

By a triangle inequality, we have

$$(1 - \mathcal{O}(\varepsilon_0))(t-s)|v-w| \le |X_1(s) - X_2(s)| \le (1 + \mathcal{O}(\varepsilon_0))(t-s)|v-w|.$$

Since $\varepsilon_0 \ll 1$, we obtain the desired result. ∎

LEMMA 2.5 (see [18]). *Suppose the main hypotheses* (H) *holds. For any fixed* (x, v, t)*, let* $[X(s), V(s)]$ *be the particle trajectory passing through the point* (x, v) *at time* t*. Then we have*

$$|V(s_1) - V(s_2)| \le \mathcal{O}(\varepsilon_0) \quad \textit{and} \quad ||V(s_1)|^2 - |V(s_2)|^2| \le \mathcal{O}(\varepsilon_0), \quad 0 \le s_1, s_2 \le \infty.$$

3. Interaction potentials. In this part, we present an interaction potential measuring possible future crossings of projected particle trajectories in the physical space. For the one-space-dimensional case $N = 1$, we consider a plasma consisting of a single species with a positive charge $q = 1$; hence the electric field

$$\partial_x E(x,t) = \int_{\mathbb{R}} f(x,v,t)\, dv \ge 0$$

is monotonic increasing in the space. In contrast, for the multidimensional case $N \ge 3$, we consider a plasma with M species. In the multispecies case, we crucially use the time-integrable decay of electric force.

3.1. 1D Vlasov–Poisson system. Consider the planar motion of single species particles with a positive charge confined inside the self-consistent potential field $E := \partial_x \phi$ as in [37, 39]:

$$\partial_t f + v\partial_x f + E\partial_v f = 0 \quad \text{and} \quad \partial_x E = \int_{\mathbb{R}} f dv. \tag{3.1}$$

Note that the electric field $E(\cdot, t)$ is nondecreasing in x. In what follows, we consider smooth solutions to (3.1); hence particle trajectories $[X(s; x, v, t), V(s; x, v, t)]$ associated with a divergence-free vector field (v, E) do not cross in the phase space. However, their projected trajectories onto the physical space can cross each other, which explains the occurrence of singularities such as shocks in the corresponding macroscopic Euler–Poisson system. For a repulsive external force, particles tend to repel each other; so it is feasible to expect that the number of future crossings of projected trajectories will be nonincreasing in time. In the following, we measure this dispersive phenomenon via an interaction potential. Throughout the paper, we suppress t-dependence in f, i.e.,

$$f(x, v) := f(x, v, t).$$

3.1.1. Construction of an interaction potential. Note that $f(x, v)f(y, v_*)$ satisfies the Vlasov equation on the two-point particle phase space $\mathbb{R}^4$:

(3.2)
$$\begin{aligned}\partial_t(f(x, v)f(y, v_*)) &+ \nabla_{(x,y)} \cdot (vf(x, v)f(y, v_*), \quad v_* f(x, v)f(y, v_*)) \\ &+ \nabla_{(v,v_*)} \cdot (E(x, t)f(x, v)f(y, v_*), E(y, t)f(x, v)f(y, v_*)) = 0.\end{aligned}$$

Define an interaction potential $\mathcal{D}_1(f)$ and its production $\Lambda_1(f)$ as follows:

$$\mathcal{D}_1(f(t)) := \underbrace{\iiiint}_{x<y, v>v_*} f(x, v)f(y, v_*)\, dv_*\, dv\, dy\, dx, \tag{3.3}$$

$$\begin{aligned}\Lambda_1(f(t)) := &\iiint_{\mathbb{R}^3} |v - v_*| f(x, v)f(x, v_*)\, dv_*\, dv\, dx \\ &+ \iiint_{\mathbb{R}^3} |E(y, t) - E(x, t)| f(x, v)f(y, v)\, dv\, dy\, dx.\end{aligned}$$

For each $t \in (0, \infty)$, $\mathcal{D}_1(f(t))$ is a priori bounded by the square of the total mass:

$$\mathcal{D}_1(f(t)) \le \|f_0\|_{L^1}^2.$$

3.1.2. Time-decay estimate. In this part, we study the time-decay of the interaction potential.

PROPOSITION 3.1. *Let f be a smooth C^1-solution with an initial datum f_0 with a finite mass, and f decays at infinity in the phase space. Then the functional $\mathcal{D}_1(f)$ satisfies a time-decay estimate*

$$\mathcal{D}_1(f(t)) + \frac{1}{2}\int_0^t \Lambda_1(f(s))\, ds = \mathcal{D}_1(f_0), \quad t \ge 0.$$

Proof. It follows from (3.2) that

$$\begin{aligned}\partial_t[f(x, v)f(y, v_*)] = &-\nabla_{(x,y)} \cdot [vf(x, v)f(y, v_*), v_* f(x, v)f(y, v_*)] \\ &-\nabla_{(v,v_*)} \cdot [E(x, t)f(x, v)f(y, v_*), E(y, t)f(x, v)f(y, v_*)].\end{aligned} \tag{3.4}$$

We integrate (3.4) and use the Fubini theorem to get

$$\begin{aligned}\frac{d}{dt}\mathcal{D}_1(f(t)) &= -\iiiint_{v>v_*,x<y} \nabla_{(x,y)} \cdot [vf(x,v)f(y,v_*), v_*f(x,v)f(y,v_*)]\, dv_*\, dv\, dy\, dx \\ &\quad - \iiiint_{x<y,v>v_*} \nabla_{(v,v_*)} \cdot [E(x,t)f(x,v)f(y,v_*), E(y,t)f(x,v)f(y,v_*)]\, dv_*\, dv\, dy\, dx \\ &= \mathcal{I}_1 + \mathcal{I}_2.\end{aligned}$$

We first estimate $\mathcal{I}_1$. The Fubini and Green theorems yield

$$\begin{aligned}\mathcal{I}_1 &= -\iiiint_{v>v_*,x<y} [\partial_x(vf(x,v)f(y,v_*)) + \partial_y(v_*f(x,v)f(y,v_*))]\, dy\, dx\, dv\, dv_* \\ &= -\iint_{v>v_*} \int_{\mathbb{R}} |v-v_*| f(x,v)f(x,v_*)\, dx\, dv_*\, dv \\ &= -\frac{1}{2}\iiint_{\mathbb{R}^3} |v-v_*| f(x,v)f(y,v_*)\, dv_*\, dv\, dx.\end{aligned}$$

Similarly, we have

$$\mathcal{I}_2 = -\frac{1}{2}\iiint_{\mathbb{R}^3} |E(y,t)-E(x,t)| f(x,v)f(y,v)\, dv\, dy\, dx,$$

where we used the fact that the electric field $E(\cdot,t)$ is nonincreasing: for each t,

$$E(x,t) \le E(y,t) \quad \text{if} \quad x \le y.$$

We finally combine the above two estimates:

$$\frac{d\mathcal{D}_1(f(t))}{dt} + \frac{1}{2}\Lambda_1(f(t)) = 0.$$

This implies

$$\mathcal{D}_1(f(t)) + \frac{1}{2}\int_0^t \Lambda_1(f(s))\, ds = \mathcal{D}_1(f_0). \qquad \square$$

REMARK 3.1. *The above lemma holds for the general* 1D *Vlasov equation with a smooth monotone increasing force field* $F(\cdot,t)$*:*

$$\partial_t f + v\partial_x f + F\partial_v f = 0 \quad \text{and} \quad \partial_x F \ge 0.$$

3.2. Multidimensional Vlasov–Poisson system. In this part, we consider the multidimensional Vlasov–Poisson system with $N \ge 3$. Compared to the 1D case in the previous section, we consider a collisionless plasma consisting of M distinct species. The main difficulty in the multidimensional case is that it is not easy to identify the region of two-point particle phase spaces where interactions occur. Hence we need to devise a more sophisticated functional which takes into account all possible interaction pairs. For this, we will see that the time-integrability of the electric field does play a key role in the following analysis.

3.2.1. Construction of an interaction functional. Let $t \geq 0$ be given, and let (x, v) be the phase space position of a given test particle at time $s = 0$. We denote $[\hat{X}(s), \hat{V}(s)]$ by the trajectory of a test particle,

$$\hat{X}(0) = x \quad \text{ and } \quad \hat{V}(0) = v,$$

and set its terminal velocity v_∞ by

$$v_\infty(x, v) \equiv v + \int_0^\infty E(\hat{X}(s), s)\, ds.$$

Then $v_\infty(x, v)$ is well defined, since the electric field E has an integrable time-decay.

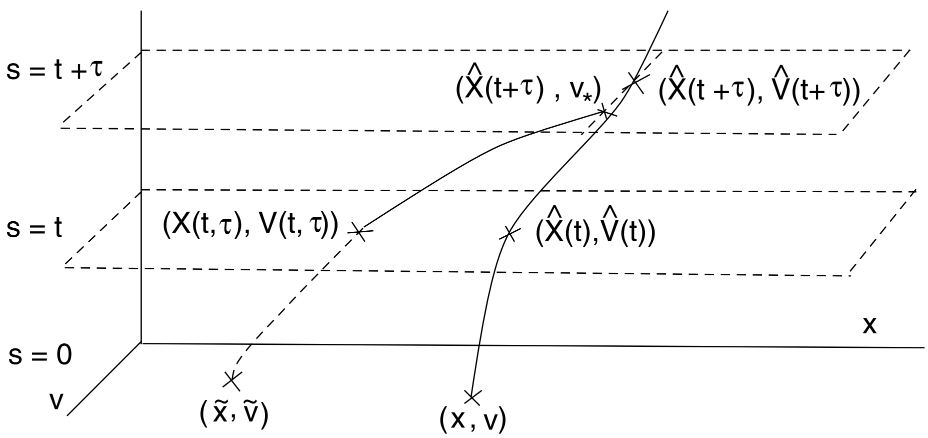


FIG. 3.1. *Schematic diagram of $X(t, \tau)$ and $V(t, \tau)$ in the phase space.*

Let $t, \tau \geq 0$ be given, and for $v_* \notin \{\hat{V}(t + \tau), v_\infty(x, v)\}$, we set (see Figure 3.1)

$$\begin{aligned} X(t, \tau) &:= X(t; t + \tau, \hat{X}(t + \tau), v_*), \quad V(t, \tau) := V(t; t + \tau, \hat{X}(t + \tau), v_*), \\ \tilde{x} &:= X(0; t + \tau, \hat{X}(t + \tau), v_*), \quad \tilde{v} := V(0; t + \tau, \hat{X}(t + \tau), v_*). \end{aligned} \tag{3.5}$$

The invariance of f_α along the trajectory gives

$$f_\alpha(X(t, \tau), V(t, \tau), t) = f_{\alpha,0}(\tilde{x}, \tilde{v}),$$

and we have

$$\begin{aligned} X(t, \tau) &= \hat{X}(t + \tau) - \tau v_* \\ &\quad + \int_t^{t+\tau} \int_{\theta_2}^{t+\tau} E(X(\theta_1; t + \tau, \hat{X}(t + \tau), v_*), \theta_1)\, d\theta_1\, d\theta_2, \end{aligned} \tag{3.6}$$

$$V(t, \tau) = v_* - \int_t^{t+\tau} E(X(\theta_1; t + \tau, \hat{X}(t + \tau), v_*), \theta_1)\, d\theta_1. \tag{3.7}$$

Define an interaction potential and its production functional: for $\alpha \in \{1, \dots, M\}$,

$$\begin{aligned}\mathcal{D}(f_\alpha(t)) &:= \iint_{\mathbb{R}^N\times\mathbb{R}^N} f_\alpha(\hat{X}(t), \hat{V}(t), t) \\ &\qquad \times \left[\int_0^\infty \int_{\mathbb{R}^N} f_\alpha(X(t,\tau), V(t,\tau), t) dv_* d\tau\right] dv\, dx, \\ \Lambda(f_\alpha(t)) &:= \iiint_{\mathbb{R}^N\times\mathbb{R}^N\times\mathbb{R}^N} f_\alpha(x,v,t) f_\alpha(x,v_*,t)\, dv_*\, dv\, dx.\end{aligned} \tag{3.8}$$

Here we used the simplified notation

$$\hat{X}(t) := X(t;0,x,v), \quad \hat{V}(t) := V(t;0,x,v).$$

In order to show that $\mathcal{D}(f_\alpha(t))$ is well defined, i.e., bounded for each t, we need to estimate the quantity inside the bracket of (3.8). For notational simplicity, we denote it by $I_{f_\alpha}(x,v,t)$:

$$I_{f_\alpha}(x,v,t) := \int_0^\infty \int_{\mathbb{R}^3} f_\alpha(X(t,\tau), V(t,\tau), t)\, dv_*\, d\tau.$$

LEMMA 3.2. *Suppose the main hypotheses* (H) *in section* 1 *hold. Then* $I_{f_\alpha}(x,v,t)$ *is finite for a.e.* (x,v,t)*. Moreover, we have the following estimate: for any fixed* $x \in \mathbb{R}^N$,

$$\max_{\alpha=1}^{M} \sup_{v,t} I_{f_\alpha}(x,v,t) \le \mathcal{O}(\varepsilon_0)(|x|+1)^2.$$

Proof. Let $(x,v,t) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}_+$ be given. Since f_α is invariant along the trajectory, it is easy to see that $I_{f_\alpha}(x,v,t)$ is nonincreasing with respect to t; hence it suffices to estimate $I_{f_\alpha}(x,v,0)$. Let $\alpha \in \{1, \dots, M\}$ be fixed and let A be a positive number to be chosen later. We decompose $I_{f_\alpha}(x,v,0)$ into three parts,

$$\begin{aligned}I_{f_\alpha}(x,v,0) \;&=\; \int_0^{A+1} \int_{\mathbb{R}^3} f_{\alpha,0}(X(0,\tau), V(0,\tau))\, dv_*\, d\tau \\ &\quad+ \int_{A+1}^\infty \int_{|v_*-\hat{V}(\tau)|\le \frac{1}{\sqrt{\tau}}} f_{\alpha,0}(X(0,\tau), V(0,\tau))\, dv_*\, d\tau \\ &\quad+ \int_{A+1}^\infty \int_{|v_*-\hat{V}(\tau)|> \frac{1}{\sqrt{\tau}}} f_{\alpha,0}(X(0,\tau), V(0,\tau))\, dv_*\, d\tau \\ &:= I^1_{f_\alpha}(x,v,0) + I^2_{f_\alpha}(x,v,0) + I^3_{f_\alpha}(x,v,0),\end{aligned}$$

where we used the simplified notation (3.5) with $t=0$:

$$X(0,\tau) := X(0;\tau,\hat{X}(\tau),v_*), \quad V(0,\tau) := V(0;\tau,\hat{X}(\tau),v_*).$$

Case 1: $I^1_{f_\alpha}(x,v,0)$. By (H2) and Lemma 2.5,

$$\begin{aligned}I^1_{f_\alpha}(x,v,0) &\le \varepsilon_0 \int_0^{A+1} \int_{\mathbb{R}^N} \frac{1}{(1+|X(0,\tau)|^2)^{\frac{\mu_1}{2}} (1+|V(0,\tau)|^2)^{\frac{\mu_2}{2}}}\, dv_*\, d\tau \\ &\le \varepsilon_0 \int_0^{A+1} \int_{\mathbb{R}^N} \frac{1}{(1+|X(0,\tau)|^2)^{\frac{\mu_1}{2}} (1+|v_*|^2+\mathcal{O}(\varepsilon_0))^{\frac{\mu_2}{2}}}\, dv_*\, d\tau \\ &\le \mathcal{O}(\varepsilon_0)(A+1),\end{aligned}$$

where we used Lemma 2.5 to get

$$\int_{\mathbb{R}^N} \frac{1}{(1+|V(0,\tau)|^2)^{\frac{\mu_2}{2}}}\, dv_* = \int_{\mathbb{R}^N} \frac{1}{(1+|v_*|^2+\mathcal{O}(\varepsilon_0))^{\frac{\mu_2}{2}}}\, dv_* = \mathcal{O}(1). \tag{3.9}$$

Case 2: $I^2_{f_\alpha}(x,v,0)$. Similarly to Case 1, we have

$$\begin{aligned} I^2_{f_\alpha}(x,v,0) &\le \varepsilon_0 \int_{A+1}^{\infty} \int_{|v_*-\hat{V}(\tau)|\le \frac{1}{\sqrt{\tau}}} \frac{1}{(1+|X(0,\tau)|^2)^{\frac{\mu_1}{2}}(1+|V(0,\tau)|^2)^{\frac{\mu_2}{2}}}\, dv_*\, d\tau \\ &\le \mathcal{O}(\varepsilon_0) \int_{A+1}^{\infty} \tau^{-\frac{N}{2}}\, d\tau \\ &= \mathcal{O}(\varepsilon_0)(A+1)^{-\frac{N-2}{2}}. \end{aligned}$$

Case 3: $I^3_{f_\alpha}(x,v,0)$. For v_* satisfying $|v_*-\hat{V}(\tau)| > \frac{1}{\sqrt{\tau}}$, we have

$$|X(0,\tau)-x| \ge \frac{\sqrt{\tau}}{C_1} \quad \text{by Lemma 2.4.}$$

This yields

$$|X(0,\tau)| \ge \frac{\sqrt{\tau}}{C_1} - |x| \ge 0 \quad \text{for } \tau \ge C_1^2|x|^2.$$

We now set

$$A := C_1^2|x|^2.$$

$$\begin{aligned} I^3_{f_\alpha}(x,v,0) &\le \varepsilon_0 \int_{C_1^2|x|^2+1}^{\infty} \int_{|v_*-\hat{V}(\tau)|> \frac{1}{\sqrt{\tau}}} \frac{1}{(1+(\sqrt{\tau}/C_1-|x|)^2)^{\frac{\mu_1}{2}}(1+|V(0,\tau)|^2)^{\frac{\mu_2}{2}}}\, dv_*\, d\tau \\ &\le \mathcal{O}(\varepsilon_0) \int_{C_1^2|x|^2+1}^{\infty} \frac{1}{(1+\sqrt{\tau}/C_1-|x|)^{\mu_1}} \int_{\mathbb{R}^N} \frac{1}{(1+|V(0,\tau)|^2)^{\frac{\mu_1}{2}}}\, dv_*\, d\tau \\ &\le \mathcal{O}(\varepsilon_0) \int_{C_1^2|x|^2+1}^{\infty} \frac{1}{(1+\sqrt{\tau}/C_1-|x|)^{\mu_1}}\, d\tau \qquad \text{by (3.9).} \end{aligned} \tag{3.10}$$

In (3.10), we have

$$\begin{aligned} \int_{C_1^2|x|^2+1}^{\infty} \frac{1}{(1+\sqrt{\tau}/C_1-|x|)^{\mu_1}}\, d\tau &\le 2C_1^2 \int_0^{\infty} \frac{\xi+|x|}{(1+\xi)^{\mu_1}}\, d\xi, \qquad \text{where } \xi := \sqrt{\tau}/C_1-|x| \\ &= \mathcal{O}(1)(1+|x|). \end{aligned}$$

Finally we combine all cases to see

$$I_{f_\alpha}(x,v,0) \le \mathcal{O}(\varepsilon_0)\Big(|x|^2 + (|x|^2+1)^{-\frac{(N-2)}{2}} + |x| + 1\Big) = \mathcal{O}(\varepsilon_0)(|x|+1)^2. \qquad \square$$

REMARK 3.2. *The above lemma implies a priori boundedness of* $\mathcal{D}(f_\alpha)$*:*

$$\mathcal{D}(f_\alpha(t)) \le \mathcal{O}(\varepsilon_0^2) \qquad \textit{for} \quad t \ge 0.$$

3.2.2. Time-decay estimate. In this part we study the time-evolution of the functional $\mathcal{D}(f_\alpha)$ along smooth solutions to (1.1). We first estimate the difference between ∂_t and ∂_τ derivatives of $X(t,\tau)$ and $V(t,\tau)$, respectively.

LEMMA 3.3. *For any fixed $(x,v,t) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}_+$ and α, let $[X(t,\tau), V(t,\tau)]$ be given by* (3.6) *and* (3.7). *Then we have*

$$\partial_t X(t,\tau) - \partial_\tau X(t,\tau) = V(t,\tau) \quad \text{and} \quad \partial_t V(t,\tau) - \partial_\tau V(t,\tau) = E(X(t,\tau),t).$$

Proof. Recall that

$$X(t,\tau) = \hat{X}(t+\tau) - \tau v_* + \int_t^{t+\tau} \int_{\theta_2}^{t+\tau} E(X(\theta_1; t+\tau, \hat{X}(t+\tau), v_*), \theta_1)\, d\theta_1\, d\theta_2.$$

We take ∂_t and ∂_τ to the above identity to get

- $$\partial_t X(t,\tau) = \hat{V}(t+\tau) + \tau E(\hat{X}(t+\tau), t+\tau) - \int_t^{t+\tau} E(X(\theta_1; t+\tau, \hat{X}(t+\tau), v_*), \theta_1)\, d\theta_1, \\ + \int_t^{t+\tau} \int_{\theta_2}^{t+\tau} \partial_t (E(X(\theta_1; t+\tau, \hat{X}(t+\tau), v_*), \theta_1))\, d\theta_1\, d\theta_2,$$
- $$\begin{aligned} \partial_\tau X(t,\tau) &= \hat{V}(t+\tau) - v_* + \tau E(\hat{X}(t+\tau), t+\tau) \\ &\quad + \int_t^{t+\tau} \int_{\theta_2}^{t+\tau} \partial_\tau (E(X(\theta_1; t+\tau, \hat{X}(t+\tau), v_*), \theta_1))\, d\theta_1\, d\theta_2 \\ &= \hat{V}(t+\tau) - \left(V(t,\tau) + \int_t^{t+\tau} E(X(\theta_1; t+\tau, \hat{X}(t+\tau), v_*), \theta_1)\, d\theta_1 \right) \\ &\quad + \tau E(\hat{X}(t+\tau), t+\tau) + \int_t^{t+\tau} \int_{\theta_2}^{t+\tau} \partial_\tau (E(X(\theta_1; t+\tau, \hat{X}(t+\tau), v_*), \theta_1)) d\theta_1 d\theta_2. \end{aligned}$$

Note that

$$\partial_t (E(X(\theta_1; t+\tau, \hat{X}(t+\tau), v_*), \theta_1)) = \partial_\tau (E(X(\theta_1; t+\tau, \hat{X}(t+\tau), v_*), \theta_1))$$

to get

$$\partial_t X(t,\tau) - V(t,\tau) = \partial_\tau X(t,\tau).$$

Again we use the relation

$$V(t,\tau) = v_* - \int_t^{t+\tau} E(X(\theta_1; t+\tau, \hat{X}(t+\tau), v_*), \theta_1)\, d\theta_1$$

to find

- $$\begin{aligned} \partial_t V(t,\tau) &= E(X(t,\tau),t) - E(\hat{X}(t+\tau), t+\tau) \\ &\quad - \int_t^{t+\tau} \partial_t (E(X(\theta_1; t+\tau, \hat{X}(t+\tau), v_*), \theta_1))\, d\theta_1, \end{aligned}$$
- $$\begin{aligned} \partial_\tau V(t,\tau) &= -E(\hat{X}(t+\tau), t+\tau) \\ &\quad - \int_t^{t+\tau} \partial_t (E(X(\theta_1; t+\tau, \hat{X}(t+\tau), v_*), \theta_1))\, d\theta_1. \end{aligned}$$

The above relations yield

$$\partial_t V(t,\tau) = \partial_\tau V(t,\tau) + E(X(t,\tau),t). \qquad \square$$

LEMMA 3.4. *For any* $(x,v,t) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}_+$ *and* $\alpha \in \{1,\dots,M\}$, *let* $[\hat{X}(t),\hat{V}(t)]$ *be the trajectory of a test particle. Then we have*

$$\begin{aligned}&\partial_t(f_\alpha(\hat{X}(t),\hat{V}(t),t)f_\alpha(X(t,\tau),V(t,\tau),t))\\ &= \partial_\tau(f_\alpha(\hat{X}(t),\hat{V}(t),t)f_\alpha(X(t,\tau),V(t,\tau),t)).\end{aligned}$$

Proof. By direct calculation, we have

$$\begin{aligned}&\partial_t(f_\alpha(\hat{X}(t),\hat{V}(t),t)f_\alpha(X(t,\tau),V(t,\tau),t))\\
&\quad= f_\alpha(\hat{X}(t),\hat{V}(t),t)(\partial_t f_\alpha + \nabla_y f_\alpha \cdot \partial_t X(t,\tau) + \nabla_{v_*} f_\alpha \cdot \partial_t V(t,\tau))\\
&\quad= f_\alpha(\hat{X}(t),\hat{V}(t),t)[\partial_t f_\alpha + \nabla_y f_\alpha \cdot (V(t,\tau) + \partial_t X(t,\tau) - V(t,\tau))\\
&\qquad+\nabla_{v_*} f_\alpha \cdot (E(X(t,\tau),t) + \partial_t V(t,\tau) - E(X(t,\tau),t))]\\
&\quad= f_\alpha(\hat{X}(t),\hat{V}(t),t)[\nabla_y f_\alpha \cdot \partial_\tau X(t,\tau) + \nabla_{v_*} f_\alpha \cdot \partial_\tau V(t,\tau)]\\
&\quad= f_\alpha(\hat{X}(t),\hat{V}(t),t)\partial_\tau(f_\alpha(X(t,\tau),V(t,\tau),t))\\
&\quad= \partial_\tau(f_\alpha(\hat{X}(t),\hat{V}(t),t)f_\alpha(X(t,\tau),V(t,\tau),t)),\end{aligned}$$

where $\nabla_y f_\alpha$ and $\nabla_{v_*} f_\alpha$ are evaluated at $(X(t,\tau),V(t,\tau),t)$, and we used

$$\frac{d}{dt} f_\alpha(\hat{X}(t),\hat{V}(t),t) = 0. \qquad \square$$

Finally, the above lemma yields the time-decay of $\mathcal{D}(f_\alpha(t))$ as follows.

Proof of Theorem 1.1. Let $t \geq 0$ be given. Then by definition of $\mathcal{D}(f_\alpha)$

$$\begin{aligned}\frac{d}{dt}\mathcal{D}(f_\alpha(t)) &= \iint_{\mathbb{R}^N\times\mathbb{R}^N}\int_0^\infty\int_{\mathbb{R}^N} \partial_t(f_\alpha(\hat{X}(t),\hat{V}(t),t)f_\alpha(X(t,\tau),V(t,\tau),t))\,dv_*\,d\tau\,dv\,dx\\
&= \iint_{\mathbb{R}^N\times\mathbb{R}^N}\int_0^\infty\int_{\mathbb{R}^N} \partial_\tau(f_\alpha(\hat{X}(t),\hat{V}(t),t)f_\alpha(X(t,\tau),V(t,\tau),t))\,dv_*\,d\tau\,dv\,dx\\
&= \iint_{\mathbb{R}^N\times\mathbb{R}^N}\int_0^\infty \partial_\tau\left(\int_{\mathbb{R}^N} f_\alpha(\hat{X}(t),\hat{V}(t),t)f_\alpha(X(t,\tau),V(t,\tau),t)\,dv_*\right)d\tau\,dv\,dx\\
&= \iint_{\mathbb{R}^N\times\mathbb{R}^N} f_\alpha(\hat{X}(t),\hat{V}(t),t)\left(\lim_{\tau\to\infty}\int_{\mathbb{R}^N} f_\alpha(X(t,\tau),V(t,\tau),t)\,dv_*\right)dv\,dx\\
&\quad- \iiint_{\mathbb{R}^N\times\mathbb{R}^N\times\mathbb{R}^N} f_\alpha(\hat{X}(t),\hat{V}(t),t)f_\alpha(\hat{X}(t),v_*,t)\,dv_*\,dv\,dx\\
&= J_1 + J_2.\end{aligned}$$

Case 1: (J_2). We use the Liouville property $dxdv = d\hat{X}(t)\,d\hat{V}(t)$ to get

$$J_2 = -\iiint_{\mathbb{R}^N\times\mathbb{R}^N\times\mathbb{R}^N} f_\alpha(x,v,t)f_\alpha(x,v_*,t)\,dv_*\,dv\,dx.$$

Case 2: (J_1). We split the integral J_1 into two parts:

$$J_1 = \iint_{\mathbb{R}^N\times\mathbb{R}^N} f_\alpha(\hat{X}(t),\hat{V}(t),t)\left(\lim_{\tau\to\infty}\int_{\mathbb{R}^N} f_\alpha(X(t,\tau),V(t,\tau),t)\,dv_*\right)dv\,dx$$
$$= \iint_{\mathbb{R}^N\times\mathbb{R}^N} f_\alpha(\hat{X}(t),\hat{V}(t),t)\lim_{\tau\to\infty}\left(\int_{|v_*-\hat{V}(t+\tau)|\le\frac{1}{\sqrt{\tau}}} f_\alpha(X(t,\tau),V(t,\tau),t)\,dv_* + \int_{|v_*-\hat{V}(t+\tau)|>\frac{1}{\sqrt{\tau}}} f_\alpha(X(t,\tau),V(t,\tau),t)\,dv_*\right)dv\,dx.$$

We claim

$$\lim_{\tau\to\infty}\left(\int_{|v_*-\hat{V}(t+\tau)|\le\frac{1}{\sqrt{\tau}}} f_\alpha(X(t,\tau),V(t,\tau),t)\,dv_* + \int_{|v_*-\hat{V}(t+\tau)|>\frac{1}{\sqrt{\tau}}} f_\alpha(X(t,\tau),V(t,\tau),t)\,dv_*\right) = 0.$$

Proof of claim. We repeat the argument used in the proof of Lemma 3.2. Let τ_0 be a positive number to be chosen later and consider $\tau\ge\max\{\tau_0, C_1^2|x|^2\}$. For v_* such that $|v_*-\hat{V}(t+\tau)|\le\frac{1}{\sqrt{\tau}}$, we use the fact

$$f_\alpha(X(t,\tau),V(t,\tau),t) = f_{\alpha,0}(\tilde{x},\tilde{v}) \le \frac{\varepsilon_0}{(1+|\tilde{x}|^2)^{\frac{\mu_1}{2}}(1+|\tilde{v}|^2)^{\frac{\mu_2}{2}}}$$

to get

$$\int_{|v_*-\hat{V}(t+\tau)|\le\frac{1}{\sqrt{\tau}}} f_\alpha(X(t,\tau),V(t,\tau),t)\,dv_* = \int_{|v_*-\hat{V}(t+\tau)|\le\frac{1}{\sqrt{\tau}}} f_{\alpha,0}(\tilde{x},\tilde{v})\,dv_* \le \mathcal{O}(\varepsilon_0)\tau^{-\frac{N}{2}}. \tag{3.11}$$

On the other hand, for v_* such that $|v_*-\hat{V}(t+\tau)|\ge\frac{1}{\sqrt{\tau}}$, we use Lemma 2.4 to find

$$\frac{\sqrt{\tau}}{C_1}\le\frac{t+\tau}{C_1}|v_*-\hat{V}(t+\tau)|\le|x-\tilde{x}|,$$

from which we have

$$1+|\tilde{x}|^2\ge 1+(|x-\tilde{x}|-|x|)^2\ge 1+\left(\frac{\sqrt{\tau}}{C_1}-|x|\right)^2 \quad\text{since}\quad |x|\le\frac{\sqrt{\tau}}{C_1}.$$

Hence if

$$\sqrt{\tau}\ge\frac{2C_1|x|+C_1^2+\sqrt{4C_1^3|x|+C_1^4}}{2} := \sqrt{\tau_0(x)},$$

we have

$$\left(\frac{\sqrt{\tau}}{C_1}-|x|\right)^2\ge\sqrt{\tau} \quad\text{or}\quad \frac{1}{(1+(\sqrt{\tau}/C_1-|x|)^2)^{\frac{\mu_1}{2}}}\le\frac{1}{(1+\sqrt{\tau})^{\frac{\mu_1}{2}}}. \tag{3.12}$$

Then the second integral in the claim can be estimated as follows: for $\tau \geq \tau_0(x)$,

$$
\begin{aligned}
(3.13)\quad &\int_{|v_*-\hat{V}(t+\tau)|>\frac{1}{\sqrt{\tau}}} f_\alpha(X(t,\tau),V(t,\tau),t)\,dv_* \\
&= \int_{|v_*-\hat{V}(t+\tau)|>\frac{1}{\sqrt{\tau}}} f_{\alpha,0}(\tilde{x},\tilde{v})\,dv_* \\
&\leq \varepsilon_0 \int_{|v_*-\hat{V}(t+\tau)|>\frac{1}{\sqrt{\tau}}} \frac{1}{(1+|\tilde{x}|^2)^{\frac{\mu_1}{2}}(1+|\tilde{v}|^2)^{\frac{\mu_2}{2}}}\,dv_* \\
&\leq \mathcal{O}(\varepsilon_0) \int_{|v_*-\hat{V}(t+\tau)|>\frac{1}{\sqrt{\tau}}} \frac{1}{(1+(\sqrt{\tau}/C_1-|x|)^2)^{\frac{\mu_1}{2}}(1+|v_*|^2+\mathcal{O}(\varepsilon_0))^{\frac{\mu_2}{2}}}\,dv_* \\
&\leq \frac{\mathcal{O}(\varepsilon_0)}{(1+\sqrt{\tau})^{\frac{\mu_1}{2}}} \int_{\mathbb{R}^N} \frac{1}{(1+|v_*|^2+\mathcal{O}(\varepsilon_0))^{\frac{\mu_2}{2}}}\,dv_* \quad \text{by (3.12)} \\
&= \frac{\mathcal{O}(\varepsilon_0)}{(1+\sqrt{\tau})^{\frac{\mu_1}{2}}}.
\end{aligned}
$$

We combine (3.11) and (3.13) to get

$$
\begin{aligned}
\lim_{\tau\to\infty} \Bigg(&\int_{|v_*-\hat{V}(t+\tau)|\leq\frac{1}{\sqrt{\tau}}} f_\alpha(X(t,\tau),V(t,\tau),t)\,dv_* \\
&+ \int_{|v_*-\hat{V}(t+\tau)|>\frac{1}{\sqrt{\tau}}} f_\alpha(X(t,\tau),V(t,\tau),t)\,dv_* \Bigg) = \lim_{\tau\to\infty}\left(\tau^{-\frac{N}{2}}+\tau^{-\frac{\mu_1}{4}}\right) = 0,
\end{aligned}
$$

which yields

$$J_1 = 0.$$

Finally we combine Cases 1 and 2 to conclude

$$\frac{d}{dt}\mathcal{D}(f_\alpha(t)) = -\Lambda(f_\alpha(t)).$$

REMARK 3.3. *The functional $\mathcal{D}(f_\alpha(t))$ is also a Lyapunov functional for the relativistic Vlasov–Maxwell system for sufficiently small and smooth decaying solutions given in* [7] *(see the appendix).*

4. Uniform ^{1}-stability. In this section, we study uniform L^1-stability for smooth solutions to (1.1). Unlike the case of the Boltzmann equation in [6, 19, 20], the nonlinear functional approach based on the interaction potential $\mathcal{D}(f_\alpha)$ seems very difficult to apply to the Vlasov–Poisson system. For large physical dimensions $N \geq 4$, instead, the direct L^1-stability estimate based on the Gronwall-type inequality can be used; we first need to estimate an integral

$$K_\alpha(x,t) := \int_{\mathbb{R}^N} \frac{1}{|x-y|^{N-1}} \left(\int_{\mathbb{R}^N} |\nabla_v f_\alpha|(y,v,t)\,dv \right) dy.$$

LEMMA 4.1. *Suppose the main hypothesis* (H) *in section 1 holds, and $N \geq 4$. Let f and $\bar{f}$ be two smooth solutions corresponding to the initial data f_0 and $\bar{f}_0$, respectively. Then we have*

$$\sum_{\alpha=1}^{M} \|K_\alpha(\cdot,t)\|_{L^\infty(\mathbb{R}^N_x)} \leq CM(1+t)^{-(N-2)}.$$

Proof. Let $\alpha \in \{1, \dots, M\}$ be given and recall Lemma 2.3:

$$\left\| \int_{\mathbb{R}^N} |\nabla_v f_\alpha(x,v,t)|\, dv \right\|_{L^\infty(\mathbb{R}^N_x)} \le \mathcal{O}(1)(1+t)^{-(N-1)}, \tag{4.1}$$

$$\left\| \int_{\mathbb{R}^N} |\nabla_v f_\alpha(x,v,t)|\, dv \right\|_{L^1(\mathbb{R}^N_x)} \le \mathcal{O}(1)(1+t). \tag{4.2}$$

Let r be a positive constant to be determined later. Then we split $K_\alpha(x,t)$ into two parts:

$$\begin{aligned} K_\alpha(x,t) &= \int_{|y-x|\le r} \frac{1}{|x-y|^{N-1}} \left(\int_{\mathbb{R}^N} |\nabla_v f_\alpha(y,v,t)|\, dv \right) dy \\ &\quad + \int_{|y-x|>r} \frac{1}{|x-y|^{N-1}} \left(\int_{\mathbb{R}^N} |\nabla_v f_\alpha(y,v,t)|\, dv \right) dy \\ &\le d(N) r \left\| \int_{\mathbb{R}^N} |\nabla_v f_\alpha(x,v,t)|\, dv \right\|_{L^\infty(\mathbb{R}^N_x)} \\ &\quad + \left\| \int_{\mathbb{R}^N} |\nabla_v f_\alpha(x,v,t)|\, dv \right\|_{L^1(\mathbb{R}^N_x)} \frac{1}{r^{N-1}}, \end{aligned} \tag{4.3}$$

where $d(N)$ is a positive constant depending only on N. In order to minimize the right-hand side of (4.3), we choose r such that

$$\left\| \int_{\mathbb{R}^N} |\nabla_v f_\alpha|\, dv \right\|_{L^\infty(\mathbb{R}^N_x)} d(N) r = \left\| \int_{\mathbb{R}^N} |\nabla_v f_\alpha|\, dv \right\|_{L^1(\mathbb{R}^N_x)} \frac{1}{r^{N-1}},$$

i.e.,

$$r = \left(\frac{\left\| \int_{\mathbb{R}^N} |\nabla_v f_\alpha|\, dv \right\|_{L^1(\mathbb{R}^N_x)}}{\left\| \int_{\mathbb{R}^N} |\nabla_v f_\alpha|\, dv \right\|_{L^\infty(\mathbb{R}^N_x)} d(N)} \right)^{\frac{1}{N}}.$$

Hence for such r, we have

$$\begin{aligned} K_\alpha(x,t) &\le 2 d(N)^{\frac{N-1}{N}} \left\| \int_{\mathbb{R}^N} |\nabla_v f_\alpha|\, dv \right\|^{1-\frac{1}{N}}_{L^\infty(\mathbb{R}^N_x)} \left\| \int_{\mathbb{R}^N} |\nabla_v f_\alpha|\, dv \right\|^{\frac{1}{N}}_{L^1(\mathbb{R}^N_x)} \\ &\le \mathcal{O}(1)(1+t)^{-(N-2)}. \end{aligned}$$

We take a supremum over x and add all α to get the desired result. ∎

REMARK 4.1. *Note that* $\sum_{\alpha=1}^M \|K_\alpha(\cdot,t)\|_{L^\infty(\mathbb{R}^N_x)}$ *is integrable in* t *for* $N \ge 4$.

Based on the above estimate, we obtain the uniform L^1-stability estimate.

Proof of Theorem 1.2. Let f_α and $\bar f_\alpha$ be smooth solutions of (1.1) corresponding to the initial data $f_{\alpha,0}$ and $\bar f_{\alpha,0}$, respectively:

$$\partial_t f_\alpha + v \cdot \nabla_x f_\alpha + E(f) \cdot \nabla_v f_\alpha = 0, \tag{4.4}$$

$$\partial_t \bar f_\alpha + v \cdot \nabla_x \bar f_\alpha + E(\bar f) \cdot \nabla_v \bar f_\alpha = 0. \tag{4.5}$$

We subtract (4.5) from (4.4) to get

$$\partial_t (f_\alpha - \bar f_\alpha) + v \cdot \nabla_x (f_\alpha - \bar f_\alpha) + E(f) \cdot \nabla_v (f_\alpha - \bar f_\alpha) = (E(\bar f) - E(f)) \cdot \nabla_v \bar f_\alpha. \tag{4.6}$$

Let $(x, v, t) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}_+$ be fixed and let $(X(s), V(s))$ be the trajectories of particles for f_α passing through the point (x, v) at time t, i.e.,

$$X(t) = x \quad \text{and} \quad V(t) = v.$$

We integrate (4.6) along the trajectory $(X(s), V(s))$ to get

$$\begin{aligned}(f_\alpha - \bar{f}_\alpha)(x, v, t) &= (f_{\alpha,0} - \bar{f}_{\alpha,0})(X(0), V(0)) \\ &\quad + \int_0^t (E(\bar{f}) - E(f))(X(s), s) \cdot \nabla_v \bar{f}_\alpha(X(s), V(s), s)\, ds.\end{aligned} \tag{4.7}$$

If we take an absolute for (4.7) and integrate it over the whole phase space, then we obtain

(4.8)

$$\begin{aligned}&\|f_\alpha(t) - \bar{f}_\alpha(t)\|_{L^1} \le \|f_{\alpha,0} - \bar{f}_{\alpha,0}\|_{L^1} \\ &\quad + \int_0^t \iint_{\mathbb{R}^N \times \mathbb{R}^N} |(E(\bar{f}) - E(f))(X(s), s)||\nabla_v \bar{f}_\alpha(X(s), V(s), s)|\, dv\, dx\, ds \\ &\quad \le \|f_{\alpha,0} - \bar{f}_{\alpha,0}\|_{L^1} + C\int_0^t \int_{\mathbb{R}^N} |\rho(y, s) - \bar{\rho}(y, s)| \\ &\qquad \times \Big[\int_{\mathbb{R}^N} \frac{dX(s)}{|X(s) - y|^{N-1}} \Big(\int_{\mathbb{R}^N} |\nabla_v \bar{f}_\alpha(X(s), V(s), t) dV(s) \Big) \Big] dy\, ds \le \|f_{\alpha,0} - \bar{f}_{\alpha,0}\|_{L^1} \\ &\qquad + C\int_0^t \iint_{\mathbb{R}^N \times \mathbb{R}^N} K_\alpha(y, s) \Big(\sum_{\alpha=1}^M |f_\alpha(y, v_*, s) - \bar{f}_\alpha(y, v_*, s)| \Big) dv_*\, dy\, ds \\ &\quad = \|f_{\alpha,0} - \bar{f}_{\alpha,0}\|_{L^1} + C\int_0^t \|K_\alpha(\cdot, s)\|_{L^\infty} \sum_{\alpha=1}^M \|f_\alpha(s) - \bar{f}_\alpha(s)\|_{L^1} ds,\end{aligned}$$

where Lemma 4.1 and the Liouville principle $dx\, dv = dX(s)dV(s)$ were employed to get

$$\begin{aligned}&\iint_{\mathbb{R}^N \times \mathbb{R}^N} \frac{1}{|X(s) - y|^{N-1}} |\nabla_v \bar{f}_\alpha(X(s), V(s), t)|\, dX(s)\, dV(s) \\ &\quad = \iint_{\mathbb{R}^N \times \mathbb{R}^N} \frac{1}{|x - y|^{N-1}} |\nabla_v \bar{f}_\alpha(x, v, t)|\, dx\, dv.\end{aligned}$$

We now add (4.8) over all α to get

$$\begin{aligned}\|f(t) - \bar{f}(t)\|_{L^1} &\le \|f_0 - \bar{f}_0\|_{L^1} + C\int_0^t \Big(\sum_{\alpha=1}^M \|K_\alpha(\cdot, s)\|_{L^\infty} \Big) \|f(s) - \bar{f}(s)\|_{L^1} ds \\ &\le \|f_0 - \bar{f}_0\|_{L^1} + CM\int_0^t (1+s)^{-(N-2)} \|f(s) - \bar{f}(s)\|_{L^1} ds.\end{aligned}$$

Hence Gronwall's inequality yields

$$\begin{aligned}\|f(t) - \bar{f}(t)\|_{L^1} &\le \|f_0 - \bar{f}_0\|_{L^1} \exp\Big(CM\int_0^t (1+s)^{-(N-2)} ds \Big) \\ &= G\|f_0 - \bar{f}_0\|_{L^1},\end{aligned} \tag{4.9}$$

where

$$G := \exp\left(CM\int_0^\infty (1+s)^{-(N-2)}ds\right) < \infty \quad \text{for} \quad N \geq 4.$$

Appendix. Relativistic Vlasov–Maxwell system. In this appendix we briefly discuss the interaction potential $\mathcal{D}(f)$ for the relativistic Vlasov–Maxwell system. For details on the global existence, we refer the reader to [5, 8, 9, 10, 11, 13, 14, 15, 16, 26]. Since the number of species is irrelevant in the following analysis, without loss of generality, we assume that a plasma consists of single species, i.e., $M = 1$; moreover, we also assume the phase space is $\mathbb{R}^3 \times \mathbb{R}^3$.

Consider a plasma consisting of ions confined to the whole phase space $\mathbb{R}^3 \times \mathbb{R}^3$ under the self-consistent electromagnetic force. We denote a one-point particle distribution function by f, and the electric field and magnetic field are denoted by E and B, respectively. The total charge ρ and the current j are defined as

$$\rho = 4\pi\int_{\mathbb{R}^3} f dv \quad \text{and} \quad j = 4\pi\int_{\mathbb{R}^3} \hat{v} f dv,$$

where the relativistic velocity is

$$\hat{v} = \frac{v}{\sqrt{1+|v|^2}}$$

if the speed of light c and the particle mass are taken to be 1. In this case, the relativistic Vlasov–Maxwell system reads

$$\text{(A.1)} \qquad \begin{cases} \partial_t f + \hat{v}\cdot\nabla_x f + (E + \hat{v}\times B)\cdot\nabla_v f = 0 \quad \text{in } \mathbb{R}^3\times\mathbb{R}^3\times\mathbb{R}_+, \\ \partial_t E = \nabla\times B - j, \quad \nabla\cdot E = \rho, \\ \partial_t B = -\nabla\times E, \quad \nabla\cdot B = 0, \end{cases}$$

subject to the initial data

$$f(x,v,0) = f_0(x,v) \quad \text{and} \quad (E,B)(x,0) = (E_0,B_0)(x) \quad (x,v)\in\mathbb{R}^3\times\mathbb{R}^3$$

satisfying constraint conditions

$$\nabla\cdot E_0 = \rho_0 \quad \text{and} \quad \nabla\cdot B_0 = 0.$$

The smooth global solution is known to exist under the small data assumption.

THEOREM A.1 (see [7, 16]). *Let f_0 be a nonnegative C^1 function with supports in $|x| \leq k, |v| \leq k$ and let E_0, B_0 be C^2 functions with supports in $|x| \leq k$. If the initial data satisfy the smallness assumption*

$$\|f_0\|_{C^1} + \|E\|_{C^2} + \|B\|_{C^2} \leq \varepsilon_0 \ll 1,$$

then there exists a unique C^1-global solution: for $(x,t)\in\mathbb{R}^3\times\mathbb{R}_+$,

$$\text{(A.2)} \quad f(t,x,v) = 0 \text{ for } |v| \geq \beta \quad \text{and} \quad |E(t,x)| + |B(t,x)| \leq \frac{\varepsilon_0}{(1+t)(t-|x|+2k)}.$$

REMARK A.1. *The compact support assumption $|v| \leq k$ on the initial data can be relaxed to include the Gaussian-type initial data* [32].

In what follows, we consider smooth solutions. For a given point $(x,v) \in \mathbb{R}^3 \times \mathbb{R}^3$, the particle path $[\bar{X}(s), \bar{V}(s)]$ is defined as the solution of the following ODE system:

$$\frac{d\bar{X}(s)}{ds} = \hat{\bar{V}}(s) \quad \text{and} \quad \frac{d\bar{V}(s)}{ds} = E(\bar{X}(s), s) + \hat{\bar{V}} \times B, \tag{A.3}$$

with the initial data

$$\bar{X}(0) = x \quad \text{and} \quad \bar{V}(0) = v,$$

where

$$\hat{\bar{V}} = \frac{\bar{V}}{\sqrt{1 + |\bar{V}|^2}}.$$

Since $\nabla_{(X,V)} \cdot (\hat{V}, E + \hat{V} \times B) = 0$, the mapping $(x,v) \longrightarrow (\bar{X}(s), \bar{V}(s))$ is measure preserving. As in section 3.2 we define $[X(t,\tau), V(t,\tau)]$ by

$$\begin{aligned} X(t,\tau) &:= X(t; t+\tau, \bar{X}(t+\tau), v_*) \quad \text{and} \\ V(t,\tau) &:= V(t; t+\tau, \bar{X}(t+\tau), v_*), \end{aligned}$$

from which we have

$$X(t,\tau) := \bar{X}(t+\tau) - \int_t^{t+\tau} \hat{V}(\theta; t+\tau, \bar{X}(t+\tau), v_*)\, d\theta, \tag{A.4}$$

$$\begin{aligned} V(t,\tau) := v_* - \int_t^{t+\tau} & E(X(\theta; t+\tau, \bar{X}(t+\tau), v_*), \theta) \\ &+ \hat{V}(\theta; t+\tau, \bar{X}(t+\tau), v_*) \times B(X(\theta; t+\tau, \bar{X}(t+\tau), v_*), \theta)\, d\theta. \end{aligned} \tag{A.5}$$

The following lemma can easily be computed.

LEMMA A.2. *Let $[X(t,\tau), V(t,\tau)]$ be the curve parameterized by τ and given by* (A.4) *and* (A.5). *Then the following estimates hold:*

(1) $\partial_t X(t,\tau) - \partial_\tau X(t,\tau) = \hat{V}(t,\tau)$,

(2) $\partial_t V(t,\tau) - \partial_\tau V(t,\tau) = E(X(t,\tau), t) + \hat{V}(t,\tau) \times B(X(t,\tau), t)$.

We also define $\mathcal{D}(f(t))$ and $\Lambda(f(t))$ by

$$\mathcal{D}(f(t)) := \iint_{\mathbb{R}^3 \times \mathbb{R}^3} f(\bar{X}(t), \bar{V}(t), t) \left(\int_0^\infty \int_{\mathbb{R}^3} f(X(t,\tau), V(t,\tau), t)\, d\tau\, dv_* \right) dv\, dx,$$

$$\Lambda(f(t)) := \iiint_{\mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^3} f(x,v,t) f(x, v_*, t)\, dv_*\, dv\, dx.$$

Below we state two lemmas whose proofs can be found in [7].

LEMMA A.3. *If $0 \le s \le t$ and $f(t,x,v) \neq 0$, then*

$$s - |X(s; t, x, v)| + 2k \ge (k+s)(2 + 2P^2(t))^{-1},$$

where the function $P(t)$ is bounded by a constant β.

LEMMA A.4. *If ε_0 is sufficiently small, there is a constant $c > 0$ such that*

$$|X(0;t,x,v) - X(0;t,x,w)| \geq ct|v-w|.$$

For us to follow the steps in section 3.2, it is enough to prove Lemma 2.5 for the relativistic Vlasov–Maxwell system or to verify directly (3.9):

$$\int_{\mathbb{R}^3} \frac{1}{(1+|V(0,\tau)|^2)^{\frac{\mu_2}{2}}}\,dv_* = \mathcal{O}(1).$$

Below for the simplicity of notation, we denote $X(s;\tau,\bar{X}(\tau),v_*)$, $V(s;\tau,\bar{X}(\tau),v_*)$ by $X(s,\tau), V(s,\tau)$, respectively. Note that $X(s,\tau), V(s,\tau)$ here are not the same as $X(s,\tau), V(s,\tau)$ defined in (3.5). It follows from (A.2) and Lemma A.3 that

$$\begin{aligned}
&\left|\int_0^\tau E(X(s,\tau),\theta) + \hat{V}(s,\tau)\times B(X(s,\tau),\theta)\,d\theta\right| \\
&\quad\leq \varepsilon_0 \int_0^\tau \frac{1}{(s+1)(s-|X(s,\tau)|+2k)}\,ds \\
&\quad\leq \varepsilon_0 \int_0^\tau \frac{2+2P^2(s)}{(s+1)(k+s)}\,ds \\
&\quad\leq \bar{C}_1\varepsilon_0 \qquad \text{for some positive constant} \quad \bar{C}_1
\end{aligned} \tag{A.6}$$

and that

$$\begin{aligned}
&\int_{\mathbb{R}^3} \frac{1}{(1+|V(0,\tau)|^2)^{\frac{\mu_2}{2}}}\,dv_* \\
&\quad= \int_{\mathbb{R}^3} \frac{1}{\left(1+\left|v_* - \int_0^\tau E(X(s,\tau),s)+\hat{V}(s,\tau)\times B(X(s,\tau),s)ds\right|^2\right)^{\frac{\mu_2}{2}}}\,dv_* \\
&\quad\leq \int_{|v_*|\leq 2\bar{C}_1\varepsilon_0} \frac{1}{\left(1+\left||v_*| - |\int_0^\tau E(X(s,\tau),s)-\hat{V}(s,\tau)\times B(X(s,\tau),s)ds|\right|^2\right)^{\frac{\mu_2}{2}}}\,dv_* \\
&\qquad+ \int_{|v_*|\geq 2\bar{C}_1\varepsilon_0} \frac{1}{\left(1+\left||v_*| - |\int_0^\tau E(X(s,\tau),s)-\hat{V}(s,\tau)\times B(X(s,\tau),s)ds|\right|^2\right)^{\frac{\mu_2}{2}}}\,dv_* \\
&\quad\leq \mathcal{O}(\varepsilon_0) + \int_{|v_*|\geq 2\bar{C}_1\varepsilon_0} \frac{1}{(1+\frac{1}{2}|v_*|^2)^{\frac{\mu_2}{2}}}\,dv_* \\
&\quad\leq \mathcal{O}(1).
\end{aligned}$$

We obtain the time decay estimates of the interaction potential.

PROPOSITION A.5. *Let f be a smooth solution to* (A.1). *Then if ε_0 is sufficiently small, $\mathcal{D}(f)$ satisfies*

$$\frac{d\mathcal{D}(f(t))}{dt} + \Lambda(f(t)) = 0.$$

Acknowledgment. The second author thanks Prof. Tai-Ping Liu for his hospitality and support.

REFERENCES

[1] A. A. Arsenev, *Global existence of a weak solution of Vlasov's system of equations*, U.S.S.R. Comput. Math. Math. Phys., 15 (1975), pp. 131–143.

[2] C. Bardos and P. Degond, *Global existence for the Vlasov-Poisson equation in three space variables with small initial data*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 2 (1985), pp. 101–118.

[3] J. Batt, *Global symmetric solutions of the initial value problem of stellar dynamic*, J. Differential Equations, 25 (1977), pp. 342–364.

[4] J. Batt and R. Rein, *A rigorous stability result for the Vlasov–Poisson system in three dimensions*, Anal. Mat. Pura Appl., 164 (1993), pp. 133–154.

[5] F. Bouchut, F. Golse, and C. Pallard, *Classical solutions and the Glassey-Strauss theorem for the 3D Vlasov-Maxwell system*, Arch. Ration. Mech. Anal., 170 (2003), pp. 1–15.

[6] M. Chae and S. Y. Ha, *Stability estimates of the Boltzmann equation with quantum effects*, submitted.

[7] R. Glassey, *The Cauchy Problem in Kinetic Theory*, SIAM, Philadelphia, PA, 1996.

[8] R. Glassey and J. Schaeffer, *The relativistic Vlasov-Maxwell system in two space dimensions: Part* I, Arch. Ration. Mech. Anal., 141 (1998), pp. 331–354.

[9] R. Glassey and J. Schaeffer, *The relativistic Vlasov-Maxwell system in two space dimensions: Part* II, Arch. Ration. Mech. Anal., 141 (1998), pp. 355–374.

[10] R. Glassey and J. Schaeffer, *The two and one-half dimensional relativistic Vlasov-Maxwell system*, Comm. Math. Phys., 185 (1997), pp. 257–284.

[11] R. Glassey and J. Schaeffer, *Global existence of the relativistic Vlasov-Maxwell system with nearly neutral initial data*, Comm. Math. Phys., 119 (1988), pp. 353–384.

[12] R. Glassey and J. Schaeffer, *On symmetric solutions to the relativistic Vlasov–Poisson system*, Comm. Math. Phys., 101 (1985), pp. 459–473.

[13] R. Glassey and W. Strauss, *Large velocities in the relativistic Vlasov-Maxwell equations*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 36 (1989), pp. 615–627.

[14] R. Glassey and W. Strauss, *High velocity particles in a collisionless plasma*, Math. Mech. Appl. Sci., 9 (1987), pp. 45–52.

[15] R. Glassey and W. Strauss, *Absence of shocks in an initially dilute collisionless plasma*, Comm. Math. Phys., 113 (1987), pp. 191–208.

[16] R. Glassey and W. Strauss, *Singularity formation in a collisionless plasma could occur only at high velocities*, Arch. Ration. Mech. Anal., 92 (1986), pp. 59–90.

[17] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[18] Y. Guo, *The Vlasov–Poisson–Boltzmann system near vacuum*, Comm. Math. Phys., 218 (2001), pp. 293–313.

[19] S. Y. Ha, *L^1-stability of the Boltzmann equation for the hard sphere model*, Arch. Ration. Mech. Anal., 173 (2004), pp. 279–296.

[20] S. Y. Ha, *Stability estimates of the Boltzmann equation and uniform stability estimates*, J. Differential Equations, 215 (2005), pp. 178–205.

[21] E. Horst, *On the classical solutions of the initial value problem for the unmodified non-linear Vlasov equation (Part* I*)*, Math. Methods Appl. Sci., 3 (1981), pp. 229–248.

[22] E. Horst, *On the classical solutions of the initial value problem for the unmodified non-linear Vlasov equation (Part* II*)*, Math. Methods Appl. Sci., 4 (1982), pp. 19–32.

[23] E. Horst, *On the asymptotic growth of the solutions of the Vlasov–Poisson system*, Math. Methods Appl. Sci., 16 (1993), pp. 75–85.

[24] E. Horst and R. Hunze, *Weak solutions of the initial value problem for the unmodified nonlinear Vlasov equation*, Math. Methods Appl. Sci., 6 (1984), pp. 262–279.

[25] S. V. Iordanskii, *The Cauchy problem for the kinetic equation of plasma*, Amer. Math. Soc. Transl. Ser., 35 (1964), pp. 351–363.

[26] S. Klainerman and G. Staffilani, *A new approach to study the Vlasov-Maxwell system*, Comm. Pure Appl. Anal., 1 (2002), pp. 103–125.

[27] P. L. Lions and B. Perthame, *Propagation of moments and regularity of solutions for the 3-dimensional Vlasov–Poisson system*, Invent. Math., 105 (1991), pp. 415–430.

[28] B. Perthame, *Mathematical tools for kinetic equations*, Bull. Amer. Math. Soc. (N.S.), 41 (2004), pp. 205–244.

[29] B. Perthame, *Time decay, propagation of low moments and dispersive effects for kinetic equations*, Comm. Partial Differential Equations, 21 (1996), pp. 659–686.

[30] K. Pfaffelmoser, *Global classical solutions of the Vlasov–Poisson system in three dimensions for general initial data*, J. Differential Equations, 95 (1992), pp. 281–303.

[31] G. Rein, *Nonlinear stability for the Vlasov–Poisson system-the energy Casimir method*, Math. Methods Appl. Sci., 17 (1994), pp. 1129–1140.
[32] J. Schaeffer, *A small data theorem for collisionless plasma that includes high velocity particles*, Indiana Univ. Math. J., 53 (2004), pp. 1–34.
[33] J. Schaeffer, *Global existence of smooth solutions to the Vlasov–Poisson system in three dimensions*, Comm. Partial Differential Equations, 16 (1991), pp. 1313–1335.
[34] J. Schaeffer, *Global existence for the Poisson-Vlasov system with nearly symmetric data*, J. Differential Equations, 69 (1987), pp. 111–148.
[35] J. Schaeffer, *The classical limit of the relativistic Vlasov-Maxwell system*, Comm. Math. Phys., 104 (1986), pp. 403–421.
[36] S. Ukai and T. Okabe, *On classical solutions in the large in time of two-dimensional Vlasov's equation*, Osaka J. Math., 15 (1978), pp. 245–261.
[37] S. Wollman, *A discrete approximation to the one-dimensional Vlasov–Poisson system*, Transport Theory Statist. Phys., 17 (1988), pp. 561–594.
[38] S. Wollman, *Global-in-time solutions to the two-dimensional Vlasov–Poisson system*, Comm. Pure Appl. Math., 33 (1980), pp. 173–197.
[39] S. Wollman, *Convergence of a numerical approximation to the one dimensional Vlasov Poisson system*, Transport Theory Statist. Phys., 19 (1990), pp. 545–562.
[40] S. Wollman, *Global-in-time solutions to the three-dimensional Vlasov–Poisson system*, J. Math. Anal. Appl., 176 (1993), pp. 76–91.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1732–1741

COARSENING RATES IN OFF-CRITICAL MIXTURES*

SERGIO CONTI†, BARBARA NIETHAMMER‡, AND FELIX OTTO§

Abstract. We study coarsening of a binary mixture within the Mullins–Sekerka evolution in the regime where one phase has small volume fraction $\phi \ll 1$. Heuristic arguments suggest that the energy density, which represents the inverse of a typical length scale, decreases as $\phi t^{-1/3}$ as $t \to \infty$. We prove rigorously a corresponding weak lower bound. Moreover, we establish a stronger result for the two-dimensional case, where we find a lower bound of the form $\phi(\ln \phi^{-1})^{1/3} t^{-1/3}$. Our approach follows closely the analysis in [R. V. Kohn and F. Otto, *Comm. Math. Phys.*, 229 (2002), pp. 375–395], which exploits the relation between two suitable length scales. Our main contribution is an isoperimetric inequality in the two-dimensional case.

Key words. Mullins–Sekerka evolution, coarsening rates, isoperimetric inequalities

AMS subject classification. 82C26

DOI. 10.1137/040620059

1. Introduction. The Mullins–Sekerka model describes the late stage coarsening in the phase separation of a binary mixture. In this model the interface between two phases is characterized by the boundary $\partial\Omega$ of a region $\Omega \subset Q \subset \mathbb{R}^n$, where Ω denotes the region covered by one phase. The Mullins–Sekerka model is based on the assumption that the diffusion field, given by the negative gradient $-\nabla u$ of the chemical potential u, is in quasi-stationary equilibrium given the phase distribution and satisfies the Gibbs–Thomson condition of local equilibrium on the interface. That is, u satisfies

$$-\Delta u = 0 \qquad \text{in } Q\backslash\partial\Omega, \tag{1.1}$$

$$u = \kappa \qquad \text{on } \partial\Omega, \tag{1.2}$$

where κ denotes the mean curvature of the interface and is defined to be positive if Ω is a ball. Then the normal velocity v of the interface is given by the jump of the normal component of the diffusion field across the interface, that is,

$$v = [\nabla u \cdot \vec{n}] := \lim_{\substack{x \in Q\backslash\overline{\Omega} \\ x \to \partial\Omega}} \nabla u(x) \cdot \vec{n} - \lim_{\substack{x \in \Omega \\ x \to \partial\Omega}} \nabla u(x) \cdot \vec{n} \qquad \text{on } \partial\Omega, \tag{1.3}$$

where $\vec{n}$ denotes the outer normal to Ω. With periodic or Neumann boundary conditions the evolution (1.1)–(1.3) reduces the total surface area and keeps the volume fraction covered by each phase constant. In the following we consider a periodic setting with period cell $Q = [0, l)^n$, where l is the system size. This is not a significant restriction, since our results will be independent of the size l.

*Received by the editors December 2, 2004; accepted for publication (in revised form) July 28, 2005; published electronically February 15, 2006. This research was supported by the Deutsche Forschungsgemeinschaft within the priority research program SPP 1095 "Analysis, modeling and simulation of multiscale problems."

http://www.siam.org/journals/sima/37-6/62005.html

†Fachbereich Mathematik, Universität Duisburg-Essen, Lotharstrasse 65, 47057 Duisburg, Germany (conti@math.uni-duisburg.de).

‡Institut für Mathematik, Humboldt Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany (niethamm@hu-berlin.de).

§Institut für Angewandte Mathematik, Universität Bonn, Wegelerstrasse 6, 53115 Bonn, Germany (otto@iam.uni-bonn.de).

Equations (1.1)–(1.3) do not in general possess a global smooth solution since the geometry might become singular—for example, when two previously separated regions touch or when pinching occurs. We assume in the following that we are in a situation where a piecewise smooth solution of the Mullins–Sekerka problem exists and for which the evolution of $\partial\Omega$ is continuous for all times. For typical scenarios in the regime of small volume fraction this assumption is satisfied. In this regime the new phase nucleates in the form of many small well-separated regions which quickly become approximately radially symmetric. If such a configuration is taken as initial data for the Mullins–Sekerka evolution, then there exists a global solution in the described sense [1, 2]. The different regions compete for surface area such that the smaller ones shrink and disappear while the larger ones grow.

We are interested in the coarsening rate within the Mullins–Sekerka evolution in the regime of small volume fraction, i.e., $\phi := |\Omega \cap Q|/|Q| \ll 1$. Here and in the following we denote by $|\omega| := \mathcal{L}^n(\omega)$ the n-dimensional Lebesgue measure of a set $\omega \subset \mathbb{R}^n$. The coarsening rate is described by the growth rate of a typical length scale $L = L(t)$. Since the Mullins–Sekerka evolution is scale invariant with respect to a rescaling $x \mapsto \lambda x, t \mapsto \lambda^3 t$, the only possible universal growth law is $L(t) \sim t^{1/3}$. This growth law has been rigorously established in the form of a weak upper bound in [7] for the Cahn–Hilliard equation in the case of equal volume fractions, that is, $\phi \sim \frac{1}{2}$.

How do we expect that this growth law depends on the volume fraction ϕ? To that aim consider as a natural inverse length scale the energy density

$$E = \text{ interfacial area per unit volume } = \frac{\mathcal{H}^{n-1}(\partial\Omega \cap Q)}{|Q|}. \tag{1.4}$$

If now, for example, the configuration consists of N approximately spherical domains of mean radius $\overline{R}(t)$, we expect

$$E(t) \sim \frac{1}{|Q|} N \overline{R}^{n-1}(t) \sim \phi \frac{1}{\overline{R}(t)}. \tag{1.5}$$

A prediction of the scaling of $\overline{R}(t)$ goes back to the classical theory by Lifshitz and Slyozov [9] and Wagner [12] (see also [10] for more details on the two-dimensional case). To that aim consider again a collection of approximately spherical regions with radius R and total volume fraction ϕ, which are well separated by a typical distance of order $d \sim \frac{R}{\phi^{1/n}}$. Due to the clear separation of length scales, the potential u should approximately be close to a slowly varying field $\bar{u}$ away from the particles. Hence, due to (1.1) and (1.2) we expect that close to a particle with center x_0 the potential is well approximated by

$$u \sim \begin{cases} (1 - R\bar{u}) \frac{R^{n-3}}{|x-x_0|^{n-2}}, & n \geq 3, \\ \left(\frac{1}{R} - \bar{u}\right) \frac{\ln \frac{d}{|x-x_0|}}{\ln\left(\frac{d}{R}\right)}, & n = 2. \end{cases}$$

Plugging the preceding formula into (1.3) gives

$$\dot{R}(t) \sim \begin{cases} \frac{n-2}{R^2(t)}(R\bar{u} - 1), & n \geq 3, \\ \frac{1}{\ln\left(\frac{d}{R}\right)} \frac{1}{R^2(t)}(R\bar{u} - 1), & n = 2. \end{cases} \tag{1.6}$$

Thus, the growth rate of a particle is, to leading order, independent of ϕ for $n \geq 3$ and depends only on $\ln\left(\frac{d}{R}\right) \sim \ln \phi^{-1}$ for $n = 2$. In particular we also find $R^2 \dot{R} \sim 1$

and $R^2\dot{R} \sim \frac{1}{\ln(\frac{d}{R})}$, respectively, such that the mean radius should follow the growth law

$$\overline{R}(t) \sim \begin{cases} t^{1/3}, & n \geq 3, \\ \frac{1}{(\ln \phi^{-1})^{1/3}} t^{1/3}, & n = 2. \end{cases} \tag{1.7}$$

This implies, together with (1.5), that

$$E(t) \sim \begin{cases} \phi t^{-1/3}, & n \geq 3, \\ \phi(\ln \phi^{-1})^{1/3} t^{-1/3}, & n = 2. \end{cases} \tag{1.8}$$

Our goal in this article is to support these heuristics by rigorously establishing a corresponding weak lower bound in the spirit of [7]. This approach has also been successfully applied to other cases, such as phase separation in multicomponent systems or epitaxial growth [8], mean-field models for coarsening [3], temperature-dependent phase field models [4], and coarsening of droplet configurations [11]. The approach is based on exploiting the relation of two suitably chosen global length scales. The first has already been introduced and is given by the inverse of the energy density E. In the following we use the notation from geometric measure theory and denote the perimeter of Ω with respect to Q by $\int_Q |\nabla\chi|$, where χ denotes here and in the following the characteristic function of the set Ω. It is well known (cf., e.g., [6]) that if $\partial\Omega$ is smooth, which we assume here for all but finitely many times, then the perimeter equals the surface area. Thus, the energy can also be written as $E = \fint_Q |\nabla\chi|$, where here and in the following $\fint_Q := \frac{1}{|Q|}\int_Q$.

The second length scale is, as in [7], a suitable negative norm of the characteristic function of Ω. We define

$$L := \left(\fint_Q |\nabla^{-1}(\chi - \phi)|^2 \, dx \right)^{1/2}, \tag{1.9}$$

where $\|\nabla^{-1}u\|_{L^2} = \|u\|_{H^{-1}}$ denotes the H^{-1}-norm for Q-periodic functions with mean value zero.

This choice of length scale is motivated by the interpretation of the Mullins–Sekerka evolution as a gradient flow on a Riemannian manifold. In fact, it is the gradient flow of the surface energy with respect to the scalar product given by the H^{-1}-norm in the bulk. In principle a natural choice for L would be the induced distance on the manifold between the phase configuration and the uniform reference state ϕ. However, we cannot easily compute this distance in the present setting due to the nonconvex constraint that χ takes only two integer values. Nevertheless one would expect that this distance is close to the distance within the Cahn–Hilliard model which is just given by (1.9). In fact, the main result formulated in the following theorem can with similar arguments also be established within the Cahn–Hilliard framework.

THEOREM 1.1. *Let E_0 and L_0 be the initial energy density and initial length, respectively, and let $\phi \ll 1$. Then we have for $T \gg \phi^{-3/2}L_0^3$ that*

$$\fint_0^T E^2(t)\,dt \gtrsim \begin{cases} \phi^2\, T^{-2/3}, & n \geq 3, \\ \phi^2\left(\ln\phi^{-1}\right)^{2/3} T^{-2/3}, & n = 2. \end{cases}$$

Here and in the following we use the notation $A \lesssim B$ or $A \gtrsim B$ if there is a constant C such that $A \leq CB$ or $A \geq CB$, respectively. The constants will be independent of ϕ but may depend on the space dimension n. Such a dependence we sometimes indicate with $C = C(n)$.

The proof of Theorem 1.1 is based exactly as in [7] on three ingredients. The first is an interpolation (or isoperimetric) inequality which relates E and L and is independent of the dynamics of the evolution.

LEMMA 1.2. *We have*

$$EL \gtrsim \begin{cases} \phi^{3/2}, & n \geq 3, \\ \phi^{3/2}\big(\ln \phi^{-1}\big)^{1/2}, & n = 2. \end{cases}$$

The proof of Lemma 1.2—in particular for the case $n = 2$—is the main contribution of this article. We provide it in the next section.

The second ingredient in the analysis of [7] is a diffusion inequality which relates the rate of change of the length L to the rate of change of the energy density.

LEMMA 1.3. *For almost all $t \in (0, T)$ we have*

$$|\dot{L}(t)|^2 \leq -\dot{E}(t).$$

Proof. We denote by u the potential of the normal velocity v of $\partial\Omega$ and by w the potential for $\chi - \phi$, that is,

$$-\Delta u = v \, d\mathcal{H}^{n-1} \lfloor \partial\Omega \qquad \text{and} \qquad -\Delta w = \chi - \phi \qquad \text{in } \mathcal{D}'(Q),$$

with periodic boundary conditions. Here $d\mathcal{H}^{n-1}\lfloor\partial\Omega$ denotes the Hausdorff measure restricted onto $\partial\Omega$.

We assume that for all but a finite number of times the solution of the Mullins–Sekerka equation is smooth, and for those times we have the relation $\partial_t w = u$. By definition $L^2(t) = \fint_Q |\nabla w|^2 \, dx$ and we find

$$\begin{aligned} L(t)\dot{L}(t) &= \fint_Q \nabla w \cdot \nabla u \, dx \\ &\leq \Big(\fint_Q |\nabla u|^2 \, dx\Big)^{1/2} \Big(\fint_Q |\nabla w|^2 \, dx\Big)^{1/2} \\ &= \Big(\fint_Q |\nabla u|^2 \, dx\Big)^{1/2} L(t). \end{aligned}$$

Hence, dividing by $L(t)$, integrating by parts, and using (1.1)–(1.3), it follows that

$$|\dot{L}(t)|^2 \leq \fint_Q |\nabla u|^2 \, dx = \frac{1}{|Q|} \int_{\partial\Omega \cap Q} v\kappa \, d\mathcal{H}^{n-1} = -\dot{E}\,.$$

The last equality follows from the well-known fact that the mean curvature is the variation of the surface area with respect to kinematically admissible normal velocities of the surface (cf., e.g., [6, Chap. 10]). ∎

The final ingredient of Theorem 1.1 is an ODE lemma which we can take directly from [7].

To that aim we consider the rescaled quantities

$$\begin{aligned} \hat{E} &= \begin{cases} \phi^{(1-n)/n} E, & n \geq 3, \\ \phi^{-1/2}\big(\ln \phi^{-1}\big)^{-1/2} E, & n = 2, \end{cases} \\ \hat{L} &= \phi^{-(n+2)/2n} L, \\ \hat{t} &= \begin{cases} \phi^{-3/n} t, & n \geq 3, \\ \phi^{-3/2}\big(\ln \phi^{-1}\big)^{1/2} t, & n = 2, \end{cases} \end{aligned} \tag{1.10}$$

and obtain

$$\hat{E}\hat{L} \gtrsim 1 \qquad \text{and} \qquad |\dot{\hat{L}}|^2 \leq -\dot{\hat{E}}. \tag{1.11}$$

In order to arrive at (1.11) one has some freedom in the choice of factors in (1.10). The present choice is just the natural one in the sense that for typical configurations the new quantities are of order one in ϕ.

The ODE lemma (see [7, Lemma 3]) now yields for $\hat{T} \gg (\hat{L}_0)^3$

$$⨍_0^{\hat{T}} (\hat{E}(\hat{t}))^2 \, d\hat{t} \gtrsim \hat{T}^{-2/3},$$

which in view of (1.10) implies the estimate in Theorem 1.1.

2. Proof of the interpolation inequality. Let χ denote the characteristic function of the set Ω. We claim that for $\phi \leq \frac{1}{2}$ the following interpolation inequality is true:

$$⨍_Q (\chi - \phi)^2 \, dx \lesssim \Big(⨍_Q |\nabla \chi| \, dx \Big)^{2/3} \Big(⨍_Q |\nabla^{-1}(\chi - \phi)|^2 \, dx \Big)^{1/3}. \tag{2.1}$$

To see that, define $(\chi - \phi)_\varepsilon := (\chi - \phi) * \eta_\varepsilon$ and $\chi_\varepsilon := \chi * \eta_\varepsilon$, where η_ε is a standard sequence of mollifiers. We have

$$⨍_Q |\nabla \chi_\varepsilon|^2 \, dx \lesssim \frac{1}{\varepsilon^2} ⨍_Q |\chi - \phi|^2 \, dx \,,$$

such that by duality

$$⨍_Q |(\chi - \phi)_\varepsilon|^2 \, dx \lesssim \frac{1}{\varepsilon^2} ⨍_Q |\nabla^{-1}(\chi - \phi)_\varepsilon|^2 \, dx \leq \frac{1}{\varepsilon^2} ⨍_Q |\nabla^{-1}(\chi - \phi)|^2 \, dx \,. \tag{2.2}$$

Hence, using $|\chi|, |\chi_\varepsilon| \leq 1$, we find

$$\begin{aligned} ⨍_Q (\chi - \phi)^2 \, dx &\leq 2 ⨍_Q |\chi - \chi_\varepsilon|^2 + 2 ⨍_Q |(\chi - \phi)_\varepsilon|^2 \, dx \\ &\overset{(2.2)}{\lesssim} ⨍_Q |\chi - \chi_\varepsilon| + \frac{1}{\varepsilon^2} ⨍_Q |\nabla^{-1}(\chi - \phi)|^2 \, dx \\ &\lesssim \varepsilon ⨍_Q |\nabla \chi| + \frac{1}{\varepsilon^2} ⨍_Q |\nabla^{-1}(\chi - \phi)|^2 \, dx. \end{aligned}$$

The preceding inequality is optimal for

$$\varepsilon = \left(\frac{⨍_Q |\nabla^{-1}(\chi - \phi)|^2 \, dx}{⨍_Q |\nabla \chi|} \right)^{1/3},$$

which proves (2.1). Since

$$⨍_Q (\chi - \phi)^2 \, dx \geq \frac{1}{2} ⨍_Q \chi^2 \, dx = \frac{1}{2} ⨍_Q \chi \, dx = \frac{1}{2}\phi,$$

inequality (2.1) implies Lemma 1.2 for the case $n \geq 3$.

In the case $n = 2$ we can even do better. We claim that

$$\fint_Q \chi^2\,dx \lesssim \frac{1}{\left(\ln\phi^{-1}\right)^{1/3}}\left(\fint_Q |\nabla\chi|\,dx\right)^{2/3}\left(\fint_Q |\nabla^{-1}(\chi-\phi)|^2\,dx\right)^{1/3} \quad \text{if } n = 2. \tag{2.3}$$

We recall that by the definition of the dual norm we have for all $\zeta \in H^1_{per}(Q)$, that is, the space of H^1 functions with periodic boundary conditions, that

$$\fint_Q |\nabla^{-1}(\chi-\phi)|^2\,dx \ge \frac{\left(\fint_Q (\chi-\phi)\zeta\,dx\right)^2}{\fint_Q |\nabla\zeta|^2\,dx}. \tag{2.4}$$

We are going to construct a periodic test function $\zeta \ge 0$ which satisfies the following properties:

$$\fint_Q \chi\zeta\,dx \ge \frac{1}{2}\fint_Q \chi\,dx, \tag{2.5}$$

$$\fint_Q \zeta\,dx \ll 1\,, \tag{2.6}$$

$$\fint_Q |\nabla\zeta|^2\,dx \lesssim \frac{\phi}{\mathcal{R}^2\ln\phi^{-1}}, \tag{2.7}$$

where

$$\mathcal{R} := \frac{\phi}{\fint_Q |\nabla\chi|}. \tag{2.8}$$

Then (2.5) and (2.6) imply

$$\fint_Q (\chi-\phi)\zeta\,dx \ge \frac{1}{4}\fint_Q \chi\,dx = \frac{1}{4}\fint_Q \chi^2\,dx,$$

and using (2.4) and (2.7) we find

$$\left(\fint_Q \chi^2\,dx\right)^2 \lesssim \left(\fint_Q (\chi-\phi)\zeta\,dx\right)^2 \le \frac{\left(\fint_Q |\nabla\chi|\right)^2}{\phi\ln\phi^{-1}}\fint_Q |\nabla^{-1}(\chi-\phi)|^2\,dx,$$

which, recalling $\fint_Q \chi^2 = \phi$, proves (2.3).

Hence it remains to construct a test function ζ with the properties (2.5)–(2.7). Fundamental for the construction is the following geometric lemma which says that for given Ω we can construct another set $\Omega_{\mathcal{R}}$ which covers a substantial part of Ω and behaves like a union of balls with radius larger than $\mathcal{R}$.

LEMMA 2.1. *Let $\Omega \subset \mathbb{R}^n$ be a given Q-periodic set and χ its characteristic function, and let $\mathcal{R}$ be such that*

$$\mathcal{R}\int_Q |\nabla\chi| \le |\Omega\cap Q|.$$

Then there exists a Q-periodic set $\Omega_{\mathcal{R}} \subset \mathbb{R}^n$ such that

$$|\Omega\cap\Omega_{\mathcal{R}}\cap Q| \ge \frac{1}{2}|\Omega\cap Q|$$

and, for all $r > 0$,

$$|\Omega_{\mathcal{R}}^r \cap Q| \le C|\Omega \cap Q| \left(1 + \left(\frac{r}{\mathcal{R}}\right)^n\right),$$

where $\Omega_{\mathcal{R}}^r = \{x \,|\, \mathrm{dist}(x, \Omega_{\mathcal{R}}) \le r\}$ *and* C *is a constant which depends only on* n.

We postpone the proof of the lemma, choose $\mathcal{R}$ as in (2.8), and define

$$\zeta(x) := \psi(r(x)),$$

where $r(x) := \mathrm{dist}(x, \Omega_{\mathcal{R}})$ and

$$\psi(r) = \begin{cases} 0, & r \ge d, \\ \frac{\ln \frac{d}{r}}{\ln \frac{d}{\mathcal{R}}}, & r \in (\mathcal{R}, d), \\ 1, & r \le \mathcal{R} \end{cases}$$

with $d = \frac{\mathcal{R}}{\phi^{1/2}} \gg \mathcal{R}$. This choice of test function is motivated by the simplest case, when $\Omega = B_{\mathcal{R}}(0)$ and $B_d(0) \subset Q$, whence $\zeta(x) := \ln \frac{d}{|x|} / \ln \frac{d}{\mathcal{R}}$ fulfills (2.5)–(2.7). (Here and in the following we use the notation $B_r(x)$ to denote a ball with radius r and center x.)

The fact that $|\nabla r(x)| = 1$ for almost all x and Theorem 2, Ch.3.4.3 and Proposition 2, Ch. 3.4.4 of [5] imply

$$\begin{aligned}
\int_Q |\nabla \zeta|^2 \, dx &= \int_Q |\psi'(r(x))|^2 |\nabla r(x)|^2 \, dx \\
&= \int_0^\infty |\psi'(s)|^2 \mathcal{H}^{n-1}(\{x \in Q \,|\, r(x) = s\}) \, ds \\
&= \int_{\mathcal{R}}^d |\psi'(s)|^2 \mathcal{H}^{n-1}(\{x \in Q \,|\, r(x) = s\}) \, ds \\
&\le - \int_{\mathcal{R}}^d \frac{d}{ds} |\psi'(s)|^2 |\{x \in Q \,|\, r(x) < s\}| \, ds \\
&\quad + |\psi'(d)|^2 |\{x \in Q \,|\, r(x) < d\}| \,.
\end{aligned}$$

Lemma 2.1 for $n = 2$ gives $|\{x \in Q \,|\, r(x) < s\}| \le C|\Omega \cap Q| \left(\frac{s}{\mathcal{R}}\right)^2$, and a straightforward computation yields

$$\begin{aligned}
\fint_Q |\nabla \zeta|^2 \, dx &\lesssim \int_{\mathcal{R}}^d \frac{\phi}{s^3 |\ln \frac{d}{\mathcal{R}}|^2} \left(\frac{s}{\mathcal{R}}\right)^2 ds \; + \; \frac{\phi}{|d \ln \frac{d}{\mathcal{R}}|^2} \left(\frac{d}{\mathcal{R}}\right)^2 \\
&\lesssim \frac{\phi}{\mathcal{R}^2 |\ln \frac{d}{\mathcal{R}}|},
\end{aligned}$$

which establishes (2.7).

Furthermore, an analogous computation yields

$$\begin{aligned}
\int_Q \zeta \, dx &= \int_{\mathcal{R}}^d \psi(s) \mathcal{H}^{n-1}(\{x \in Q \,|\, r(x) = s\}) \, ds + |\Omega_{\mathcal{R}}^{\mathcal{R}} \cap Q| \\
&= - \int_{\mathcal{R}}^d \psi'(s) |\{x \in Q \,|\, r(x) < s\}| \, ds + \psi(s) |\{x \in Q \,|\, r(x) < s\}| \Big|_{\mathcal{R}}^d + |\Omega_{\mathcal{R}}^{\mathcal{R}} \cap Q| \,.
\end{aligned} \tag{2.9}$$

Since $\psi(d) = 0$ and $\psi(\mathcal{R}) = 1$ the last two terms cancel. Lemma 2.1 and the choice of d imply

$$\fint_Q \zeta\, dx \lesssim \frac{1}{\ln \frac{d}{\mathcal{R}}} \int_{\mathcal{R}}^d \frac{\phi s}{\mathcal{R}^2}\, ds \lesssim \frac{d^2}{\mathcal{R}^2 \ln \frac{d}{\mathcal{R}}} \phi = \frac{1}{\ln \phi^{-1/2}} \ll 1,$$

so that the desired property (2.6) follows.

Proof of Lemma 2.1. We first observe that it suffices to prove the statement for $|Q \cap \Omega| \le \frac{1}{2}|Q|$; otherwise one can take $\Omega_{\mathcal{R}} = Q$. Next, it is convenient to redefine $\mathcal{R}$ such that

$$\mathcal{R} \int_Q |\nabla \chi| \le \frac{1}{4} |\Omega \cap Q|. \tag{2.10}$$

We notice that the isoperimetric inequality in particular implies that

$$\mathcal{R} \le \frac{1}{4} |\Omega \cap Q|^{1/n} \le \frac{1}{4} |Q|^{1/n}. \tag{2.11}$$

We set

$$\Omega_{\mathcal{R}} := \left\{ x \,|\, |\Omega \cap B_{\mathcal{R}}(x)| > \frac{1}{2} |B_{\mathcal{R}}(x)| \right\} \tag{2.12}$$

and claim that

$$|\Omega \cap \Omega_{\mathcal{R}} \cap Q| \ge \frac{1}{2} |\Omega \cap Q|. \tag{2.13}$$

We first notice that

$$\Omega_{\mathcal{R}} = \left\{ x \,|\, \chi_{\mathcal{R}}(x) > \frac{1}{2} \right\}, \qquad \text{where} \qquad \chi_{\mathcal{R}}(x) := \frac{|\Omega \cap B_{\mathcal{R}}(x)|}{|B_{\mathcal{R}}(x)|},$$

and $\chi_{\mathcal{R}}$ can be considered as a convolution of the characteristic function χ of Ω.

Furthermore we have

$$\chi - \chi_{\mathcal{R}} \ge 1 - \frac{1}{2} = \frac{1}{2} \qquad \text{on } \Omega \backslash \Omega_{\mathcal{R}},$$

so that

$$\begin{aligned} |(\Omega \backslash \Omega_{\mathcal{R}}) \cap Q| &\le 2 \int_Q |\chi - \chi_{\mathcal{R}}|\, dx \\ &\le 2 \sup_{|h| \le \mathcal{R}} \int_Q |\chi(x) - \chi(x+h)|\, dx \\ &\le 2\mathcal{R} \int_Q |\nabla \chi| \\ &\le \frac{1}{2} |\Omega \cap Q|, \end{aligned}$$

where the last inequality follows from (2.10). Thus

$$|\Omega \cap \Omega_{\mathcal{R}} \cap Q| = |\Omega \cap Q| - |(\Omega \backslash \Omega_{\mathcal{R}}) \cap Q| \ge \frac{1}{2} |\Omega \cap Q|,$$

which establishes (2.13).

Next we claim that there exists a set $A \subset \Omega_{\mathcal{R}} \cap Q$ of points such that

$$\#A \leq C\frac{|\Omega \cap Q|}{|B_{\mathcal{R}}|} \qquad \text{and} \qquad \Omega_{\mathcal{R}} \cap Q \subset \bigcup_{x\in A} B_{2\mathcal{R}}(x), \tag{2.14}$$

where $\#A$ denotes the cardinality of the set A. To that aim let $A \subset \Omega_{\mathcal{R}} \cap Q$ be a maximal family such that

$$\{B_{\mathcal{R}}(x)\}_{x\in A} \quad \text{are disjoint.} \tag{2.15}$$

We claim that

$$\Omega_{\mathcal{R}} \cap Q \subset \bigcup_{x\in A} B_{2\mathcal{R}}(x) \cap Q. \tag{2.16}$$

Indeed, assume that (2.16) were wrong. Then there exists $y \in \Omega_{\mathcal{R}} \cap Q$ such that for all $x \in A$ we have $y \notin B_{2\mathcal{R}}(x)$; that is, for all $x \in A$ we find $B_{\mathcal{R}}(y) \cap B_{\mathcal{R}}(x) = \emptyset$. This contradicts the maximality of A.

We now have, because of $A \subset \Omega_{\mathcal{R}} \cap Q$,

$$\begin{aligned} \#A|B_{\mathcal{R}}| \quad &= \quad \sum_{x\in A} |B_{\mathcal{R}}(x)| \\ &\overset{(2.12)}{<} \quad 2\sum_{x\in A} |\Omega \cap B_{\mathcal{R}}(x)| \\ &\overset{(2.15),(2.11)}{\leq} \quad C(n)|\Omega \cap Q|. \end{aligned}$$

Finally we claim

$$|\Omega_{\mathcal{R}}^r \cap Q| \leq C(n)|\Omega \cap Q|\Big(2 + \frac{r}{\mathcal{R}}\Big)^n. \tag{2.17}$$

In view of (2.16) we have the inclusion $\Omega_{\mathcal{R}} \cap Q \subset \bigcup_{x\in A} B_{2\mathcal{R}}(x) \cap Q$, which implies $\Omega_{\mathcal{R}}^r \subset \bigcup_{x\in A} B_{2\mathcal{R}+r}(x) \cap Q$. Thus

$$\begin{aligned} |\Omega_{\mathcal{R}}^r \cap Q| \quad &\leq \quad \sum_{x\in A} |B_{2\mathcal{R}+r}(x) \cap Q| \\ &\overset{(2.14)}{\leq} \quad C\frac{|\Omega \cap Q|}{|B_{\mathcal{R}}|}|B_{2\mathcal{R}+r}| \\ &= \quad C|\Omega \cap Q|\Big(\frac{2\mathcal{R}+r}{\mathcal{R}}\Big)^n, \end{aligned}$$

which proves (2.17) and thus completes the proof of Lemma 2.1. ∎

REFERENCES

[1] N. Alikakos and G. Fusco, *Ostwald ripening for dilute systems under quasistationary dynamics*, Comm. Math. Phys., 238 (2003), pp. 429–479.

[2] N. Alikakos, G. Fusco, and G. Karali, *Ostwald ripening in two dimensions*, J. Differential Equations, 205 (2004), pp. 1–49.

[3] S. Dai and R. L. Pego, *Universal bound on coarsening rates for mean-field models of phase transitions*, SIAM J. Math. Anal., 37 (2005), pp. 347–371.

[4] S. Dai and R. L. Pego, *An upper bound on the coarsening rate for mushy zones in a phase-field model*, Interfaces Free Bound., 7 (2005), pp. 187–197.
[5] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math., CRC Press, Boca Raton, 1992.
[6] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Basel, 1984.
[7] R. V. Kohn and F. Otto, *Upper bounds for coarsening rates*, Comm. Math. Phys., 229 (2002), pp. 375–395.
[8] R. V. Kohn and X. Yan, *Coarsening rates for models of multicomponent phase separation*, Interfaces Free Bound., 6 (2004), pp. 135–149.
[9] I. M. Lifshitz and V. V. Slyozov, *The kinetics of precipitation from supersaturated solid solutions*, J. Phys. Chem. Solids, 19 (1961), pp. 35–50.
[10] B. Niethammer and F. Otto, *Domain coarsening in thin films*, Comm. Pure Appl. Math., 54 (2001), pp. 361–384.
[11] F. Otto, T. Rump, and D. Slepčev, *Coarsening rates for a droplet model*, SIAM J. Math. Anal., to appear.
[12] C. Wagner, *Theorie der Alterung von Niederschlägen durch Umlösen*, Z. Elektrochemie, 65 (1961), pp. 581–594.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1742–1760

UNIQUENESS OF WEAK SOLUTIONS OF THE NAVIER–STOKES EQUATIONS OF MULTIDIMENSIONAL, COMPRESSIBLE FLOW*

DAVID HOFF†

Abstract. We prove uniqueness and continuous dependence on initial data of weak solutions of the Navier–Stokes equations of compressible flow in two and three space dimensions. The solutions we consider may display codimension-one discontinuities in density, pressure, and velocity gradient, and consequently are the generic singular solutions of this system. The key point of the analysis is that solutions with minimal regularity are best compared in a Lagrangean framework; that is, we compare the instantaneous states of corresponding fluid particles in two different solutions rather than the states of different fluid particles instantaneously occupying the same point of space-time. Estimates for H^{-1} differences in densities and L^2 differences in velocities are obtained by duality from bounds for the corresponding adjoint system.

Key words. uniqueness, continuous dependence, Navier–Stokes equations, compressible flow

AMS subject classifications. 76N10, 35Q30

DOI. 10.1137/040618059

1. Introduction.

We prove uniqueness and continuous dependence on initial data of weak solutions of the Navier–Stokes equations of compressible flow

$$
\begin{cases} \rho_t + \mathrm{div}(\rho u) = 0, \\ (\rho u^j)_t + \mathrm{div}(\rho u^j u) + P(\rho)_{x_j} = \mu \Delta u^j + \lambda \mathrm{div} u_{x_j} + \rho f^j \end{cases} \tag{1.1}
$$

for $x \in \mathbb{R}^n$, $n = 2, 3$, and $t > 0$, and with initial data

$$
(\rho, u)|_{t=0} = (\rho_0, u_0). \tag{1.2}
$$

Here ρ and $u = (u^1, \dots, u^n)$ are the unknown functions of x and t representing density and velocity, $P = P(\rho)$ is the pressure, f is a given external force, $\mu > 0$ and $\lambda \geq 0$ are viscosity constants, and div and Δ are the usual spatial divergence and Laplace operators.

We compare solutions with only minimal regularity, including the generic singular solutions of the Navier–Stokes system—solutions with codimension-one discontinuities in density, pressure, and velocity gradient (see Hoff [6]). Our result extends that of Danchin [1], which applies to solutions in certain Besov spaces of continuous functions. Danchin makes a direct comparison by subtracting the equations satisfied by different solutions and controlling the linearization errors by the assumed regularity. This regularity is absent for solutions in the class considered here, however, and a different approach is required.

One could hope to prove uniqueness and continuous dependence for solutions in the largest class for which existence is known, that is, solutions with finite energies and nonnegative densities (see Feireisl [3] and Feireisl, Novotny, and Petzeltova [4], for example). Such a result may be unattainable, however, and may be in fact be

*Received by the editors November 1, 2004; accepted for publication (in revised form) August 15, 2005; published electronically February 15, 2006. This research was supported in part by the NSF under grant DMS–0305072.
http://www.siam.org/journals/sima/37-6/61805.html

†Department of Mathematics, Indiana University, Bloomington, IN 47405 (hoff@indiana.edu).

false: it is known that certain anomalies can arise in solutions of (1.1) when densities are zero on nonnegligible sets, and there are examples of nonphysical solutions. To be more specific, we say that a solution (ρ, u) is *locally momentum conserving* if, whenever E and V are bounded open sets in $\mathbb{R}^n$ with $\bar{E} \subset V$ and with $\rho = 0$ a.e. in $(V - E) \times [t_1, t_2]$, then the change in the momentum of the fluid in E from time t_1 to time t_2 should be the impulse $\int_{t_1}^{t_2} \int_E \rho f \, dxdt$ applied by the external force to the fluid in E. Weak solutions violating this condition do in fact exist, one such solution being constructed in Hoff and Serre [8]: initial data (ρ_0, u_0) is given corresponding to two fluids initially separated by a vacuum; it is shown that, if ρ_0 is replaced by $\rho_0 + \delta$, then the limit as $\delta \to 0$ of the corresponding perturbed solutions exists and is a weak solution in the entire physical space but fails to be locally momentum conserving. It is also shown that there is a different, locally momentum conserving solution corresponding to the same initial data, but this solution is defined only on the support of the density and cannot be extended as a weak solution to the entire space. We also point out that it appears likely, although it is not presently known, that even if the initial density is strictly positive, large energies can cause spontaneous cavitation. (This cannot occur in one space dimension, however, but the argument, given in Hoff and Smoller [9], depends upon restoring forces which scale unfavorably with dimension.) It therefore remains an interesting open problem to select physical solutions from the most general class of known weak solutions with large energies and possible vacuum states and to establish their uniqueness and continuous dependence on initial data. In the present paper we avoid these issues and difficulties by assuming that at least one of the solutions being compared has strictly positive density.

Our regularity assumptions are listed below in (1.4)–(1.14). These have been made as weak as the analysis will allow and consequently are somewhat technical. We discuss these assumptions in some detail below following the statement of the theorem, giving at the very end of this section six somewhat stronger *sufficient* conditions under which our result applies. Five of these are fairly mild a priori conditions on the system parameters P, μ, and λ and on the forces and initial data, and are easily checked.

The sixth condition is an a posteriori condition imposed on solutions rather than on data, and therefore deserves special comment. A key point of our analysis is that weak solutions are best compared in a Lagrangean framework: it is more natural to compare the instantaneous states of corresponding fluid particles in different solutions rather than the states of different fluid particles which are instantaneously at the same point of space-time. Consequently, the function which maps the position x of a fluid particle in one solution to the position of the corresponding particle in the other solution at a given time plays a special role in the analysis. Our proof requires that this function be Lipschitz continuous with respect to x, a condition which follows from the assumption that velocity gradients are in the space $L^1((0,T); L^\infty(\mathbb{R}^n))$; this is the sixth condition. Now, a large class of weak solutions is constructed in Hoff [6] that satisfy this regularity condition and that at the same time exhibit the generic singularities described above. However, a simple characterization of singular data for which there exist corresponding solutions satisfying this sixth condition has been elusive. The difficulty is essentially equivalent to the classical fact that Calderon–Zygmund operators map L^p into L^p for $p \in (1, \infty)$ but not for $p = 1$ or ∞. We shall discuss this point in greater detail below.

We compare two weak solutions (ρ, u, f) and $(\overline{\rho}, \overline{u}, \overline{f})$ defined on $\mathbb{R}^n \times (0,T)$ and obtain bounds for

$$\left(\int_0^T \int_{\mathbb{R}^n} |u - \overline{u}|^2 dxdt \right)^{1/2} + \sup_{0 \le t \le T} \|(\rho - \overline{\rho})(\cdot, t)\|_{H^{-1}(\mathbb{R}^n)}$$

in terms of $\|\rho_0-\overline{\rho}_0\|_{L^2}$, $\|\rho_0 u_0-\overline{\rho}_0\overline{u}\|_{L^2}$, and $\left(\int_0^T\int|f-\overline{f}|^2dxdt\right)^{1/2}$. Now, a cursory examination of the second equation in (1.1) suggests that L^2 differences in u are driven by H^{-1} differences in $P(\rho)$, whereas H^{-1} bounds for differences in ρ, not $P(\rho)$, are natural in this theory. These are the same when P is linear or more generally when the divided difference $(\rho-\overline{\rho})^{-1}[P(\rho)-P(\overline{\rho})]$ is somewhat regular. Our theorem will therefore apply either when P is the pressure $P=K\rho$ of an ideal isothermal fluid, or for more general pressures when somewhat greater regularity conditions are imposed on solutions; see (1.16) below and the corresponding discussion following (3.13).

Before describing our result in more detail, we introduce some standard notation. First, the usual convective derivative $\frac{d}{dt}$ of a given function $w:\mathbb{R}^n\times(0,T)\to\mathbb{R}$ with respect to a velocity field u is given by $\frac{dw}{dt}=\dot{w}=w_t+u\cdot\nabla w$, where ∇w is the spatial gradient of w, and for $w:\mathbb{R}^n\times(0,T)\to\mathbb{R}^n$, $\frac{dw}{dt}=\dot{w}=w_t+\nabla wu$, where ∇w is the $n\times n$ matrix of partial derivatives of w. We apply this notation to rewrite the momentum equation in (1.1): subtracting u^j times the first equation in (1.1) from the second, we obtain

$$\rho\dot{u}^j+P(\rho)_{x_j}=\mu\Delta u^j+\lambda\,\mathrm{div}\,u_{x_j}+\rho f^j,$$

and adding and subtracting terms, that

$$\begin{aligned}\rho\dot{u}^j&=[(\mu+\lambda)\mathrm{div}\,u-P(\rho)+P(\widetilde{\rho})]_{x_j}+\mu(u^j_{x_k}-u^k_{x_j})_{x_k}+\rho f^j\\&\equiv F_{x_j}+\mu\omega^{j,k}_{x_k}+\rho f^j\end{aligned}\tag{1.3}$$

(summation over repeated indices is understood, and $\widetilde{\rho}$ is a constant, positive reference density). Here ω is the vorticity matrix, which together with the scalar quantity F plays an important role in the existence theory: F and ω are more regular than are ∇u and ρ in general, and (1.3) gives a representation of the internal surface forces acting in the fluid as the sum of a divergence-free field and the gradient of a scalar field. As we shall see, F and ω play an important role in the uniqueness theory as well.

We now give a precise formulation of our results. Let $\widetilde{\rho}$ be a fixed, positive, constant reference density as above and let $\widetilde{P}=P(\widetilde{\rho})$. The solutions we consider will be weak solutions in the following sense.

Definition. *A weak solution on $\mathbb{R}^n\times[0,T]$ of the system (1.1)–(1.2) is a triple (ρ,u,f) satisfying the following:*

(1.4) *$\rho-\widetilde{\rho}$ is a bounded map from $[0,T]$ into $L^1_{loc}(\mathbb{R}^n)\cap H^{-1}(\mathbb{R}^n)$ and $\rho\ge 0$ a.e.;*

(1.5)
$\rho_0u_0\in L^2(\mathbb{R}^n)$; ρu, $P-\widetilde{P}$, ∇u, $\rho f\in L^2(\mathbb{R}^n\times(0,T))$; $\rho|u|^2\in L^1(\mathbb{R}^n\times(0,T))$;

(1.6) *For $0\le t_1\le t_2\le T$ and for test functions φ which are Lipschitz on $\mathbb{R}^n\times[t_1,t_2]$ and for which $\mathrm{supp}\,\varphi(\cdot,t)\subset K$, $t_1\le t\le t_2$, where K is compact,*

$$\int_{\mathbb{R}^n}\rho\varphi dx\Big|_{t_1}^{t_2}=\int_{t_1}^{t_2}\int_{\mathbb{R}^n}(\rho\varphi_t+\rho u\cdot\nabla\varphi)dxdt$$

(it is understood here that $\rho(\cdot,0)=\rho_0$);

(1.7) The weak form of the momentum equation

$$-\int_{\mathbb{R}^n}\rho_0 u_0\cdot\psi(\cdot,0)dx=\int_0^T\int_{\mathbb{R}^n}[\rho u\cdot(\psi_t+\nabla\psi u)+(P(\rho)-\widetilde{P})\mathrm{div}\,\psi \\ -\mu\nabla u^j\cdot\nabla\psi^j-\lambda(\mathrm{div}\,u)(\mathrm{div}\,\psi)+\rho f\cdot\psi]dxdt$$

holds for test functions ψ which are locally Lipschitz on $\mathbb{R}^n\times[0,T]$ and for which $\psi,\psi_t,\nabla\psi\in L^2(\mathbb{R}^n\times(0,T))$, $\nabla\psi\in L^\infty(\mathbb{R}^n\times(0,T))$, and $\psi(\cdot,T)=0$.

The two solutions (ρ,u,f) and $(\overline{\rho},\overline{u},\overline{f})$ we compare will also be assumed to satisfy

(1.8) $$u,\overline{u}\in C(\mathbb{R}^n\times(0,T])\cap L^1((0,T);W^{1,\infty}(\mathbb{R}^n))\cap L^\infty_{loc}((0,T];L^\infty(\mathbb{R}^n))$$

and

(1.9) $$\rho-\widetilde{\rho},\ \overline{\rho}-\widetilde{\rho},\ u,\ \overline{u},\ f,\overline{f}\in L^2(\mathbb{R}^n\times(0,T)).$$

One of the solutions (ρ,u,f) will have to satisfy

(1.10) $$\rho,\rho^{-1}\in L^\infty(\mathbb{R}^n\times(0,T))$$

and

(1.11) $$\int_0^T\int_{\mathbb{R}^n}|u|^r dxdt<\infty$$

for some $r>n$, and the other solution $(\overline{\rho},\overline{u},\overline{f})$ will have to satisfy

(1.12)
$$\int_0^T\left[\|\overline{u}(0,t)\|^2_{L^\infty}+t\|\nabla\overline{u}(\cdot,t)\|^2_{L^\infty}+t\|\nabla\overline{F},\nabla\overline{\omega}\|^2_{L^2}+(t\|\nabla\overline{F},\nabla\overline{\omega}\|^2_{L^4})^a\right]dt<\infty,$$

where $\bar{F}$ and $\bar{\omega}$ are as in (1.3), the gradients are with respect to x, and $a=2/3$ for $n=2$ and $a=4/5$ for $n=3$; and

(1.13) $$\overline{f}\in L^1((0,T);L^{2q}(\mathbb{R}^n))$$

for some $q\in[1,\infty]$. Finally, we shall assume that

(1.14) $$\rho_0-\overline{\rho}_0\in(L^2\cap L^{2p})(\mathbb{R}^n),$$

where p is the Hölder conjugate of q.

We now state our main result.

Theorem. *Given $P=K\rho$, M, T, and $r>n$, where n is 2 or 3, there is a constant C depending on K, M, T, and r such that if (ρ,u,f) and $(\overline{\rho},\overline{u},\overline{f})$ are weak solutions of* (1.1) *satisfying* (1.4)–(1.9)*, with (ρ,u,f) satisfying* (1.10) *and* (1.11) *and $(\overline{\rho},\overline{u},\overline{f})$ satisfying* (1.12) *and* (1.13)*, if* (1.14) *holds, and if all norms occurring in these conditions are bounded by M, then*

(1.15)
$$\left(\int_0^T\int_{\mathbb{R}^n}|u-\overline{u}|^2dxdt\right)^{1/2}+\sup_{0\le t\le T}\|(\rho-\overline{\rho})(\cdot,t)\|_{H^{-1}(\mathbb{R}^n)} \\ \le C\left[\|\rho_0-\overline{\rho}_0\|_{L^2\cap L^{2p}}+\|\rho_0u_0-\overline{\rho}_0\overline{u}_0\|_{L^2}+\left(\int_0^T\int_{\mathbb{R}^n}|f-\overline{f}\circ S|^2dxdt\right)^{1/2}\right]$$

(the notation $\overline{f}\circ S$ *is explained below in* (1.18)*). If* $\int_0^T t\|\nabla\overline{f}(\cdot,t)\|_{L^\infty}dt \le M$*, then* $\overline{f}\circ S$ *may be replaced by* $\overline{f}$ *in* (1.15)*. Also, these same results hold with no restriction on the pressure function* P *provided that*

$$\sup_{0\le t\le T}\left\|\nabla\left(\frac{P(\rho(\cdot,t))-P(\overline{\rho}(\cdot,t))}{\rho(\cdot,t)-\overline{\rho}(\cdot,t)}\right)\right\|_{L^\alpha}<\infty, \tag{1.16}$$

where $\alpha>2$ *when* $n=2$ *and* $\alpha=3$ *when* $n=3$*; in this case the constant* C *depends additionally on* α *and on the above* sup.

The key point of the analysis, and the major premise of this paper, is that solutions with minimal regularity are best compared in a Lagrangean framework, as mentioned earlier. Now, in one space dimension, the initial position of a fluid particle and the cumulative mass are equivalent Lagrangean coordinates, and the corresponding Lagrangean formulation of the system (1.1) consists of simple equations in divergence form for which a straightforward weak interpretation is possible. In several space variables, however, the only Lagrangean coordinate is the initial position, and the system (1.1), written in Lagrangean coordinates, is neither simple nor in divergence form, making a weak interpretation problematic. We deal with this difficulty as follows. First, by imposing the condition (1.8), we ensure the existence of particle trajectories $X(y,t,s)$ satisfying

$$\begin{cases}\dfrac{\partial X}{\partial t}(y,t,s)=u(X(y,t,s),t),\\ X(y,s,s)=y,\end{cases} \tag{1.17}$$

and similarly for $\overline{X}$. $X(y,t,s)$ is therefore the position at time t of the fluid particle whose position at time s is y. In particular, the fluid particle in (ρ,u,f) at (x,t) was at $X(x,0,t)$ initially and so corresponds to the particle in $(\overline{\rho},\overline{u},\overline{f})$ which at time t is at the point

$$\overline{X}(X(x,0,t),t,0)\equiv S(x,t) \tag{1.18}$$

(this is the S appearing in (1.15)). The "Lagrangean" comparison therefore consists of an estimate for $u-\overline{u}\circ S$, where we abuse notation slightly and abbreviate $\overline{u}(S(x,t),t)=(\overline{u}\circ S)(x,t)$. The advantages of the Lagrangean formulation may be understood briefly as follows. First, from its definition, $S(X(y,t,0),t)=\overline{X}(y,t,0)$, and thus $S_t+\nabla Su=\overline{u}\circ S$. It then follows easily that if $w=\overline{u}\circ S$,

$$(\overline{u}_t+\nabla\overline{u}\,\overline{u})\circ S=w_t+\nabla wu,$$

a result which could have been predicted from the meaning of convective differentiation. Given the form of the momentum equation, the comparison of u with $w=\overline{u}\circ S$ now appears quite natural and proceeds in the usual way: bounds for $u-\overline{u}\circ S$ are obtained by duality from estimates for solutions of the adjoint of the weak equation satisfied by $u-\overline{u}\circ S$. A bound for the Eulerian difference $u-\overline{u}$ can then be derived via the regularity assumption in (1.12) for $\|\nabla\overline{u}\|_{L^\infty}$ (which is required elsewhere in the analysis) and the simple fact that

$$\int|x-S(x,t)|^2dx\le Ct\int_0^t\int|u-\overline{u}\circ S|^2dxdt.$$

The convenience of the Lagrangean framework comes at a price, however: we require at several points in the analysis a bound for $\sup_t\|\nabla S(\cdot,t)\|_{L^\infty}$. This bound follows from corresponding bounds for $\nabla X(\cdot, 0, t)$ and $\nabla \overline{X}(\cdot, t, 0)$, and these in turn are consequences of the hypothesis in (1.8) that $u, \overline{u} \in L^1((0,T); W^{1,\infty}(\mathbb{R}^n))$, about which more below.

We now discuss the existence of solutions satisfying the conditions listed in (1.4)–(1.12). First we recall the results of Hoff [7], which apply to flows both in half-spaces and in the whole space: under somewhat technical but fairly mild restrictions on P, μ, λ, and f, which we refer to as conditions (H), the system (1.1)–(1.2) has a global weak solution (ρ, u) provided that

$$\begin{cases} 0 \le \rho_0 \text{ a.e.}, \quad \rho_0 \in L^\infty, \\ \displaystyle\int \left[\rho_0 |u_0|^q + \varphi(\rho_0 |u_0|^2 + (\rho_0 - \tilde{\rho})^2)\right] dx < \infty \\ \displaystyle\int [\rho_0 |u_0|^2 + (\rho_0 - \widetilde{\rho})^2] dx \ll 1, \end{cases} \tag{1.19}$$

where $q > 6$ and $\varphi = 1$ for $n = 3$ and $q > 2$ and $\varphi = (1 + |x|^2)^\beta, \beta > 0$, for $n = 2$. The solution is shown to satisfy all the conditions in (1.4)–(1.9) as well as the energy estimates

$$\begin{aligned} &\sup_{0 \le t \le T} \int \left[\rho |u|^2 + t|\nabla u|^2 + t^n \rho |\dot{u}|^2\right] dx \\ &\qquad + \int_0^T \int \left[|\nabla u|^2 + t|(\rho u^j)_t + \operatorname{div}(\rho u^j u)|^2 + t^n |\nabla \dot{u}|\right]^2 dx dt \\ &\quad \le C(T), \end{aligned}$$

with the exception of the condition in (1.8) that $u \in L^1((0,T); W^{1,\infty}(\mathbb{R}^n))$. Moreover, if $\inf \rho_0 > 0$, then $\rho \ge C^{-1} > 0$ a.e., and as a consequence, the conditions (1.10) and (1.11) hold. To summarize, *the results of* [7] *guarantee the existence of a solution* (ρ, u, f) *satisfying all the conditions* (1.4)–(1.11)*, with the exception of the requirement in* (1.8) *that* $u \in L^1((0,T); W^{1,\infty}(\mathbb{R}^n))$*, provided only that the system conditions* (H) *hold, that* (ρ_0, u_0) *satisfies* (1.19)*, and that* $\inf \rho_0 > 0$.

Next, by applying the interpolation techniques of [6], we can show that if $\inf \overline{\rho}_0 > 0$ and $\overline{u}_0 \in H^s(\mathbb{R}^n)$ for some $s \in [0,1]$, then the improved energy estimates

$$\begin{aligned} &\sup_{0 \le t \le T} \int \left[\overline{\rho} |\overline{u}|^2 + (\overline{\rho} - \tilde{\rho})^2 + t^{1-s} |\nabla \overline{u}|^2 + t^\sigma |\dot{\overline{u}}|^2\right] dx \\ &\qquad + \int_0^T \int \left[|\nabla \overline{u}|^2 + t^{1-s} |\dot{\overline{u}}|^2 + t^\sigma |\nabla \dot{\overline{u}}|^2\right] dx dt \\ &\quad \le C(T) \end{aligned} \tag{1.20}$$

hold, where

$$\sigma = \begin{cases} 2 - s, & n = 2, \\ \max\{2 - s, 3 - 3s\}, & n = 3. \end{cases} \tag{1.21}$$

It then follows that the bounds in (1.12) hold if $s > 0$ when $n = 2$ and $s > 1/2$ when $n = 3$. For example, when $n = 2$ we can apply the fact that $\nabla \overline{F} = \operatorname{div}(\overline{\rho}\, \dot{\overline{u}} - \overline{\rho} \overline{f})$ (see (1.3)) together with standard elliptic theory and Sobolev estimates to obtain that, modulo constants depending on $\overline{f}$,

$$\int_0^T t^{2/3}\|\nabla\overline{F}(\cdot,t)\|_{L^4}^{4/3}dt \le C\int_0^T t^{2/3}\left(\int|\dot{\overline{u}}|^2dx\right)^{1/3}\left(\int|\nabla\dot{\overline{u}}|^2dx\right)^{1/2}dt$$
$$\le C\left(\int_0^T t^{2s-1}dt\right)^{1/3}\left(\int_0^T\int t^{1-s}|\dot{\overline{u}}|^2dxdt\right)^{1/3}\left(\int_0^T\int t^{2-s}|\nabla\dot{\overline{u}}|^2dxdt\right)^{1/3},$$

which is finite if $s > 0$. A similar result holds for $n = 3$ if $s > 1/2$. To summarize, *the results of* [6] *and* [7] *guarantee the existence of a solution* $(\overline{\rho},\overline{u})$ *satisfying all the conditions* (1.4)–(1.9) *and* (1.12)–(1.13)*, with the exception of the condition in* (1.8) *that* $u \in L^1((0,T);W^{1,\infty}(\mathbb{R}^n))$*, provided that the system conditions* (H) *hold, that* $(\overline{\rho}_0,\overline{u}_0)$ *satisfies* (1.19)*, and that* $\inf\overline{\rho}_0 > 0$ *and* $\overline{u}_0 \in H^s(\mathbb{R}^n)$*, where* $s > 0$ *for* $n = 2$ *and* $s > 1/2$ *for* $n = 3$. Recall that functions in H^s may be discontinuous across hypersurfaces of $\mathbb{R}^n$ for $s < 1/2$ but not for $s > 1/2$. The above restrictions on $\overline{u}_0$ therefore allow locally Riemann-like initial data in $\mathbb{R}^2$, but codimension-one singularities only in ρ_0 in $\mathbb{R}^3$.

The assumption in (1.8) that ∇u, $\nabla\overline{u} \in L^1((0,T);L^\infty(\mathbb{R}^n))$ is the most restrictive condition on the list. In order to examine it in more detail we decompose the velocity field u by splitting off its most singular part: recalling the definitions in (1.3) of F and ω, we write

$$\begin{aligned}\Delta u^j &= u^j_{x_kx_k} = u^k_{x_kx_j} + (u^j_{x_k} - u^k_{x_j})_{x_k}\\ &= \left[(\mu+\lambda)^{-1}F_{x_j} + \omega^{j,k}_{x_k}\right] + (\mu+\lambda)^{-1}(P-\widetilde{P})_{x_j}\\ &\equiv \Delta u^j_{F,\omega} + \Delta u^j_P,\end{aligned}\tag{1.22}$$

so that $u = u_{F,\omega} + u_P$. We can easily show that $\int_0^T\|\nabla u_{F,\omega}(\cdot,t)\|_{L^\infty}dt < \infty$ if (1.20) holds for the same values of s as above, which again is a consequence of the assumptions (1.19), $\inf\rho_0 > 0$, and $u_0 \in H^s$. To see this, we apply standard elliptic theory in the definitions (1.22) and (1.3) of $u_{F,\omega}$ and F, ignoring the contribution of ω, which turns out to be the same as that of F, and of f, which is a lower-order term, to obtain that, in $\mathbb{R}^3$, for some $\gamma > 3$ and some $\varepsilon > 0$ determined by γ,

$$\|\nabla u_{F,\omega}(\cdot,t)\|_{L^\infty} \le C\|D_x^2u_{F,\omega}(\cdot,t)\|_{L^\gamma} \le C\|\nabla F(\cdot,t)\|_{L^\gamma}$$
$$\le C\|\dot{u}(\cdot,t)\|_{L^\gamma} \le C\|\dot{u}(\cdot,t)\|_{L^2}^{(1-\varepsilon)/2}\|\nabla\dot{u}(\cdot,t)\|_{L^2}^{(1+\varepsilon)/2},$$

so that

$$\int_0^T\|\nabla u_{F,\omega}(\cdot,t)\|_{L^\infty}dt \le C\int_0^T t^\beta\left(t^{1-s}\int|\dot{u}|^2dx\right)^{(1-\varepsilon)/4}\left(t^\sigma\int|\nabla\dot{u}|^2dx\right)^{(1+\varepsilon)/4}dt$$
$$\le C\left(\int_0^T t^{2\beta}dt\right)^{1/2}$$

by (1.20), where $4\beta = (s-1)(1-\varepsilon) - \sigma$. The above integral is therefore finite if $2\beta > -1$ for some $\varepsilon > 0$ and hence some $\gamma > 3$, and this holds if $s > 1/2$. A similar result holds for $n = 2$ with $s > 0$. Thus for the solution constructed in [7], $\int_0^T\|\nabla u_{F,\omega}(\cdot,t)\|_{L^\infty}dt$ is finite if (1.16) holds, $\inf\rho_0 > 0$, and $u_0 \in H^s$, where $s > 1/2$ for $n = 3$ and $s > 0$ for $n = 2$.

There remains the condition that $\int_0^T \|\nabla u_P(\cdot,t)\|_{L^\infty} dt < \infty$, that is, that

$$\int_0^T \|\nabla \Gamma_{x_j} * (P(\rho(\cdot,t)) - \widetilde{P})\|_{L^\infty} dt < \infty,$$

where Γ is the fundamental solution of the Laplace operator on $\mathbb{R}^n$. Now, it is a standard result of Calderon–Zygmund theory that $\nabla\Gamma_{x_j}*$ is a bounded linear transformation from L^p to L^p for $p \in (1,\infty)$ but not for $p = 1$ or ∞ (see [5, pp. 230–235], for example). It is also true that if $P(\rho(\cdot,t)) \in L^\infty$, then $u_P^j(\cdot,t) = (\mu+\lambda)^{-1}\Gamma_{x_j}*(P(\rho(\cdot,t))-\widetilde{P})$ is log-Lipschitz, if not Lipschitz. This is sufficient to guarantee the existence and uniqueness of the particle trajectories $X(y,t,s)$ in (1.17). It does not ensure, however, that X is Lipschitz with respect to y, and this is essential for our analysis. Of course, we could obtain that $\nabla u_P(\cdot,t) \in L^\infty$ by imposing the condition that $P(\rho(\cdot,t)) \in C^\alpha$, but this would exclude solutions with codimension-one singularities and is altogether too strong. On the other hand, a large class of solutions is constructed in Hoff [6] in which densities are *piecewise* C^α and, as a consequence, $\nabla u_P(\cdot,t) \in L^\infty$. The results of the present paper therefore do apply to this class, which includes solutions with Riemann-like initial data.

We make one final observation concerning (1.8) for the isothermal case $P = K\rho$, namely that for the solutions constructed in Hoff [6, 7], the condition $\nabla u_P(\cdot,t) \in L^\infty$ can be shown to hold if in addition to (1.19) it is assumed that $\inf \rho_0 > 0$ and there is a constant C such that

$$\left| \int \rho_0(x)\varphi(x)dx \right| \leq C\|\varphi\|_{L^1_w} \tag{1.23}$$

for integrable functions φ. Here $\|\varphi\|_{L^1_w}$ is the weak-L^1 norm of φ, which is the smallest number $M \geq 0$ such that $k|\{x : |\varphi(x)| \geq k\}| \leq M$ for all $k > 0$. L^1_w contains L^1 but is not a Banach space, and $\|\cdot\|_{L^1_w}$ is not a norm. However, $A \equiv \Gamma_{x_j x_k}* : L^1 \to L^1_w$ is "bounded," and the adjoint of the mass equation in (1.1) is easily seen to preserve the class L^1_w. That is, if φ is the solution of the initial-value problem

$$\begin{cases} \varphi_t + \nabla\varphi \cdot u = 0, \\ \varphi(x,\bar{t}) = \Phi(x), \end{cases}$$

then as a simple consequence of the conservation of mass, $\|\varphi(\cdot,t)\|_{L^1_w} \leq C\|\Phi\|_{L^1_w}$ for a constant C determined by the assumed pointwise bounds on ρ. Taking $\psi \in L^1$ and $\Phi = A\psi \in L^1_w$, we then obtain that

$$\begin{aligned} \left| \int (A\rho)(x,\bar{t})\psi(x)dx \right| &= \left| \int \rho(x,\bar{t})\Phi(x)dx \right| \\ &= \left| \int \rho_0(x)\varphi(x,0)dx \right| \leq C\|\varphi(\cdot,0)\|_{L^1_w} \\ &\leq C\|\Phi\|_{L^1_w} \leq C\|\psi\|_{L^1} \end{aligned}$$

by (1.23), so that $\|(A\rho)(\cdot,t)\|_{L^\infty} \leq C$, as required. While fairly simple and appealing, this argument is apparently of little practical importance because of the extreme difficulty in verifying the condition (1.23).

We summarize the above considerations as follows: solutions (ρ,u) and $(\overline{\rho},\overline{u})$ satisfying all the conditions (1.4)–(1.14) are known to exist under the following hypotheses:

- The conditions (H) described in [7] on the system parameters P, f, $\overline{f}$, μ, and λ hold;
- Both (ρ_0, u_0) and $(\overline{\rho}_0, \overline{u}_0)$ satisfy (1.19);
- $\inf \rho_0 > 0$ and $\inf \overline{\rho}_0 > 0$;
- The initial densities and forces satisfy (1.13) and (1.14);
- $u_0, \overline{u}_0 \in H^s(\mathbb{R}^n)$ where $s > 0$ for $n = 2$ and $s > 1/2$ for $n = 3$;
- $\nabla u_P, \nabla \overline{u}_P \in L^1([0,T]; L^\infty(\mathbb{R}^n))$, where u_P and $\overline{u}_P$ are as defined in (1.22).

The first five of these conditions may be verified a priori and, again, the last is known to hold for the large class of solutions exhibiting codimension-one singularities constructed in Hoff [6].

There is now a large and growing literature on various mathematical aspects of the Navier–Stokes equations of compressible fluid flow. See the books of Feireisl [2] and Lions [10] for general discussions and more complete bibliographies relating to the existence of weak solutions, including for the nonbarotropic case in which temperature appears and an energy balance equation is appended to (1.1). See also Xin [11] for a result concerning the blowup of smooth solutions of (1.1) when the initial density has compact support.

2. Lagrangean structure. In this section we discuss properties of the particle trajectories $X(y,t,s)$ required for the proof of the theorem. We shall make repeated use both in this section and in the next of Rademacher's theorem and some of its consequences (see Ziemer [12, pp. 49–53]): a Lipschitz function has a differential almost everwhere; the composition of a function in $W^{1,p}$ with a bi-Lipschitz map is again in $W^{1,p}$, and the chain rule holds; and the usual change of variables formula holds for integrable functions composed with univalent Lipschitz mappings.

LEMMA 2.1. *Let u satisfy* (1.8). *Then there is a unique function $X \in C(\mathbb{R}^n \times [0,T]^2)$ satisfying*

$$\begin{cases} \dfrac{\partial X}{\partial t}(y,t,s) = u(X(y,t,s),t) \\ X(y,s,s) = y. \end{cases} \tag{2.1}$$

$X(\cdot,t,s)$ is Lipschitz on $\mathbb{R}^n$ for $(t,s) \in [0,T]^2$, and there is a constant C such that

$$\left\| \frac{\partial X}{\partial y}(\cdot,t,s) \right\|_{L^\infty} \le C, \qquad (t,s) \in [0,T]^2.$$

Also, given $\tau > 0$ there is a constant $C = C(\tau)$ such that

$$\left| \frac{\partial X}{\partial t}(y,t,s) \right| \le C(\tau)$$

for all $(y,s) \in \mathbb{R}^n \times [0,T]$ and almost all $t \in [\tau,T]$, and

$$\left| \frac{\partial X}{\partial s}(y,t,s) \right| \le C(\tau)$$

for all $(y,t) \in \mathbb{R}^n \times [0,T]$ and almost all $s \in [\tau,T]$.

Proof. The proof is elementary and thus is omitted. We do note, however, the importance of the assumption that $u \in L^1((0,T); W^{1,\infty})$ in the derivation of the first bound above:

$$\begin{aligned} \frac{1}{2}\frac{\partial}{\partial t}\left|\nabla_y X(y,t,s)\right|^2 &= \nabla_y X(y,t,s) \cdot \nabla_y u(X(y,t,s),t) \\ &\le \|\nabla u(\cdot,t)\|_{L^\infty} \left|\nabla_y X(y,t,s)\right|^2. \qquad \square \end{aligned}$$

In the following lemma we compute the Jacobian determinant of $X(\cdot, t, s)$.

LEMMA 2.2. *Let u be as in (1.8) and ρ as in (1.4). Assume that ρu is locally integrable on $\mathbb{R}^n \times [0,T]$ and that the weak form (1.6) of the mass equation holds. If E_{t_0} is a bounded measurable set in $\mathbb{R}^n$ and if $E_t = \{X(y,t,t_0) : y \in E_{t_0}\}$, then*

$$\int_{E_t} \rho(x,t)dx = \int_{E_{t_0}} \rho(y,t_0)dy. \tag{2.2}$$

Also,

$$\rho(X(\cdot,t,t_0),t)\,|\det \nabla_y X(\cdot,t,t_0)| = \rho(\cdot,t_0) \tag{2.3}$$

a.e. on $\mathbb{R}^n$.

Proof. The conservation of mass (2.2) is proved by applying (1.6) to the solution $\varphi^{\eta,\delta}$ of the adjoint system

$$\begin{cases} \varphi_t^{\eta,\delta} + \nabla\varphi^{\eta,\delta}\cdot u^\delta = 0, \\ \varphi^{\eta,\delta}(\cdot,t_0) = \chi^\eta_{E_{t_0}}, \end{cases}$$

where u^δ and $\chi^\eta_{E_{t_0}}$ are smooth approximations to u and to the characteristic function of E_{t_0}, and then letting first δ, then η tend to zero. The bound $\|\nabla\varphi^{\eta,\delta}\|_{L^\infty} \le C(\eta)$, which follows from the condition $\int_0^T \|\nabla u(\cdot,t)\|_{L^\infty}dt < \infty$ as in the proof of Lemma 2.1, ensures that the approximation errors disappear in the limit. This proves (2.2). Then as a consequence,

$$\begin{aligned} \int_{E_{t_0}} \rho(y,t_0)dy &= \int_{E_t} \rho(x,t)dx \\ &= \int_{E_{t_0}} \rho(X(y,t,t_0),t)\,|\det \nabla_y X(y,t,t_0)|\,dy. \end{aligned}$$

This holds for all bounded measureable sets E_{t_0}; the integrands therefore agree a.e. $\square$

In the following lemma we list properties of the mapping S defined in (1.18) that are required for the proof of the theorem.

LEMMA 2.3. *Let both (ρ,u) and $(\overline{\rho},\overline{u})$ satisfy the hypotheses of Lemmas 2.1 and 2.2, let X and $\overline{X}$ be as in Lemma 2.1, and define*

$$\begin{cases} S(x,t) = \overline{X}(X(x,0,t),t,0), \\ S^{-1}(x,t) = X(\overline{X}(x,0,t),t,0). \end{cases} \tag{2.4}$$

Then

(a) *$S^{\pm1}$ is continuous on $\mathbb{R}^n \times [0,T]$ and Lipschitz continuous on $\mathbb{R}^n \times [\tau,T]$ for all $\tau > 0$, and there is a constant C such that*

$$\|\nabla S^{\pm1}(\cdot,t)\|_{L^\infty} \le C, \qquad t \in [0,T];$$

(b) *$(S_t + \nabla S u)(x,t) = \overline{u}(S(x,t),t)$ a.e. in $\mathbb{R}^n \times (0,T)$;*

(c) *$\overline{\rho}(S(x,t),t)\rho_0(X(x,0,t)) \det \nabla S(x,t) = \rho(x,t)\overline{\rho}_0(X(x,0,t))$ a.e. in $\mathbb{R}^n \times (0,T)$;*

(d) *if u, $\overline{u} \in L^2(\mathbb{R}^n \times (0,T))$, then*

$$\int |x - S(x,t)|^2 dx \le Ct\int_0^t \int |u(x,s) - \overline{u}(S(x,s),s)|^2 dxds,$$

$$\int |x - S^{-1}(x,t)|^2 dx \le Ct\int_0^t \int |u(S^{-1}(x,s),s) - \overline{u}(x,s)|^2 dxds.$$

The constants C in (a) *and* (d) *depend on upper bounds for u and $\overline{u}$ in $L^1((0,T);W^{1,\infty}(\mathbb{R}^n))$.*

Proof. (a) and (b) follow directly from Lemmas 2.1 and 2.2. To prove (c) we differentiate the relation $S(X(y,t,0),t) = \overline{X}(y,t,0)$ to obtain

$$\det\nabla S(X(y,t,0),t)\det\nabla_y X(y,t,0) = \det\nabla_y \bar{X}(y,t,0)$$

a.e. Multiplying by $\rho(X(y,t,0),t)\bar{\rho}(\bar{X}(y,t,0),t)$ and applying (2.3) with $t_0 = 0$, we then get

$$\bar{\rho}(\bar{X}(y,t,0),t)\rho_0(y)\det\nabla S(X(y,t,0),t) = \rho(X(y,t,0))\bar{\rho}_0(y)\,.$$

We then set $x = X(y,t,0)$ so that $y = X(x,0,t)$ and $S(x,t) = \bar{X}(y,t,0)$ to obtain (c). To prove (d) we compute

$$\begin{aligned}
\int |x - S(x,t)|^2 dx &\le C\int |X(y,t,0) - \overline{X}(y,t,0)|^2 dy \\
&= C\int \left|\int_0^t [u(X(y,s,0),s) - \overline{u}(\overline{X}(y,s,0),s)]ds\right|^2 dy \\
&\le Ct\int_0^t \int |u(x,s) - \overline{u}(S(x,s),s)|^2 dxds.
\end{aligned}$$

Notice that the Jacobian determinants occurring in the above changes of variables are bounded by (a). The proof for S^{-1} is similar. □

3. Proof of the theorem. We now fix two solutions (ρ,u,f) and $(\overline{\rho},\overline{u},\overline{f})$ as described in the theorem of section 1 and define the mappings X, $\overline{X}$, and S as in (2.1) and (2.4). We shall make frequent use of the conclusions of Lemma 2.3 and the properties of Lipschitz mappings described at the beginning of section 2.

We begin by deriving the weak form of the equation satisfied by $u - \overline{u}\circ S$, where again we abuse notation slightly and write $\overline{u}(S(x,t),t) = (\overline{u}\circ S)(x,t)$. Let $\psi : \mathbb{R}^n \times [0,T] \to \mathbb{R}^n$ be a test function satisfying the conditions in (1.7) and which additionally is in $(H^2\cap C^2)(\mathbb{R}^n\times[0,T])$. Then

$$\begin{aligned}
&-\int_{\mathbb{R}^n} \rho_0(x)u_0(x)\cdot\psi(x,0)dx \\
&\qquad = \int_0^T\int_{\mathbb{R}^n}\Big[\rho u\cdot(\psi_t + \nabla\psi u) + (P(\rho) - \widetilde{P})\operatorname{div}\psi \\
&\qquad\qquad - \mu\nabla u^j\cdot\nabla\psi^j - \lambda\operatorname{div}u\operatorname{div}\psi + \rho f\cdot\psi\Big]dxdt.
\end{aligned}\tag{3.1}$$

Now let $\overline{\psi} \equiv \psi\circ S^{-1}$ and note that $\overline{\psi}$ also satisfies the conditions in (1.7). Thus

$$\begin{aligned}
&-\int_{\mathbb{R}^n} \overline{\rho}_0(x)\overline{u}_0(x)\cdot\overline{\psi}(x,0)dx \\
&\qquad = \int_0^T\int_{\mathbb{R}^n}\Big[\overline{\rho u}\cdot(\overline{\psi}_t + \nabla\overline{\psi}\,\overline{u}) + (P(\overline{\rho}) - \widetilde{P})\operatorname{div}\overline{\psi} \\
&\qquad\qquad - \mu\nabla\overline{u}^j\cdot\nabla\overline{\psi}^j - \lambda\operatorname{div}\overline{u}\operatorname{div}\overline{\psi} + \overline{\rho}\overline{f}\cdot\overline{\psi}\Big]dxdt.
\end{aligned}$$

We make the change of variables $x = S(y,t)$ in the first and last terms on the right here, noting that, by Lemma 2.3,

$$(\overline{\psi}_t + \nabla\overline{\psi}\,\overline{u}) \circ S = \psi_t + \nabla\psi u$$

and

$$(\overline{\rho} \circ S)|\det \nabla S| = A_0\rho,$$

where

$$A_0(x,t) = \overline{\rho}_0(X(x,0,t))/\rho_0(X(x,0,t)) \tag{3.2}$$

(recall that ρ_0 is strictly positive by (1.10)). The result is that

$$\begin{aligned}
&-\int_{\mathbb{R}^n} \overline{\rho}_0(x)\overline{u}_0(x)\cdot\overline{\psi}(x,0)dx \\
&= \int_0^T\int_{\mathbb{R}^n}\Big[A_0\rho(\overline{u}\circ S)\cdot(\psi_t+\nabla\psi u) + (P(\overline{\rho})-\widetilde{P})\mathrm{div}\overline{\psi} \\
&\qquad -\mu\nabla\overline{u}^j\cdot\nabla\overline{\psi}^j - \lambda\,\mathrm{div}\,\overline{u}\,\mathrm{div}\,\overline{\psi} + A_0\rho(\overline{f}\circ S)\cdot\psi\Big]\,dxdt.
\end{aligned} \tag{3.3}$$

Next we subtract (3.3) from (3.1), letting

$$z = u - \overline{u}\circ S$$

and noting that $\overline{\psi}(\cdot,0) = \psi(\cdot,0)$:

$$\begin{aligned}
&\int_{\mathbb{R}^n}(\overline{\rho}_0\overline{u}_0 - \rho_0 u_0)\cdot\psi(\cdot,0)dx \\
&= \int_0^T\int_{\mathbb{R}^n}\big[\rho z\cdot(\psi_t+\nabla\psi u) + (1-A_0)\rho(\overline{u}\circ S)(\psi_t+\nabla\psi u) \\
&\qquad + (P(\rho)-\widetilde{P})\mathrm{div}\,\psi - (P(\overline{\rho})-\widetilde{P})\mathrm{div}\,\overline{\psi} \\
&\qquad - \mu(\nabla u^j\cdot\nabla\psi^j - \nabla\overline{u}^j\cdot\nabla\overline{\psi}^j) \\
&\qquad - \lambda(\mathrm{div}\,u\,\mathrm{div}\,\psi - \mathrm{div}\,\overline{u}\,\mathrm{div}\,\overline{\psi}) \\
&\qquad + \rho(f-\overline{f}\circ S)\cdot\psi + (1-A_0)\rho(\overline{f}\circ S)\cdot\psi\big]\,dxdt.
\end{aligned} \tag{3.4}$$

We rewrite three of the terms on the right here as follows (recall the definitions of $\overline{F}$ and $\overline{\omega}$ in (1.3)):

$$\begin{aligned}
&\iint\Big[(\widetilde{P}-P(\overline{\rho}))\mathrm{div}\overline{\psi} + \mu\nabla\overline{u}^j\cdot\nabla\overline{\psi}^j + \lambda\,\mathrm{div}\,\overline{u}\,\mathrm{div}\,\overline{\psi}\Big] \\
&\quad = \iint\Big[((\mu+\lambda)\mathrm{div}\,\overline{u} - P(\overline{\rho}) + \widetilde{P})\mathrm{div}\,\overline{\psi} + \mu(\overline{u}^j_{x_k} - \overline{u}^k_{x_j})\overline{\psi}^j_{x_k}\Big] \\
&\quad = -\iint(\nabla\overline{F}\cdot\overline{\psi} + \mu\,\overline{\omega}^{j,k}_{x_k}\overline{\psi}^j) \\
&\quad = -\iint(\nabla\overline{F}\cdot\psi + \mu\,\overline{\omega}^{j,k}_{x_k}\psi^j) + E_1 \\
&\quad = \iint\Big[(\widetilde{P}-P(\overline{\rho}))\mathrm{div}\,\psi + \mu\nabla\overline{u}^j\cdot\nabla\psi^j + \lambda\,\mathrm{div}\,\overline{u}\,\mathrm{div}\,\psi\Big] + E_1,
\end{aligned}$$

where

$$E_1 = \int_0^T \int \left[\nabla \overline{F} \cdot (\psi - \psi \circ S^{-1}) + \mu \overline{\omega}^{j,k}_{x_k} (\psi^j - \psi^j \circ S^{-1})\right] dxdt. \tag{3.5}$$

Defining E_2 by

$$E_2 = \iint \left[\rho(f - \overline{f} \circ S) \cdot \psi + (1 - A_0)\rho(\overline{f} \circ S) \cdot \psi\right],$$

we then have from (3.4) that

$$\begin{aligned}
&\int_{\mathbb{R}^n} (\overline{\rho}_0 \overline{u}_0 - \rho_0 u_0) \cdot \psi(\cdot, 0) dx \\
&= \int_0^T \int_{\mathbb{R}^n} [\rho z \cdot (\psi_t + \nabla \psi u) + (1 - A_0)\rho(\overline{u} \circ S)(\psi_t + \nabla \psi u) \\
&\qquad + (P(\rho) - P(\overline{\rho}))\mathrm{div}\, \psi \\
&\qquad - \mu(\nabla u^j - \nabla \overline{u}^j) \cdot \nabla \psi^j - \lambda(\mathrm{div}\, u - \mathrm{div}\, \overline{u})\mathrm{div}\, \psi\big] \, dxdt \\
&\quad + E_1 + E_2.
\end{aligned} \tag{3.6}$$

We write

$$\begin{aligned}
&\iint \rho z \cdot [(\psi_t + \nabla \psi u) - \mu(\nabla u^j - \nabla \overline{u}^j) \cdot \nabla \psi^j - \lambda(\mathrm{div}\, u - \mathrm{div}\, \overline{u})\mathrm{div}\, \psi] \\
&= \iint z \cdot [\rho(\psi_t + \nabla \psi u) + \mu \Delta \psi + \lambda \nabla \mathrm{div} \psi] + E_3 ,
\end{aligned}$$

where

$$E_3 = \iint (\overline{u} \circ S - \overline{u}) \cdot (\mu \Delta \psi + \lambda \nabla \mathrm{div}\, \psi).$$

We then have from (3.6) that

$$\begin{aligned}
&\int_{\mathbb{R}^n} (\overline{\rho}_0 \overline{u}_0 - \rho_0 u_0) \cdot \psi(\cdot, 0) dx \\
&= \int_0^T \int_{\mathbb{R}^n} z \cdot [\rho(\psi_t + \nabla \psi u) + \mu \Delta \psi + \lambda \nabla \mathrm{div}\, \psi] \, dxdt \\
&\quad + \int_0^T \int_{\mathbb{R}^n} [(1 - A_0)\rho(\overline{u} \circ S)(\psi_t + \nabla \psi u) + (P(\rho) - P(\overline{\rho}))\mathrm{div}\, \psi] \, dxdt \\
&\quad + E_1 + E_2 + E_3.
\end{aligned} \tag{3.7}$$

Now extend ρ and u to be constant in t outside $[0, T]$, let ρ^δ and u^δ be corresponding smooth approximations obtained by mollifying in both x and t, and let $\psi^\delta : \mathbb{R}^n \times [0, T] \to \mathbb{R}^n$ be the solution of the adjoint system

$$\begin{cases} \rho^\delta(\psi^\delta_t + \nabla \psi^\delta u^\delta) + \mu \Delta \psi^\delta + \lambda \nabla \mathrm{div}\, \psi^\delta = G, \\ \psi^\delta(\cdot, T) = 0 \end{cases} \tag{3.8}$$

for a given $G \in H^\infty(\mathbb{R}^n \times [0, T])$. The existence of the solution ψ^δ is a straightforward exercise; bounds for various of its norms are given in the following lemma.

LEMMA 3.1. *There is a constant C as described in the theorem of section 1 and independent of δ and G such that the solution ψ^δ of (3.8) satisfies*

$$\sup_{0\le t\le T}\int_{\mathbb{R}^n}[|\psi^\delta(x,t)|^2+|\nabla\psi^\delta(x,t)|^2]dx+\int_0^T\int_{\mathbb{R}^n}\left[|\psi_t^\delta+\nabla\psi^\delta u^\delta|^2+|D_x^2\psi^\delta|^2\right]dxdt \le C\int_0^T\int_{\mathbb{R}^n}|G|^2dxdt, \tag{3.9}$$

and there is a constant $C=C(G)$ depending on higher derivatives of G but independent of δ, such that

$$\sup_{0\le t\le T}\|\nabla\psi^\delta(\cdot,t)\|_{L^\infty}+\int_0^T\int_{\mathbb{R}^n}|\psi^\delta|^r dxdt\le C(G), \tag{3.10}$$

where r is as in (1.11).

The proof, which is elementary, is deferred to the end of this section.

We now take $\psi=\psi^\delta$ in (3.7) to obtain finally that

$$\begin{aligned}&\int_{\mathbb{R}^n}(\overline{\rho}_0\overline{u}_0-\rho_0u_0)\cdot\psi^\delta(\cdot,0)dx\\ &=\int_0^T\int_{\mathbb{R}^n}z\cdot G\,dxdt+E_1+E_2+E_3+E_4+E_5+E_6\,,\end{aligned} \tag{3.11}$$

where

$$\begin{aligned}E_4&=\iint z\cdot\left[(\rho-\rho^\delta)\psi_t^\delta+\nabla\psi^\delta(\rho u-\rho^\delta u^\delta)\right],\\ E_5&=\iint(1-A_0)\rho(\overline{u}\circ S)\cdot(\psi_t^\delta+\nabla\psi^\delta u),\end{aligned}$$

and

$$E_6=\iint(P(\rho)-P(\overline{\rho}))\operatorname{div}\psi^\delta.$$

We now apply the bounds of Lemma 3.1 to estimate each of the terms $E_1,\dots,E_6$ and the term on the left-hand side of (3.11), and then take the limit as $\delta\to0$. First, it is easy to see that $E_4\to0$ as $\delta\to0$ and that the term on the left-hand side of (3.11) is bounded by

$$\|\rho_0u_0-\overline{\rho}_0\overline{u}_0\|_{L^2}\left(\iint|G|^2\right)^{1/2}.$$

Next we note that, from the definition (3.2) of A_0,

$$\sup_{0\le t\le T}\|1-A_0(\cdot,t)\|_{L^{2p}}\le C\|\rho_0-\overline{\rho}_0\|_{L^{2p}}, \tag{3.12}$$

so that, if q is the Hölder conjugate of p as in (1.13)–(1.14), then the force terms E_2

are bounded by

$$
\begin{aligned}
&C\Bigg[\left(\iint |f-\overline{f}\circ S|^2\right)^{1/2}\left(\iint |\psi^\delta|^2\right)^{1/2} \\
&\qquad + \sup_t \|\rho_0-\overline{\rho}_0\|_{L^{2p}} \sup_t \left(\int |\psi^\delta|^2 dx\right)^{1/2} \int_0^T \|\overline{f}(\cdot,t)\|_{L^{2q}} dt\Bigg] \\
&\quad \le C\left[\left(\iint |f-\overline{f}\circ S|^2\right)^{1/2} + \|\rho_0-\overline{\rho}_0\|_{L^{2p}}\right]\left(\iint |G|^2\right)^{1/2}
\end{aligned}
$$

by (3.9).

Next we bound the term E_1, disregarding the contribution of $\overline{\omega}$, which is essentially identical to that of $\overline{F}$. The argument requires that we subdivide time intervals into smaller subintervals. Thus let J be a large positive integer and for $t \in [0,T]$ define $t_j = jt/J$, $j = 0,\ldots,J$. Then

$$
\begin{aligned}
|E_1| &\le \int_0^T \int_{\mathbb{R}^n} |\nabla\overline{F}(x,t)|\,|\psi^\delta(x,t) - \psi^\delta(S^{-1}(x,t),t)|dxdt \\
&\le \sum_j \int_0^T \int_{\mathbb{R}^n} |\nabla\overline{F}(x,t)|\,|\psi^\delta(S^{-1}(x,t_{j+1}),t) - \psi^\delta(S^{-1}(x,t_j),t)|dxdt \\
&\le \sum_j \int_0^T \|\nabla\overline{F}(\cdot,t)\|_{L^4} \|S^{-1}(\cdot,t_{j+1}) - S^{-1}(\cdot,t_j)\|_{L^2} \\
&\qquad \times \left(\int_0^1 \int_{\mathbb{R}^n} |\nabla\psi^\delta(\theta S^{-1}(x,t_{j+1}) + (1-\theta)S^{-1}(x,t_j),t)|^4 dxd\theta\right)^{1/4} dt.
\end{aligned}
$$

Now, the mapping $x \mapsto S^{-1}(x,t_j)$ is bi-Lipschitz on $\mathbb{R}^n$, so that the inner double integral here is bounded by

$$
C\int_0^1 \int_{\mathbb{R}^n} |\nabla\psi^\delta(x+\theta T(x),t)|^4 dxd\theta,
$$

where $T(x) = S^{-1}(S(x,t_j),t_{j+1}) - x$. It is easy to check from the definitions (2.4) and the condition (1.8) that $\|\nabla T\|_{L^\infty}$ is arbitrarily small if t_{j+1} is sufficiently close to t_j, i.e., if J is sufficiently large. The mapping $x \mapsto x + \theta T(x)$ is then one-to-one (and bi-Lipschitz) on $\mathbb{R}^n$, so that the change of variables $y = x + \theta T(x)$ is valid. The inner double integral above is thus bounded by $C\int |\nabla\psi^\delta(x,t)|^4 dx$, and therefore

$$
|E_1| \le C\left(\iint |z|^2 dxdt\right)^{1/2} \int_0^T t^{1/2} \|\nabla\overline{F}(\cdot,t)\|_{L^4} \|\nabla\psi^\delta(\cdot,t)\|_{L^4} dt
$$

by Lemma 2.3(d). Finally we apply the standard Sobolev embedding $H^1(\mathbb{R}^3) \to$

$L^4(\mathbb{R}^3)$ for $n = 3$ to obtain

$$
\begin{aligned}
|E_1| &\leq C \left(\iint |z|^2 dxdt \right)^{1/2} \int_0^T t^{1/2} \|\nabla \overline{F}(\cdot,t)\|_{L^4} \left(\int |\nabla \psi^\delta|^2 dx \right)^{1/8} \left(\int |D_x^2 \psi^\delta|^2 \right)^{3/8} dt \\
&\leq C \left(\iint |z|^2 dxdt \right)^{1/2} \left(\iint |G|^2 dxdt \right)^{1/8} \left(\int t^{4/5} \|\nabla \overline{F}(\cdot,t)\|_{L^4}^{8/5} dt \right)^{5/8} \\
&\qquad \times \left(\iint |D_x^2 \psi^\delta|^2 dxdt \right)^{3/8} \\
&\leq \left(\iint |z|^2 dxdt \right)^{1/2} \left(\iint |G|^2 dxdt \right)^{1/2} \left(\int_0^T \beta(t) dt \right)^{5/8},
\end{aligned}
$$

where $\beta(t) = Ct^{4/5} \|\nabla \overline{F}(\cdot,t)\|_{L^4}^{8/5}$, which is in $L^1((0,T))$ by the hypothesis (1.12). A similar bound holds for $n = 2$. More generally, we can define $E_1(t_1,t_2)$ to be the integral in (3.5), but with $[0,T]$ replaced by $[t_1,t_2] \subseteq [0,T]$. It is then clear from the above analysis that

$$
|E_1(t_1,t_2)| \leq \left(\int_0^{t_2} \int |z|^2 dxdt \right)^{1/2} \left(\int_0^{t_2} \int |G|^2 dxdt \right)^{1/2} \left(\int_{t_1}^{t_2} \beta(t) dt \right)^s \tag{3.13}
$$

for the same β and s, where β is determined solely by the properties of (ρ, u, f) and $(\overline{\rho}, \overline{u}, \overline{f})$ and s is determined by n. In particular, C, β, and s are independent of G, t_1, and t_2.

The term E_3 is bounded in almost the same way as E_1 (the hypothesis on the second term in (1.12) is required here), and the bound for E_5 is trivial. For E_6 we first define the divided difference

$$
a(x,t) = \int_0^1 P'(\rho(x,t) + \theta(\overline{\rho}(x,t) - \rho(x,t))) d\theta
$$

as in (1.16). Then for $\varphi \in \mathcal{D}(\mathbb{R}^n)$ and at a fixed time t,

$$
\begin{aligned}
\left| \int_{\mathbb{R}^n} [P(\rho) - P(\overline{\rho})] \varphi dx \right| &= \left| \int_{\mathbb{R}^n} a(\rho - \overline{\rho}) \varphi dx \right| \\
&\leq \|(\rho - \overline{\rho})(\cdot,t)\|_{H^{-1}} \|a\varphi\|_{H^1} \\
&\leq C \|(\rho - \overline{\rho})(\cdot,t)\|_{H^{-1}} \left(\|\varphi\|_{H^1} + \|\varphi \nabla a\|_{L^2} \right).
\end{aligned}
$$

Now, for an ideal isothermal fluid, $P(\rho) = K\rho$, a is constant, and $\nabla a = 0$. For more general P we apply the hypothesis (1.16) for the case $n = 3$ that $\nabla a \in L^3$ and the embedding $H^1 \to L^6$ to obtain

$$
\|\varphi \nabla a\|_{L^2} \leq \|\varphi\|_{L^6} \|\nabla a\|_{L^3} \leq C \|\nabla \varphi\|_{L^2} \leq C \|\varphi\|_{H^1}.
$$

A similar result holds for $n = 2$. It follows that

$$
\|(P(\rho) - P(\overline{\rho}))(\cdot,t)\|_{H^{-1}} \leq C \|(\rho - \overline{\rho})(\cdot,t)\|_{H^{-1}}
$$

when either $P(\rho) = K\rho$ or for more general P when (1.16) holds. Thus in either case,

$$
\begin{aligned}
|E_6| &\leq \int_0^T \|P(\rho) - P(\overline{\rho})\|_{H^{-1}} \|\operatorname{div} \psi^\delta\|_{H^1} dt \\
&\leq CT^{1/2} \sup_{0 \leq t \leq T} \|(\rho - \overline{\rho})(\cdot,t)\|_{H^{-1}} \left(\iint |G|^2 dxdt \right)^{1/2}
\end{aligned} \tag{3.14}
$$

by Lemma 3.1. In order to complete the bound for E_6 we therefore need to obtain a bound for $\|\rho-\overline{\rho}\|_{H^{-1}}$. Again let $\varphi\in\mathcal{D}(\mathbb{R}^n)$ so that at a fixed time,

$$\begin{aligned}\left|\int(\rho-\overline{\rho})\varphi dx\right| &= \left|\int\left[\rho\varphi-(\overline{\rho}\circ S)(\varphi\circ S)|\det\nabla S|\right]dx\right|\\ &= \left|\int\rho(\varphi-A_0(\varphi\circ S))dx\right|\\ &\le \int\rho|\varphi-\varphi\circ S|dx+\int\rho|\varphi\circ S|\,|1-A_0|dx\end{aligned}$$

by Lemma 2.3(c) and the definition (3.2). An argument similar to that given above in the bound for E_1 shows that the first term on the right here is bounded by

$$Ct^{1/2}\left(\iint|z|^2dxdt\right)^{1/2}\left(\int|\nabla\varphi|^2dx\right)^{1/2},$$

and the second term is clearly bounded by $C\|\rho_0-\overline{\rho}_0\|_{L^2}\|\varphi\|_{L^2}$, by (3.12). We thus obtain that

$$\sup_{0\le t\le T}\|(\rho-\overline{\rho})(\cdot,t)\|_{H^{-1}}\le C\left[\|\rho_0-\overline{\rho}_0\|_{L^2}+T^{1/2}\left(\iint|z|^2dxdt\right)^{1/2}\right], \tag{3.15}$$

and, substituting into (3.14), that

$$|E_6|\le CT^{1/2}\left[\|\rho_0-\overline{\rho}_0\|_{L^2}+\left(\iint|z|^2dxdt\right)^{1/2}\right]\left(\iint|G|^2dxdt\right)^{1/2}.$$

Combining all the above bounds, we conclude from (3.11) that

$$\left|\int_0^T\int z\cdot Gdxdt\right|\le C\left[M_0\left(\int_0^T\int|G|^2dxdt\right)^{1/2}+E_1(0,T)\right], \tag{3.16}$$

where

$$\begin{aligned}M_0 &= \|\rho_0-\overline{\rho}_0\|_{L^2\cap L^{2p}}+\|\rho_0u_0-\overline{\rho}_0\overline{u}_0\|_{L^2}\\ &\quad+\left(\int_0^T\int|f-\overline{f}\circ S|^2dxdt\right)^{1/2}+T^{1/2}\left(\int_0^T\int|z|^2dxdt\right)^{1/2},\end{aligned}$$

E_1 is as in (3.13), and C is now fixed.

We note that (3.16) holds with T replaced by any time $\tau\le T$ with the same constant C, and we recall that β and s in (3.13) are fixed and are independent of G. To complete the bound for z we therefore choose a time $\tau>0$ so that $\left(\int_{t_1}^{t_2}\beta\right)^s\le 1/(2C)$ when $0\le t_2-t_1\le\tau$. Then applying (3.16) with T replaced by τ and bounding $E_1(0,\tau)$ as in (3.13), we obtain that

$$\left(\int_0^\tau\int|z|^2dxdt\right)^{1/2}\le 2CM_0$$

and as a consequence that

$$|E_1(0,\tau)| \le M_0 \left(\int_0^\tau \int |G|^2 dx dt \right)^{1/2}$$

for any G, again by (3.13). We now apply (3.16) with T replaced by 2τ:

$$\begin{aligned}
\left| \int_0^{2\tau} \int z \cdot G dx dt \right| &\le CM_0 \left(\int_0^{2\tau} \int |G|^2 dx dt \right)^{1/2} + C[E_1(0,\tau) + E_1(\tau, 2\tau)] \\
&\le 2CM_0 \left(\int_0^{2\tau} \int |G|^2 dx dt \right)^{1/2} \\
&\quad + C \left(\int_0^{2\tau} \int |z|^2 dx dt \right)^{1/2} \left(\int_0^{2\tau} \int |G|^2 dx dt \right)^{1/2} \cdot (2C)^{-1}
\end{aligned}$$

by (3.13), so that

$$\left(\int_0^{2\tau} \int |z|^2 dx dt \right)^{1/2} \le 4CM_0.$$

Since $\tau > 0$ is fixed, we can exhaust $[0,T]$ in a finite number of such steps to obtain that $\left(\int_0^T \int |z|^2 \right)^{1/2} \le CM_0$ for a new constant C. The term $T^{1/2} \left(\int_0^T \int |z|^2 \right)^{1/2}$ can then be eliminated from the definition of M_0 by a Gronwall-type argument. This together with (3.15) then proves the required bound (1.15), but with $u - \overline{u}$ replaced by $z = u - \overline{u} \circ S$. However, by Lemma 2.3(d) and the assumed bound for the second term in (1.12),

$$\begin{aligned}
\iint |\overline{u} - \overline{u} \circ S|^2 dx dt &\le \int \|\nabla \overline{u}(\cdot,t)\|_{L^\infty}^2 \int |x - S(x,t)|^2 dx dt \\
&\le C \left(\int t \|\nabla \overline{u}(\cdot,t)\|_{L^\infty}^2 dt \right) \left(\iint |z|^2 dx dt \right) \\
&\le C \iint |z|^2 dx dt,
\end{aligned}$$

which has just been shown to satisfy the bound in (1.15). This completes the proof of the theorem. $\square$

Proof of Lemma 3.1. We omit the superscripts. Multiplying the differential equation in (3.8) by ψ^j and adding the equation $\frac{1}{2}|\psi|^2(\rho_t + \operatorname{div}(\rho u)) = 0$, we obtain

$$\left(\tfrac{1}{2}\rho|\psi|^2 \right)_t + \operatorname{div}\left(\tfrac{1}{2}\rho|\psi|^2 u \right) + \psi \cdot (\mu \Delta \psi + \lambda \nabla \operatorname{div} \psi) = \psi \cdot G,$$

from which it follows easily that

$$\sup_{0 \le t \le T} \int |\psi(x,t)|^2 dx + \int_0^T \int_{\mathbb{R}^n} |\nabla \psi|^2 dx dt \le C \int_0^T \int_{\mathbb{R}^n} |G|^2 dx dt,$$

since $C^{-1} \le \rho \le C$. Next, writing $\dot{\psi} = \psi_t + \nabla \psi u$, we obtain from (3.8) that

$$\rho |\dot{\psi}|^2 + \left(\tfrac{1}{2}\mu |\Delta \psi|^2 + \tfrac{1}{2}\lambda (\operatorname{div} \psi)^2 \right) + \{\dots\}_x \le \psi \cdot G + C|\nabla \psi|^2 |\nabla u|,$$

so that

$$\sup_{0\le t\le T}\int|\nabla\psi(x,t)|^2dx+\int_0^T\int|\dot\psi|^2dxdt\le C\int_0^T\int|G|^2dxdt \tag{3.17}$$

since $\nabla u\in L^1((0,T);L^\infty(\mathbb{R}^n))$ by (1.8). It then follows from (3.8) that

$$\iint|\mu\Delta\psi+\lambda\operatorname{div}\psi|^2dxdt\le C\iint|G|^2dxdt$$

and therefore that

$$\iint|D_x^2\psi|^2dxdt\le C\iint|G|^2dxdt$$

because the operator $\mu\Delta+\lambda\nabla\mathrm{div}$ is elliptic. This proves (3.9). The a priori bound required for the proof of (3.10) is similar but slightly more technical: we take the convective derivative in (3.8) then multiply by $\operatorname{sgn}(\dot{\psi^j})|\dot{\psi^j}|^{s-1}$, where $s>n$, to obtain a bound for $\sup_t\int|\dot\psi|^sdx$. This gives a bound for $\sup_t\int|D_x^2\psi|^sdx$, as above, hence a bound for $\sup_t\|\nabla\psi(\cdot,t)\|_{L^\infty}$. $\square$

REFERENCES

[1] R. Danchin, *Local theory in critical spaces for compressible viscous and heat-conductive gases*, Comm. Partial Differential Equations, 26 (2001), pp. 1183–1233.
[2] E. Feireisl, *Dynamics of Viscous Compressible Fluids*, Oxford Lecture Ser. Math. Appl. 26, Oxford University Press, Oxford, UK, 2004.
[3] E. Feireisl, *On the motion of a viscous, compressible, and heat conducting fluid*, Indiana Univ. Math. J., 53 (2004), pp. 1707–1740.
[4] E. Feireisl, A. Novotny, and H. Petzeltova, *On the existence of globally defined weak solutions to the Navier–Stokes equations of compressible isentropic fluids*, J. Math. Fluid Mech., 3 (2001), pp. 358–392.
[5] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.
[6] D. Hoff, *Dynamics of singularity surfaces for compressible, viscous flows in two space dimensions*, Comm. Pure Appl. Math., 55 (2002), pp. 1365–1407.
[7] D. Hoff, *Compressible flow in a half-space with Navier boundary conditions*, J. Math. Fluid Mech., 7 (2005), pp. 315–338.
[8] D. Hoff and D. Serre, *The failure of continuous dependence on initial data for the Navier–Stokes equations of compressible flow*, SIAM J. Appl. Math., 51 (1991), pp. 887–898.
[9] D. Hoff and J. Smoller, *Non-formation of vacuum states for compressible Navier–Stokes equations*, Comm. Math. Phys., 216 (2001), pp. 255–276.
[10] P. L. Lions, *Mathematical Topics in Fluid Mechanics*, Vol. 2, Oxford Lecture Ser. Math. Appl. 10, Oxford University Press, New York, 1998.
[11] Z. Xin, *Blowup of smooth solutions to the compressible Navier–Stokes equation with compact density*, Comm. Pure Appl. Math., 51 (1998), pp. 229–240.
[12] W. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1761–1779

A NONLINEAR FOURTH-ORDER PARABOLIC EQUATION WITH NONHOMOGENEOUS BOUNDARY CONDITIONS*

MARIA PIA GUALDANI†, ANSGAR JÜNGEL†, AND GIUSEPPE TOSCANI‡

Abstract. A nonlinear fourth-order parabolic equation with nonhomogeneous Dirichlet–Neumann boundary conditions in one space dimension is analyzed. This equation appears, for instance, in quantum semiconductor modeling. The existence and uniqueness of strictly positive classical solutions to the stationary problem are shown. Furthermore, the existence of global nonnegative weak solutions to the transient problem is proved. The proof is based on an exponential transformation of variables and new "entropy" estimates. Moreover, it is proved by the entropy–entropy production method that the transient solution converges exponentially fast to its steady state in the L^1 norm as time goes to infinity, under the condition that the logarithm of the steady state is concave. Numerical examples show that this condition seems to be purely technical.

Key words. fourth-order parabolic equation, fourth-order elliptic equation, existence and uniqueness of nonnegative solutions, entropy–entropy production method, exponential decay in time

AMS subject classifications. 35K30, 35K35, 35B40

DOI. 10.1137/S0036141004444615

1. Introduction.

In recent years, the nonlinear fourth-order parabolic equation

$$u_t + (u(\log u)_{xx})_{xx} = 0, \quad u(\cdot, 0) = u_I \geq 0 \quad \text{in } \Omega,\ t > 0, \tag{1.1}$$

in a bounded interval $\Omega = (0, 1)$ with periodic or Dirichlet–Neumann boundary conditions or in the whole space $\Omega = \mathbb{R}$, has attracted the interest of many mathematicians since it possesses some interesting mathematical properties. For instance, the solutions are nonnegative, there are several Lyapunov functionals, and related logarithmic Sobolev inequalities can be derived [4, 10].

Equation (1.1) was first derived in the context of fluctuations of a stationary nonequilibrium interface [9]. It also appears as an approximation of the so-called quantum drift-diffusion model for semiconductors [1], which can be derived by a quantum moment method from a Wigner-BGK (Bhatnagar–Gross–Krook) equation [8]. More precisely, the quantum drift-diffusion model for the electron density u and the electron current density J reads as

$$u_t - J_x = 0, \quad J = -\frac{\varepsilon^2}{2} u \left(\frac{(\sqrt{u})_{xx}}{\sqrt{u}} \right)_x + T u_x + uE,$$

where ε is the scaled Planck constant, T the temperature, and $E = E(x, t)$ the electric field. Then (1.1) follows from this equation for $\varepsilon = 1$, zero temperature, and zero electric field since $u((\sqrt{u})_{xx}/\sqrt{u})_x = 2(u(\log u)_{xx})_x$.

*Received by the editors June 22, 2004; accepted for publication (in revised form) November 1, 2005; published electronically February 15, 2006. The authors acknowledge partial support from the Project "Hyperbolic and Kinetic Equations" of the European Union, grant HPRN-CT-2002-00282, and from the DAAD-Vigoni Program.

http://www.siam.org/journals/sima/37-6/44461.html

†Institut für Mathematik, Universität Mainz, Staudingerweg 9, 55099 Mainz, Germany (gualdani@mathematik.uni-mainz.de, juengel@mathematik.uni-mainz.de). These authors have been supported by the Deutsche Forschungsgemeinschaft, grants JU359/3 (Gerhard-Hess Award) and JU359/5 (Priority Program "Multi-Scale Problems").

‡Dipartimento di Matematica, Università di Pavia, Via Ferrara 1, 27100 Pavia, Italy (giuseppe.toscani@unipv.it).

The first analytical result for (1.1) has been presented in [4]; more precisely, the existence of local-in-time positive classical solutions with periodic boundary conditions has been proved. This result has been generalized to global nonnegative weak solutions in [10]. The existence of global weak periodic solutions in several space dimensions has been proved very recently employing Wasserstein space techniques [12].

In quantum semiconductor modeling, Dirichlet–Neumann boundary conditions of the type

$$u(0,t) = u(1,t) = 1, \quad u_x(0,t) = u_x(1,t) = 0, \quad t > 0, \tag{1.2}$$

have been employed to model resonant tunneling diodes in $\Omega = (0,1)$ [14]. Here, the function $u(x,t)$ signifies the (nonnegative) electron density in the semiconductor device. The existence of global weak solutions to (1.1)–(1.2) has been proved in [13].

The boundary conditions (1.2) simplify the analysis considerably. Indeed, one of the main ideas of the existence proof is to employ an exponential transformation of variables, $u = e^y$. In the new variable y, the boundary conditions are homogeneous. Thus, using, for instance, the test function y in the weak formulation of (1.1), no integrals with boundary data appear.

The boundary conditions (1.2) follow from physical considerations like the charge neutrality at the boundary contacts, i.e., $u - C = 0$ at $x = 0, 1$, where $C = C(x)$ models fixed background charges. Numerical results show that the Neumann boundary conditions for the density u should be nonhomogeneous for ultrasmall semiconductor devices (see section 4 in [16]). Moreover, when the values of the doping profile $C(x)$ are different at the contacts, the Dirichlet boundary conditions satisfy $u(0,t) \neq u(1,t)$. Therefore, we wish to study the more general *nonhomogeneous* boundary conditions

$$u(0,t) = u_0, \quad u(1,t) = u_1, \quad u_x(0,t) = w_0, \quad u_x(1,t) = w_1, \quad t > 0, \tag{1.3}$$

where u_0, $u_1 > 0$ and w_0, $w_1 \in \mathbb{R}$. The treatment of the nonhomogeneities is also interesting from a mathematical point of view. Indeed, almost all results for (1.1) (and for related fourth-order equations like the thin-film model [3]) are shown only for periodic or no-flux boundary conditions or for whole-space problems, in order to avoid integrals with boundary data. In this paper, we show how to deal with nonhomogeneous boundary conditions for (1.1).

More precisely, we show (i) the existence and uniqueness of a classical positive solution u_∞ to the stationary problem corresponding to (1.1), (ii) the existence of global nonnegative weak solutions $u(\cdot,t)$ to the transient problem (1.1), (1.3), and (iii) the exponential convergence of $u(\cdot,t)$ to its steady state u_∞ as $t \to \infty$ in the L^1 norm, under the assumption that the boundary data is such that $\log u_\infty$ is concave. The long-time behavior is illustrated by numerical experiments. Notice that this is the first result of the *stationary* problem corresponding to (1.1) in the literature (if (1.2) or periodic boundary conditions are assumed, the steady state is constant). We also remark that the Wasserstein techniques of [12] cannot be applied to (1.1), (1.3) since this technique relies on the conservation of the L^1 norm which is not the case here.

The long-time behavior of solutions to (1.1) has been studied for periodic boundary conditions [5, 10] and for the boundary conditions (1.2) [15]. In particular, it could be shown that the solutions converge exponentially fast to their (constant) steady states. The decay rate has been numerically computed in [6]. No results are available up to now for the case of the nonhomogeneous boundary conditions (1.3).

Our first main result is the existence and uniqueness of stationary solutions needed in the existence proof for the transient problem.

THEOREM 1.1. *Let u_0, $u_1 > 0$ and w_0, $w_1 \in \mathbb{R}$. Then there exists a unique classical solution $u \in C^\infty([0,1])$ to*

$$(u(\log u)_{xx})_{xx} = 0 \quad \text{in } (0,1), \quad u(0) = u_0,\ u(1) = u_1, \\ u_x(0) = w_0,\ u_x(1) = w_1, \tag{1.4}$$

satisfying $u(x) \geq m > 0$ for all $x \in [0,1]$, and the constant $m > 0$ depends only on the boundary data.

The existence proof is based on a fixed-point argument and appropriate a priori estimates, using the structure of the equation and the one-dimensionality heavily. More precisely, we perform the exponential transformation $u = e^y$ and write the equation in (1.4) as $y_{xx} = (ax+b)e^{-y}$ for some a, $b \in \mathbb{R}$. The key point is to derive uniform bounds on a and b. This implies a uniform H^1 bound for y and, in view of the one-dimensionality, a uniform L^∞ bound for $y = \log u$, hence showing the positivity of u. For the uniqueness we employ a monotonicity property of the operator $\sqrt{u} \mapsto -(u(\log u)_{xx})_{xx}/(2\sqrt{u})$ for suitable functions u (the monotonicity property was first observed in [13]).

The second main result is the existence of solutions to the transient problem. For simplicity, we consider time-independent boundary data only.

THEOREM 1.2. *Let u_0, $u_1 > 0$ and w_0, $w_1 \in \mathbb{R}$. Let $u_I(x) \geq 0$ be integrable such that $\int_0^1 (u_I - \log u_I)dx < \infty$. Then there exists a weak solution u to* (1.1), (1.3) *satisfying $u(x,t) \geq 0$ in $(0,1) \times (0,\infty)$ and*

$$u \in L^{5/2}_{\text{loc}}(0,\infty; W^{1,1}(0,1)) \cap W^{1,10/9}_{\text{loc}}(0,\infty; H^{-2}(0,1)),\ \log u \in L^2_{\text{loc}}(0,\infty; H^2(0,1)).$$

For the proof of this theorem we semidiscretize (1.1) in time and solve at each time step a nonlinear elliptic problem. The main difficulty is to obtain uniform estimates. The idea of [13] is to derive these estimates from a special Lyapunov functional,

$$E_1(t) = \int_0^1 \left(\frac{u}{u_\infty} - \log \frac{u}{u_\infty} \right) dx,$$

which is also called an "entropy" functional. Indeed, a formal computation (made precise in section 3) shows that

$$\frac{dE_1}{dt} + \int_0^1 (\log u)^2_{xx} dx = \int_0^1 u(\log u)_{xx} \left(\frac{1}{u_\infty} \right)_{xx} dx, \tag{1.5}$$

implying that E_1 is nonincreasing if $(1/u_\infty)_{xx} = 0$, which is the case in [13] where $u_\infty = \text{const}$ holds. However, in the general case $(1/u_\infty)_{xx} \neq 0$, the right-hand side of (1.5) still needs to be estimated.

The key idea is to employ the *new* "entropy"

$$E_2(t) = \int_0^1 (\sqrt{u} - \sqrt{u_\infty})^2 dx.$$

A formal computation yields

$$\frac{dE_2}{dt} + 2\int_0^1 \left(\sqrt[4]{\frac{u_\infty}{u}} (\sqrt{u})_{xx} - \sqrt[4]{\frac{u}{u_\infty}} (\sqrt{u_\infty})_{xx} \right)^2 dx = 0. \tag{1.6}$$

With this estimate the right-hand side of (1.5) can be treated. Indeed, the above entropy production integral allows us to find the bound

$$\int_0^1 \left(\sqrt{u}(\log u)_{xx}^2 + (\sqrt[8]{u})_x^4\right) dx \le c \tag{1.7}$$

for some constant $c > 0$ depending only on the boundary data; see Lemma 3.2 for details. (Here and in the following, the notation $(f(u))_x^4$ means $[(f(u))_x]^4$.) Then, using Young's inequality, the right-hand side of (1.5) is bounded from above by

$$\int_0^1 \sqrt{u}(\log u)_{xx}^2 dx + \|1/u_\infty^2\|_{W^{2,\infty}(0,1)} \int_0^1 u^{3/2} dx,$$

which is bounded in view of (1.7). We stress the fact that this idea is new in the literature.

The above estimates are only valid if u is nonnegative. However, no maximum principle is generally available for fourth-order equations. We prove the nonnegativity property by employing the same idea as in the stationary case: after introducing an exponential variable $u = e^y$, we obtain a uniform H^2 bound by (1.5) and (1.7) and hence an L^∞ bound for $y = \log u$, which shows that u is positive. Letting the parameter of the time discretization tend to zero, we conclude the nonnegativity of u.

We notice that, interestingly, the new entropy E_2 is connected with the monotonicity property of $\sqrt{u} \mapsto -(u(\log u)_{xx})_{xx}/(2\sqrt{u})$ since the proof of this property also relies on the estimate (1.6) (see Lemma 2.3 in [13] and (2.7) below).

The physical (relative) entropy

$$E_3(t) = \int_0^1 \left(u \log \frac{u}{u_\infty} - u + u_\infty\right) dx$$

is still another Lyapunov functional. It is used in the proof of the long-time behavior of solutions, which is our final main result.

THEOREM 1.3. *Let the assumptions of Theorem* 1.2 *hold and let* $\int_0^1 u_I(\log u_I - 1)dx < \infty$. *Let* u *be the solution to* (1.1), (1.3) *constructed in Theorem* 1.2 *and let* u_∞ *be the unique solution to* (1.4). *We assume that the boundary data is such that* $\log u_\infty$ *is concave. Then there exist constants* $c, \lambda > 0$ *depending only on the boundary and initial data such that for all* $t > 0$,

$$\|u(\cdot, t) - u_\infty\|_{L^1(0,1)} \le ce^{-\lambda t}.$$

In order to prove this result, we take formally the time derivative of the relative entropy E_3. It can be shown (see section 4 for details) that the assumption $(\log u_\infty)_{xx} \le 0$ allows us to derive

$$\frac{dE_3}{dt} + P \le 0,$$

where $P \ge 0$ denotes the entropy production term involving second derivatives of u. This term can be estimated similarly as in [15] in terms of the entropy yielding

$$\frac{dE_3}{dt} + 2\lambda E_3 \le 0$$

for some $\lambda > 0$. Gronwall's inequality implies the exponential convergence in terms of the relative entropy. A Csiszar–Kullback-type inequality then gives the assertion. The assumption on the concavity of $\log u_\infty$ can be slightly relaxed (see Remark 4.4).

The paper is organized as follows. Section 2 is devoted to the proof of Theorem 1.1. Then the existence of transient solutions (Theorem 1.2) is shown in section 3. Theorem 1.3 is proved in section 4, and finally in section 5, some numerical results are presented.

2. Existence and uniqueness of stationary solutions. In this section we will prove Theorem 1.1. First, we perform the transformation of variables $y = \log u$ and consider the problem

$$(e^y y_{xx})_{xx} = 0 \quad \text{in } (0,1), \quad y(0) = y_0,\ y(1) = y_1,\ y_x(0) = \alpha,\ y_x(1) = \beta, \tag{2.1}$$

where $y_0 = \log u_0$, $y_1 = \log u_1$, $\alpha = w_0/u_0$, and $\beta = w_1/u_1$. Clearly, any classical solution of (2.1) is a positive classical solution of (1.4). We show first some a priori estimates for the solution of (2.1).

LEMMA 2.1. *Let y be a classical solution to* (2.1). *Then*

$$y(x) \le M := \max\{y_0, y_1\} + |\alpha| + |\beta|. \tag{2.2}$$

Proof. First we observe that there exist constants a, $b \in \mathbb{R}$ such that y solves the equation $y_{xx} = (ax+b)e^{-y}$. This implies that y_{xx} can change its sign at most once. In the following we consider several cases for the sign of $y_{xx}(0)$ and $y_{xx}(1)$.

First case. Let $y_{xx}(0) \ge 0$ and $y_{xx}(1) \ge 0$. Since y_{xx} can change the sign at most once it follows that $y_{xx} \ge 0$ in $(0,1)$. We conclude that $y(x) \le \max\{y_0, y_1\}$ for all $x \in [0,1]$.

Second case. Let $y_{xx}(0) \ge 0$ and $y_{xx}(1) < 0$. There exists $x_1 \in [0,1)$ such that $y_{xx}(x_1) = 0$. Therefore, $ax + b \ge 0$ for all $x \in [0, x_1]$ and $ax + b \le 0$ for all $x \in [x_1, 1]$. A Taylor expansion gives for all $x \in [x_1, 1]$

$$\begin{aligned} y(x) &= y(1) + y_x(1)(x-1) + \int_x^1 (s-x) y_{xx}(s)\,ds \\ &= y_1 + \beta(x-1) + \int_x^1 (s-x)(as+b)e^{-y(s)}ds \le \max\{y_0, y_1\} + |\beta|. \end{aligned}$$

We claim that $y(x) \le \max\{y_0, y_1\} + |\beta|$ holds for all $x \in [0, x_1]$. For this, let $x_2 \in [0, x_1]$ be such that $y(x_2) = \max\{y(x) : x \in [0, x_1]\}$. Suppose that $y(x_2) > \max\{y_0, y_1\} + |\beta|$. Then $x_2 \in (0, x_1)$ and, since $y(x)$ reaches a maximum at the interior point x_2, $y_{xx}(x_2) \le 0$. Since $x_2 \in (0, x_1)$, we have $y_{xx}(x_2) = (ax_2 + b)e^{-y(x_2)} \ge 0$. This shows that $y_{xx}(x_2) = 0$. But then $y_{xx}(x_2) = (ax_2 + b)e^{-y(x_2)}$ implies that $ax_2 + b = 0$. Since also $ax_1 + b = 0$, it follows that $a = b = 0$ and thus $y_{xx}(x) = 0$ for all $x \in [0,1]$; this is a contradiction to $y_{xx}(1) < 1$. Hence, $y(x) \le \max\{y_0, y_1\} + |\beta|$ for all $x \in [0,1]$.

Third case. Let $y_{xx}(0) < 0$ and $y_{xx}(1) \ge 0$. By similar arguments as in the second case, it can be shown that $y(x) \le \max\{y_0, y_1\} + |\alpha|$ for all $x \in [0,1]$.

Fourth case. Let $y_{xx}(0) < 0$ and $y_{xx}(1) < 0$. This implies that $ax + b < 0$ for all $x \in [0,1]$ and, by a Taylor expansion,

$$y(x) = y_0 + \alpha x + \int_0^x (x-s)(as+b)e^{-y(s)}ds \le y_0 + |\alpha|, \quad x \in [0,1].$$

The lemma is proved. ∎

LEMMA 2.2. *Let y be a classical solution to* (2.1). *Then there exists a constant $K > 0$ depending only on y_0, y_1, α, and β such that*

$$\|y\|_{H^2(0,1)} \le K.$$

Proof. There exist constants a, $b \in \mathbb{R}$ such that y solves the equation

$$y_{xx} = (ax+b)e^{-y} \quad \text{in } (0,1), \tag{2.3}$$

and $b = e^{y_0} y_{xx}(0)$, $a = e^{y_1} y_{xx}(1) - e^{y_0} y_{xx}(0)$. In order to obtain a uniform estimate for y_{xx} we first have to find uniform estimates for a and b. For this, we multiply (2.3) by y_x^2 and integrate over $(0,1)$:

$$\int_0^1 (ax+b)e^{-y} y_x^2 dx = \int_0^1 y_{xx} y_x^2 dx = \frac{1}{3}\int_0^1 (y_x^3)_x dx = \frac{1}{3}(\beta^3 - \alpha^3).$$

Next we multiply (2.3) by y_{xx}, integrate over $(0,1)$, integrate by parts, and use the above equality:

$$\begin{aligned}\int_0^1 y_{xx}^2 dx &= \int_0^1 (ax+b)e^{-y} y_{xx} dx \\ &= \int_0^1 (ax+b)e^{-y} y_x^2 dx - a\int_0^1 e^{-y} y_x dx + [(ax+b)e^{-y(x)} y_x(x)]_0^1 \\ &= \frac{1}{3}(\beta^3 - \alpha^3) + a(e^{-y_1} - e^{-y_0}) + (a+b)e^{-y_1}\beta - be^{-y_0}\alpha.\end{aligned}$$

By Young's inequality this becomes

$$\int_0^1 y_{xx}^2 dx \le C + \frac{1}{60}e^{-2M} a^2 + \frac{1}{12}e^{-2M} b^2, \tag{2.4}$$

where $C := (\beta^3 - \alpha^3)/3 + 15e^{2M}((1+\beta)e^{-y_1} - e^{-y_0})^2 + 3e^{2M}(\beta e^{-y_1} - \alpha e^{-y_0})^2$. Taking the square of (2.3) and integrating over $(0,1)$ yields, by Lemma 2.1,

$$\begin{aligned}\int_0^1 y_{xx}^2 dx &= \int_0^1 (ax+b)^2 e^{-2y} dx \ge e^{-2M}\int_0^1 (ax+b)^2 dx \\ &= \frac{1}{3}e^{-2M}(a^2 + 3ab + 3b^2) \ge \frac{1}{3}e^{-2M}\left(\frac{a^2}{10} + \frac{b^2}{2}\right),\end{aligned} \tag{2.5}$$

where we have used the Young inequality $3ab \ge -9a^2/10 - 5b^2/2$. Putting together (2.4) and (2.5), we obtain

$$\frac{a^2}{10} + \frac{b^2}{2} \le 3e^{2M}\int_0^1 y_{xx}^2 dx \le 3e^{2M} C + \frac{a^2}{20} + \frac{b^2}{4}. \tag{2.6}$$

Therefore, a and b are bounded by a constant which depends only on y_0, y_1, α, and β. By (2.4) this gives a uniform estimate for $\|y_{xx}\|_{L^2(0,1)}$ and, employing Poincaré's inequality, also for $\|y\|_{H^2(0,1)}$. ◻

Proof of Theorem 1.1. We wish to employ the Leray–Schauder fixed-point theorem. For this let $\sigma \in [0,1]$ and $z \in H^1(0,1)$ and let $y \in H^2(0,1)$ be the unique solution of

$$(e^z y_{xx})_{xx} = 0 \quad \text{in } (0,1), \quad y(0) = \sigma y_0,\ y(1) = \sigma y_1,\ y_x(0) = \sigma\alpha,\ y_x(1) = \sigma\beta.$$

This defines a fixed-point operator $S : H^1(0,1) \times [0,1] \to H^1(0,1)$, $S(z,\sigma) = y$. Clearly, $S(z,0) = 0$ for all z. Moreover, by standard arguments, S is continuous and compact, since the embedding $H^2(0,1) \hookrightarrow H^1(0,1)$ is compact. It remains to show that there exists a constant $K > 0$ such that for all $\sigma \in [0,1]$ and for any fixed point y of $S(\cdot,\sigma)$, the estimate $\|y\|_{H^1(0,1)} \le K$ holds. Lemma 2.2 settles the case $\sigma = 1$. For $\sigma < 1$, a similar proof as in Lemma 2.2 shows the existence of a constant $K > 0$ such that $\|y\|_{H^2(0,1)} \le K$ holds. By the Leray–Schauder theorem, this proves the existence of a solution $y \in H^2(0,1)$ to (2.1).

Actually, the solution y is a classical solution. Indeed, y satisfies $y_{xx} = (ax + b)e^{-y} \in H^2(0,1)$ for some a, $b \in \mathbb{R}$, and hence, $y \in H^4(0,1)$. By bootstrapping, $y \in H^n(0,1)$ for all $n \in \mathbb{N}$ and y is a classical solution.

In order to prove the uniqueness of solutions, we extend an idea of [13]. Let u_1 and u_2 be two positive classical solutions to (1.4). We multiply (1.4) for u_1 by $1 - \sqrt{u_2/u_1}$ and (1.4) for u_2 by $\sqrt{u_1/u_2} - 1$, integrate, and take the difference. This yields, by integrating by parts,

$$
\begin{aligned}
(2.7)\quad 0 &= \int_0^1 \Big[(u_1(\log u_1)_{xx})_{xx}(1 - \sqrt{u_2/u_1}) - (u_2(\log u_2)_{xx})_{xx}(\sqrt{u_1/u_2} - 1)\Big]\,dx \\
&= 2\int_0^1 \Big[(\sqrt{u_1})_{xxxx} - \frac{1}{\sqrt{u_1}}(\sqrt{u_1})_{xx}^2 - (\sqrt{u_2})_{xxxx} \\
&\qquad\qquad + \frac{1}{\sqrt{u_2}}(\sqrt{u_2})_{xx}^2\Big](\sqrt{u_1} - \sqrt{u_2})dx \\
&= 2\int_0^1 \Big[((\sqrt{u_1})_{xx} - (\sqrt{u_2})_{xx})(\sqrt{u_1} - \sqrt{u_2})_{xx} \\
&\quad - (\sqrt{u_1})_{xx}^2\left(1 - \sqrt{\tfrac{u_2}{u_1}}\right) + (\sqrt{u_2})_{xx}^2\left(\sqrt{\tfrac{u_1}{u_2}} - 1\right)\Big]dx \\
&= 2\int_0^1 \left(\sqrt[4]{\frac{u_2}{u_1}}(\sqrt{u_1})_{xx} - \sqrt[4]{\frac{u_1}{u_2}}(\sqrt{u_2})_{xx}\right)^2.
\end{aligned}
$$

Therefore,

$$0 = \sqrt[4]{\frac{u_2}{u_1}}(\sqrt{u_1})_{xx} - \sqrt[4]{\frac{u_1}{u_2}}(\sqrt{u_2})_{xx} \quad \text{in } (0,1).$$

Writing $u_1 = e^{y_1}$ and $u_2 = e^{y_2}$, this identity is equal to

$$
\begin{aligned}
0 &= e^{(y_2-y_1)/4}(e^{y_1/2})_{xx} - e^{(y_1-y_2)/4}(e^{y_2/2})_{xx} \\
&= \frac{1}{2}e^{(y_2+y_1)/4}\left(y_{1,xx} + \frac{1}{2}y_{1,x}^2\right) - \frac{1}{2}e^{(y_1+y_2)/4}\left(y_{2,xx} + \frac{1}{2}y_{2,x}^2\right),
\end{aligned}
$$

and hence

$$(2.8)\qquad y_{1,xx} - y_{2,xx} = -\frac{1}{2}(y_{1,x}^2 - y_{2,x}^2) \quad \text{in } (0,1).$$

We integrate this equation over $(0, x_0)$, use the boundary condition $y_{1x}(0) = y_{2x}(0)$, and take the supremum,

$$\|(y_1 - y_2)_x\|_{L^\infty(0,x_0)} \le \int_0^{x_0} |(y_1 + y_2)_x| \cdot |(y_1 - y_2)_x| dx \le x_0 L \|(y_1 - y_2)_x\|_{L^\infty(0,x_0)},$$

where $L = \|y_{1,x}\|_{L^\infty(0,1)} + \|y_{2,x}\|_{L^\infty(0,1)}$. Choosing $x_0 = 1/2L$ gives $(y_1 - y_2)_x = 0$ and hence $y_1 - y_2 = 0$ in $[0, x_0]$. In particular, $(y_1 - y_2)_x(x_0) = 0$. Therefore, integrating (2.8) over $(x_0, 2x_0)$ we obtain by the same arguments that $y_1 - y_2 = 0$ in $[x_0, 2x_0]$.

After a finite number of steps we achieve $y_1 - y_2 = 0$ in $[0,1]$. This proves the uniqueness of solutions. ∎

Remark 2.3. Equation (2.3) with $y(0) = y(1)$ and $y_x(0) = -y_x(1) \le 0$ is formally related to a combustion problem. Indeed, the boundary conditions imply that y is symmetric around $x = \frac{1}{2}$ and that $y(x) \le y(0) = y_0$ holds for any $x \in [0,1]$. The symmetry implies further $a = e^{y_0}(y_{xx}(1) - y_{xx}(0)) = 0$ and moreover, $b = e^{y_0} y_{xx}(0) \ge 0$. Thus we can write (2.3) as $y_{xx} = be^{-y}$ or, introducing $z(x) = -y(x)$,

$$z_{xx} + be^z = 0 \quad \text{in } (0,1), \quad z(0) = z(1) = -y_0.$$

This is the solid fuel ignition model of [2]. It is well known that there exists $b^* > 0$ such that this problem has exactly two solutions if $b \in (0, b^*)$, it has exactly one solution if $b = b^*$, and it has no solution if $b > b^*$ [2, 11]. This relation provides a better bound for b (for the above special boundary conditions) than the estimate (2.6). Indeed, $a = 0$ and b is uniformly bounded by a number $b^* > 0$ *independently* of the boundary conditions (and depending only on the domain $(0,1)$).

3. Existence of transient solutions. In order to prove Theorem 1.2 we again perform the exponential change of unknowns and we semidiscretize (1.1) in time. For this, we divide the time interval $(0,T]$ for some $T > 0$ into N subintervals $(t_{k-1}, t_k]$, with $k = 1, \ldots, N$, where $0 = t_0 < \cdots < t_N = T$. Define $\tau_k = t_k - t_{k-1} > 0$ and $\tau = \max\{\tau_k : k = 1, \ldots, N\}$. We assume that $\tau \to 0$ as $N \to \infty$.

Let $u_\infty > 0$ be the unique classical solution to (1.4) and set $y_\infty = \log u_\infty$. We perform the transformation $z = \log(u/u_\infty)$ and $z_0 = \log(u_I/u_\infty)$. For given $k \in \{1, \ldots, N\}$ and z_{k-1} we first solve the semidiscrete problem

$$\frac{e^{y_\infty}}{\tau_k}(e^{z_k} - e^{z_{k-1}}) = -\left(e^{z_k + y_\infty}(z_k + y_\infty)_{xx}\right)_{xx}, \quad z_k \in H_0^2(0,1). \tag{3.1}$$

Proposition 3.1. *For each $k = 1, \ldots, N$, there exists a unique weak solution $z_k \in H_0^2(0,1)$ to* (3.1).

For the proof of this proposition we first show some a priori estimates.

Lemma 3.2. *Let $z_k \in H_0^2(0,1)$ be a weak solution to* (3.1). *Then there exists a constant $c > 0$ depending only on T, u_I, and u_∞ such that*

$$\|e^{z_k/2}\|_{L^2(0,1)} \le c, \tag{3.2}$$

$$\sum_{i=1}^N \tau_i \int_0^1 e^{z_i/2}\left((z_i + y_\infty)_{xx}^2 + (z_i + y_\infty)_x^4\right) \le c, \tag{3.3}$$

$$\sum_{i=1}^N \tau_i \|e^{z_i}\|_{L^\infty(0,1)} \le c. \tag{3.4}$$

Proof. Similarly as in the uniqueness proof of Theorem 1.1 we use the test functions $1 - e^{-z_k/2} \in H_0^2(0,1)$ in the weak formulation of the semidiscretized equation (3.1) and $e^{z_k/2} - 1 \in H_0^2(0,1)$ in the weak formulation of the stationary equation (1.4) and take the sum of the corresponding equations:

$$\begin{aligned} \frac{1}{\tau_k}\int_0^1 e^{y_\infty}(e^{z_k} - e^{z_{k-1}})(1 - e^{-z_k/2})dx &= \int_0^1 e^{z_k + y_\infty}(z_k + y_\infty)_{xx}(e^{-z_k/2})_{xx}dx \\ &\quad + \int_0^1 e^{y_\infty} y_{\infty,xx}(e^{z_k/2})_{xx}dx. \end{aligned} \tag{3.5}$$

The right-hand side is equal to the first integral in (2.7) with $u_1 = e^{z_k+y_\infty}$ and $u_2 = e^{y_\infty}$. Therefore, the right-hand side is equal to the expression

$$-2\int_0^1 \left(e^{-z_k/4}(e^{(z_k+y_\infty)/2})_{xx} - e^{z_k/4}(e^{y_\infty/2})_{xx}\right)^2 dx.$$

For the left-hand side of (3.5) we write

$$\begin{aligned}
&\frac{1}{\tau_k}\int_0^1 e^{y_\infty}(e^{z_k} - e^{z_{k-1}})(1-e^{-z_k/2})dx \\
&= \frac{1}{\tau_k}\int_0^1 e^{y_\infty}((e^{z_k/2}-1)^2 - (e^{z_{k-1}/2}-1)^2)dx + \frac{1}{\tau_k}\int_0^1 e^{y_\infty}(e^{z_k/4} - e^{z_{k-1}/2-z_k/4})^2 dx \\
&\geq \frac{1}{\tau_k}\int_0^1 e^{y_\infty}\left((e^{z_k/2}-1)^2 - (e^{z_{k-1}/2}-1)^2\right)dx.
\end{aligned}$$

Therefore, we conclude from (3.5), for all $k = 1, \ldots, N$,

$$\begin{aligned}
&\frac{1}{\tau_k}\int_0^1 e^{y_\infty}(e^{z_k/2}-1)^2 dx + 2\int_0^1 \left(e^{-z_k/4}(e^{(z_k+y_\infty)/2})_{xx} - e^{z_k/4}(e^{y_\infty/2})_{xx}\right)^2 dx \\
&\qquad \leq \frac{1}{\tau_k}\int_0^1 e^{y_\infty}(e^{z_{k-1}/2}-1)^2 dx. \qquad (3.6)
\end{aligned}$$

This yields

$$\int_0^1 e^{y_\infty}(e^{z_k/2}-1)^2 dx \leq \int_0^1 e^{y_\infty}(e^{z_0/2}-1)^2 dx = \int_0^1 (\sqrt{u_I} - \sqrt{u_\infty})^2 dx < \infty \qquad (3.7)$$

and thus (3.2). Moreover, after summing up (3.6),

$$2\sum_{i=1}^k \tau_i \int_0^1 \left(e^{-z_i/4}(e^{(z_i+y_\infty)/2})_{xx} - e^{z_i/4}(e^{y_\infty/2})_{xx}\right)^2 dx \leq \int_0^1 e^{y_\infty}(e^{z_0/2}-1)^2 dx.$$

Young's inequality gives

$$4\sum_{i=1}^k \tau_i \int_0^1 e^{-z_i/2}\left(e^{(z_i+y_\infty)/2}\right)_{xx}^2 dx \leq c + c\sum_{i=1}^k \tau_i \int_0^1 e^{z_i/2} dx,$$

where here and in the following, $c > 0$ denotes a generic constant depending only on T, u_I, and u_∞. In view of (3.7), the right-hand side is uniformly bounded. Hence

$$\begin{aligned}
&\sum_{i=1}^k \tau_i \int_0^1 e^{-(z_i+y_\infty)/2}\left(e^{(z_i+y_\infty)/2}\right)_{xx}^2 dx \\
&\leq \|e^{y_\infty/2}\|_{L^\infty(0,1)} \sum_{i=1}^k \tau_i \int_0^1 e^{-z_i/2}\left(e^{(z_i+y_\infty)/2}\right)_{xx}^2 dx \leq c.
\end{aligned}$$

Now the assertion (3.3) follows since, by integration by parts,

$$\int_0^1 e^{u/2}u_x^2 u_{xx} dx = -\frac{1}{6}\int_0^1 e^{u/2}u_x^4 + \frac{1}{3}(e^{u(1)/2}u_x(1)^3 - e^{u(0)/2}u_x(0)^3)$$

for all $u \in H^2(0,1)$, and hence,

$$\begin{aligned}
&\int_0^1 e^{-(z_i+y_\infty)/2}\left(e^{(z_i+y_\infty)/2}\right)_{xx}^2 dx \\
&= \frac{1}{4}\int_0^1 e^{(z_i+y_\infty)/2}\left((z_i+y_\infty)_{xx}^2 + \frac{1}{4}(z_i+y_\infty)_x^4 + (z_i+y_\infty)_{xx}(z_i+y_\infty)_x^2\right)dx \\
&= \frac{1}{4}\int_0^1 e^{(z_i+y_\infty)/2}\left((z_i+y_\infty)_{xx}^2 + \frac{1}{12}(z_i+y_\infty)_x^4\right)dx + \frac{1}{12}(e^{y_1/2}\beta^3 - e^{y_0/2}\alpha^3).
\end{aligned}$$

Finally, (3.4) is a consequence of (3.3) and the Poincaré–Sobolev inequality since

$$\int_0^1 e^{z_i/2}(z_i)_x^4 dx = 8^4\int_0^1 (e^{z_i/8})_x^4 \geq c\|e^{z_i/8}\|_{L^\infty(0,1)}^4.$$

This shows the lemma. ∎

LEMMA 3.3. *Let $z_k \in H_0^2(0,1)$ be a weak solution to* (3.1). *Then there exists a constant $c > 0$ depending only on T, u_I, and u_∞ such that*

$$\int_0^1 (e^{z_k} - z_k)dx + \sum_{i=1}^k \tau_i \int_0^1 (z_i + y_\infty)_{xx}^2 dx \leq c. \tag{3.8}$$

Proof. We choose the test function $e^{-y_\infty}(1-e^{-z_k}) \in H_0^2(0,1)$ in the weak formulation of (3.1). Then, by Young's inequality,

$$\begin{aligned}
&\int_0^1 (e^{z_k} - e^{z_{k-1}})(1-e^{-z_k})dx \\
&= -\tau_k\int_0^1 e^{z_k}(z_k+y_\infty)_{xx}(y_{\infty,x}^2 - y_{\infty,xx})dx - \tau_k\int_0^1 (z_k+y_\infty)_{xx}^2 dx \\
&\quad + \tau_k\int_0^1 (z_k+y_\infty)_x^2(z_k+y_\infty)_{xx}dx \\
&\leq \tau_k\int_0^1 e^{z_k/2}(z_k+y_\infty)_{xx}^2 dx + \tau_k\int_0^1 e^{3z_k/2}(y_{\infty,x}^2 - y_{\infty,xx})^2 dx \\
&\quad - \tau_k\int_0^1 (z_k+y_\infty)_{xx}^2 dx + \frac{\tau_k}{3}(\beta^3-\alpha^3).
\end{aligned}$$

In view of (3.3) and (3.4), the right-hand side is uniformly bounded. The left-hand side can be estimated by

$$\int_0^1 (e^{z_k} - e^{z_{k-1}})(1-e^{-z_k})dx \geq \int_0^1 (e^{z_k}-z_k)dx - \int_0^1 (e^{z_{k-1}} - z_{k-1})dx,$$

which is a consequence of the elementary inequality $e^x - 1 \geq x$ for all $x \in \mathbb{R}$. Thus, (3.8) is proved. ∎

Proof of Proposition 3.1. The existence of a solution to (3.1) is shown by the Leray–Schauder fixed-point theorem. For this, let $k \in \{1,\ldots,N\}$ and z_{k-1} be given. Furthermore, let $w \in H^1(0,1)$ and $\sigma \in [0,1]$, and define the linear forms

$$\begin{aligned}
a(z,\phi) &= \int_0^1 e^{w+y_\infty} z_{xx}\phi_{xx}dx, \\
F(\phi) &= -\frac{1}{\tau_k}\int_0^1 e^{y_\infty}(e^w - e^{z_{k-1}})\phi dx - \int_0^1 e^{w+y_\infty}y_{\infty,xx}\phi_{xx}dx,
\end{aligned}$$

where $\phi \in H_0^2(0,1)$. Consider the linear problem

$$a(z,\phi) = \sigma F(\phi) \quad \text{for all } \phi \in H_0^2(0,1).$$

By the Lax–Milgram lemma, there exists a unique solution $z \in H_0^2(0,1)$ to this problem. This defines the fixed-point operator $S : H^1(0,1)\times[0,1] \to H^1(0,1)$, $S(w,\sigma) = z$. It is not difficult to show that S is continuous and compact, since the embedding $H_0^2(0,1) \hookrightarrow H^1(0,1)$ is compact. Moreover, $S(w,0) = 0$ for all $w \in H^1(0,1)$. It remains to prove that any fixed point of S satisfies a uniform bound in $H^1(0,1)$. In fact, Lemma 3.3 shows that any fixed point $z \in H_0^2(0,1)$ is uniformly bounded if $\sigma = 1$. The estimate for $\sigma < 1$ is similar (and, in fact, independent of σ). This provides the wanted bound in $H^1(0,1)$, and the Leray–Schauder theorem can be applied to yield the existence of a solution to (3.1). ◻

For the proof of Theorem 1.2 we need some more uniform estimates. Let $z^{(N)}$ be defined by $z^{(N)}(x,t) = z_k(x)$ if $t \in (t_{k-1},t_k]$, $x \in (0,1)$.

Lemma 3.4. *The following estimates hold:*

$$\|z^{(N)}\|_{L^\infty(0,T;L^1(0,1))} + \|z^{(N)}\|_{L^2(0,T;H^2(0,1))} \le c, \tag{3.9}$$

$$\|z^{(N)}\|_{L^{5/2}(0,T;W^{1,\infty}(0,1))} + \|e^{z^{(N)}}\|_{L^{5/2}(0,T;W^{1,1}(0,1))} \le c, \tag{3.10}$$

where $c > 0$ depends only on u_I and the boundary data.

Proof. The inequality $e^x - x \ge |x|$ for all $x \in \mathbb{R}$ and the estimate (3.8) imply that $z^{(N)}$ is uniformly bounded in $L^\infty(0,T;L^1(0,1))$ which, together with (3.8), shows (3.9). Then, using the Poincaré and Gagliardo–Nirenberg inequalities, we obtain from (3.8)

$$\begin{aligned}\|z^{(N)}\|_{L^{5/2}(0,T;W^{1,\infty}(0,1))} &\le c\|z_x^{(N)}\|_{L^{5/2}(0,T;L^\infty(0,1))} \\ &\le c\|z^{(N)}\|^{1/5}_{L^\infty(0,T;L^1(0,1))}\|z^{(N)}\|^{4/5}_{L^2(0,T;H^2(0,1))} \le c.\end{aligned}$$

This estimate, (3.2), and the first bound in (3.9) imply (3.10) since

$$\begin{aligned}\|e^{z^{(N)}}\|_{L^{5/2}(0,T;W^{1,1}(0,1))} &\le c\left(\|e^{z^{(N)}}\|_{L^{5/2}(0,T;L^1(0,1))} + \|(e^{z^{(N)}})_x\|_{L^{5/2}(0,T;L^1(0,1))}\right) \\ &\le c\|e^{z^{(N)}}\|_{L^{5/2}(0,T;L^1(0,1))} \\ &\quad + c\|e^{z^{(N)}}\|_{L^\infty(0,T;L^1(0,1))}\|z_x^{(N)}\|_{L^{5/2}(0,T;L^\infty(0,1))} \\ &\le c.\end{aligned}$$

The lemma is proved. ◻

We also need an estimate for the discrete time derivative. For this, introduce the shift operator $(\sigma_N(z^{(N)}))(\cdot,t) = z_{k-1}$ for $t \in (t_{k-1},t_k]$.

Lemma 3.5. *The following estimate holds:*

$$\|e^{z^{(N)}} - e^{\sigma_N(z^{(N)})}\|_{L^{10/9}(0,T;H^{-2}(0,1))} \le c\tau, \tag{3.11}$$

where $c > 0$ depends only on u_I and u_∞.

Proof. From (3.1) and Hölder's inequality we obtain

$$\begin{aligned}\frac{1}{\tau_k}\|e^{z^{(N)}} - e^{\sigma_N(z^{(N)})}\|_{L^{10/9}(0,T;H^{-2}(0,1))} &\le \|e^{z^{(N)}+y_\infty}(z^{(N)}+y_\infty)_{xx}\|_{L^{10/9}(0,T;L^2(0,1))} \\ &\le \|e^{z^{(N)}+y_\infty}\|_{L^{5/2}(0,T;L^\infty(0,1))}\|(z^{(N)}+y_\infty)_{xx}\|_{L^2(0,T;L^2(0,1))}\end{aligned}$$

and the right-hand side is uniformly bounded by (3.9) and (3.10) since $W^{1,1}(0,1) \hookrightarrow L^\infty(0,1)$. $\square$

Proof of Theorem 1.2. For any $N \in \mathbb{N}$, there exists a solution $z^{(N)} \in L^2(0,T; H_0^2(0,1))$ to the sequence of recursive equations (3.1) satisfying $z^{(N)}(\cdot,0) = z_0$. The uniform bounds (3.10) and (3.11) and the compact embedding $W^{1,1}(0,1) \hookrightarrow L^1(0,1)$ allow us to apply Theorem 5 of [17] (Aubin's lemma) yielding the existence of a subsequence of $e^{z^{(N)}}$ (not relabeled) such that $e^{z^{(N)}} \to \rho$ strongly in $L^1(0,T;L^1(0,1))$ and hence also in $L^1(0,T;H^{-2}(0,1))$. The above results give, using (3.2) and $L^1(0,1) \hookrightarrow H^{-2}(0,1)$,

$$\begin{aligned}
\|e^{z^{(N)}} - \rho\|^2_{L^2(0,T;H^{-2}(0,1))} &\le \|e^{z^{(N)}} - \rho\|_{L^\infty(0,T;H^{-2}(0,1))}\|e^{z^{(N)}} - \rho\|_{L^1(0,T;H^{-2}(0,1))} \\
&\le c\left(\|e^{z^{(N)}}\|_{L^\infty(0,T;L^1(0,1))} + \|\rho\|_{L^\infty(0,T;L^1(0,1))}\right) \\
&\quad \times \|e^{z^{(N)}} - \rho\|_{L^1(0,T;H^{-2}(0,1))} \\
&\le c\|e^{z^{(N)}} - \rho\|_{L^1(0,T;H^{-2}(0,1))} \to 0 \quad \text{as } N \to \infty. \qquad (3.12)
\end{aligned}$$

Moreover, the estimate (3.9) provides the existence of a subsequence, also denoted by $z^{(N)}$, such that

$$z^{(N)} \rightharpoonup z \quad \text{weakly in } L^2(0,T;H^2(0,1)) \text{ as } N \to \infty. \qquad (3.13)$$

We claim now that $e^z = \rho$. For this, let w be a smooth function. Letting $N \to \infty$ in

$$0 \le \int_0^T \langle e^{z^{(N)}} - e^w, z^{(N)} - w\rangle_{H^{-2},H_0^2} dt$$

and using the convergence results (3.12) and (3.13) yield

$$0 \le \int_0^T \int_0^1 (\rho - e^w)(w - z)dxdt.$$

The strict monotonicity of $x \mapsto e^x$ then implies that $e^z = \rho$.

Thus, $e^{z^{(N)}} \to e^z$ strongly in $L^1(0,T;L^1(0,1))$ and (maybe for a subsequence) a.e. The uniform bound (3.10) implies that (after extracting a subsequence) $e^{z^{(N)}} \rightharpoonup e^z$ weakly* in $L^{5/2}(0,T;L^\infty(0,1))$ since $W^{1,1}(0,1) \hookrightarrow L^\infty(0,1)$. Therefore, we conclude via Lebesgue's convergence theorem that

$$e^{z^{(N)}} \to e^z \quad \text{strongly in } L^2(0,T;L^2(0,1)). \qquad (3.14)$$

Finally, the uniform estimate (3.11) gives for a subsequence

$$\frac{1}{\tau}(e^{z^{(N)}} - e^{\sigma_N(z^{(N)})}) \rightharpoonup (e^z)_t \quad \text{weakly in } L^{10/9}(0,T;H^{-2}(0,1)). \qquad (3.15)$$

The convergence results (3.13)–(3.15) allow us to pass to the limit $N \to \infty$ in the weak formulation of (3.1) to obtain a weak solution $z \in L^2(0,T;H_0^2(0,1))$ to

$$e^{y_\infty}(e^z)_t = -(e^{z+y_\infty}(z + y_\infty)_{xx})_{xx} \quad \text{in } (0,1),\ t > 0,$$

such that $z(\cdot,0) = z_0 = \log(u_I/u_\infty)$ in the sense of $H^{-2}(0,1)$. Transforming back to the variable $u = e^{z+y_\infty}$ gives the assertion. $\square$

4. Long-time behavior of the solutions. This section is devoted to the proof of Theorem 1.3. The proof is based on the entropy–entropy production method. For this we need the following lemma for lower and upper estimates for the entropy

$$E_3 = \int_0^1 e^{y_\infty}(e^z(z-1)+1)dx.$$

LEMMA 4.1. *Let z, $y_\infty \in L^\infty(0,1)$. Then*

$$c_1\left(\int_0^1 e^{y_\infty}|e^z-1|dx\right)^2 \le E_3 \le c_2\|e^{z/2}-1\|^2_{L^\infty(0,1)}, \tag{4.1}$$

where c_1, $c_2 > 0$ depend on $\|e^{y_\infty}\|_{L^\infty(0,1)}$ and $\|e^z\|_{L^1(0,1)}$.

The lower bound for E_3 is a Csiszar–Kullback-type inequality. A similar version of this lemma is shown in [15].

Proof. The upper bound is proved by expanding the function $f(x) = x^2(\log x^2 - 1) + 1$ around $x = 1$,

$$\begin{aligned} f(e^{z/2}) &= f(1) + f'(1)(e^{z/2}-1) + \frac{1}{2}f''(\theta)(e^{z/2}-1)^2 \\ &= 2(\log\theta+1)(e^{z/2}-1)^2 \le 2(e^{z/2}+1)(e^{z/2}-1)^2, \end{aligned}$$

where θ lies between $e^{z/2}$ and 1, and using the inequality $\log\theta \le \theta - 1 \le \max\{e^{z/2}, 1\} - 1 \le e^{z/2}$. Then

$$E_3 \le 2\int_0^1 e^{y_\infty}(e^{z/2}+1)(e^{z/2}-1)^2dx \le 2\|e^{y_\infty}\|_{L^\infty(0,1)}(\|e^z\|^{1/2}_{L^1(0,1)}+1)\|e^{z/2}-1\|^2_{L^\infty(0,1)},$$

and we set $c_2 = 2\|e^{y_\infty}\|_{L^\infty(0,1)}(\|e^z\|^{1/2}_{L^1(0,1)} + 1)$.

For the lower bound we observe that a Taylor expansion of $f(x) = x(\log x - 1) + 1$ around $x = 1$ yields

$$e^{2y_\infty}(e^z(z-1)+1) = \frac{e^{2y_\infty}}{2\theta}(e^z-1)^2,$$

and $\theta = \theta(z)$ lies between e^z and 1. Then, by the Cauchy–Schwarz inequality,

$$\begin{aligned} \int_0^1 e^{y_\infty}|e^z-1|dx &\le \int_{\{z<0\}} e^{y_\infty}(1-e^z)dx + \int_{\{z>0\}} e^{y_\infty}(e^z-1)dx \\ &\le \int_{\{z<0\}} e^{y_\infty}\frac{1-e^z}{\theta(z)^{1/2}}dx + \int_{\{z>0\}} e^{y_\infty}\frac{e^z-1}{\theta(z)^{1/2}}\theta(z)^{1/2}dx \\ &\le \operatorname{meas}\{z<0\}^{1/2}\left(\int_{\{z<0\}} e^{2y_\infty}\frac{(1-e^z)^2}{\theta(z)}dx\right)^{1/2} \\ &\quad + \left(\int_{\{z>0\}} e^{2y_\infty}\frac{(e^z-1)^2}{\theta(z)}dx\right)^{1/2}\left(\int_{\{z>0\}}\theta(z)dx\right)^{1/2} \\ &\le (1+\|e^z\|^{1/2}_{L^1(0,1)})\left(\int_0^1 e^{2y_\infty}\frac{(e^z-1)^2}{\theta(z)}dx\right)^{1/2} \\ &\le \sqrt{2}\|e^{y_\infty}\|^{1/2}_{L^\infty(0,1)}(1+\|e^z\|^{1/2}_{L^1(0,1)})E_3^{1/2}, \end{aligned}$$

and the assertion follows with $c_1^{-1} = 2\|e^{y_\infty}\|_{L^\infty(0,1)}(1+\|e^z\|_{L^1(0,1)}^{1/2})^2$. $\square$

Proof of Theorem 1.3. The idea is to differentiate the entropy E_3 of the introduction with respect to time and to use the differential equation (1.1). Since we do not have enough regularity for the solution u to (1.1), we need to regularize. We set as in the proof of Theorem 1.2 $u_\infty = e^{y_\infty}$, where u_∞ is the unique solution to (1.4). There exist numbers a, $b \in \mathbb{R}$ such that $e^{y_\infty} y_{\infty,xx} = ax + b \le 0$ for all $x \in (0,1)$ since $y_\infty = \log u_\infty$ is assumed to be concave. This implies that $y_\infty \ge \min\{y_\infty(0), y_\infty(1)\}$ and hence $e^{y_\infty} \ge \min\{u_0, u_1\}$ in $(0,1)$. Furthermore, let $z_k \in H_0^2(0,1)$ be a solution to (3.1) for given z_{k-1}. We assume for simplicity that $\tau = \tau_k$ for all $k \in \mathbb{N}$.

Using z_k as a test function in the weak formulation of (3.1), we obtain, after integrating by parts,

$$\begin{aligned}
&\frac{1}{\tau}\int_0^1 e^{y_\infty}(e^{z_k} - e^{z_{k-1}})z_k dx = -\int_0^1 e^{z_k + y_\infty}(z_k + y_\infty)_{xx} z_{k,xx} dx \\
&= -\int_0^1 e^{z_k+y_\infty} z_{k,xx}^2 dx - \int_0^1 e^{z_k} z_{k,xx}(ax+b)dx \\
&= -\int_0^1 e^{z_k+y_\infty} z_{k,xx}^2 dx + \int_0^1 e^{z_k} z_{k,x}^2 (ax+b)dx + a\int_0^1 e^{z_k} z_{k,x} dx \\
&\le -\min\{u_0,u_1\}\int_0^1 e^{z_k} z_{k,xx}^2 dx,
\end{aligned} \tag{4.2}$$

since $ax + b \le 0$ in $(0,1)$ and $e^{z_k(x)} = 1$ for $x = 0,1$. The left-hand side is estimated from below by employing the elementary inequality $e^x \ge x + 1$ for all $x \in \mathbb{R}$:

$$\begin{aligned}
&\frac{1}{\tau}\int_0^1 e^{y_\infty}(e^{z_k} - e^{z_{k-1}})z_k dx \\
&= \frac{1}{\tau}\int_0^1 e^{z_k+y_\infty}(z_k - 1)dx - \frac{1}{\tau}\int_0^1 e^{z_{k-1}+y_\infty}(z_{k-1}-1)dx \\
&\quad + \frac{1}{\tau}\int_0^1 e^{z_{k-1}+y_\infty}(e^{z_k - z_{k-1}} + z_{k-1} - z_k - 1)dx \\
&\ge \frac{1}{\tau}\int_0^1 e^{z_k+y_\infty}(z_k - 1)dx - \frac{1}{\tau}\int_0^1 e^{z_{k-1}+y_\infty}(z_{k-1}-1)dx.
\end{aligned} \tag{4.3}$$

This shows that the sequence $E^{(k)} = \int_0^1 e^{y_\infty}(e^{z_k}(z_k - 1) + 1)dx$ is nonincreasing and bounded from below by $E^{(0)} = \int_0^1 (u_I(\log(u_I/u_\infty) - 1) + 1)dx$, which is finite.

We relate the entropy production term on the right-hand side of (4.2) to the entropy itself. We first claim that

$$\int_0^1 e^{z_k} z_{k,xx}^2 dx \ge 4\int_0^1 (e^{z_k/2})_{xx}^2 dx. \tag{4.4}$$

To see this we set $u = e^{z_k}$ and observe that an integration by parts yields

$$\int_0^1 \frac{u_{xx}u_x^2}{u^2}dx = \frac{2}{3}\int_0^1 \frac{u_x^4}{u^3}dx.$$

Then

$$\int_0^1 e^{z_k} z_{k,xx}^2 dx = \int_0^1 \left(\frac{u_{xx}^2}{u} - \frac{1}{3}\frac{u_x^4}{u^3} \right) dx \geq \int_0^1 \left(\frac{u_{xx}^2}{u} - \frac{5}{12}\frac{u_x^4}{u^3} \right) dx \tag{4.5}$$
$$= 4\int_0^1 (\sqrt{u})_{xx}^2 dx = 4\int_0^1 (e^{z_k/2})_{xx}^2 dx.$$

We need the Poincaré inequalities

$$\|u\|_{L^2(0,1)} \leq \frac{1}{\pi}\|u_x\|_{L^2(0,1)}, \quad \|u\|_{L^\infty(0,1)} \leq \|u_x\|_{L^2(0,1)}$$

for all $u \in H_0^1(0,1)$. Therefore, using Lemma 4.1, we infer

$$\int_0^1 e^{z_k} z_{k,xx}^2 dx \geq 4\pi^2 \int_0^1 (e^{z_k/2} - 1)_x^2 dx \geq 4\pi^2 \|e^{z_k/2} - 1\|_{L^\infty(0,1)}^2 \geq \frac{4\pi^2}{c_2} E^{(k)}. \tag{4.6}$$

Setting $\gamma = 4\pi^2 \min\{u_0, u_1\}/c_2$, we obtain from (4.2) the difference inequality

$$E^{(k)} \leq E^{(k-1)} - \gamma\tau E^{(k)},$$

from which

$$E^{(k)} \leq (1+\gamma\tau)^{-1} E^{(k-1)} \leq (1+\gamma\tau)^{-k} E^{(0)} \leq (1+\gamma\tau)^{-t/\tau} E^{(0)} \tag{4.7}$$

follows. The parameter γ depends on $\|e^{z_k}\|_{L^1(0,1)}$ through c_2. However, since $e^{z^{(N)}}$ is uniformly bounded in $L^\infty(0,T;L^1(0,1))$ in view of Lemma 3.3, γ is bounded uniformly in k. We have shown in the proof of Theorem 1.2 that $e^{z_k} \to e^z$ a.e. Then the uniform boundedness of e^{z_k} and z_k and Lebesgue's dominated convergence theorem imply that

$$E^{(k)} \to E_3(t) = \int_0^1 e^{y_\infty}(e^{z(\cdot,t)}(z(\cdot,t)-1)+1)dx.$$

Hence, after letting $\tau \to 0$, we conclude from (4.7) that $E_3(t) \leq E_3(0)e^{-\gamma t}$. The first inequality in (4.1) gives the assertion with $\lambda = \gamma/2$. $\square$

Remark 4.2. The decay rate λ is not optimal. For instance, we neglected the term $\int_0^1 u_x^4/12u^3 dx$ in (4.5) and the constants in (4.1) are not the best ones. For optimal constants in logarithmic Sobolev inequalities related to (1.1) with periodic boundary conditions, we refer the reader to [10].

Remark 4.3. It is not easy to find conditions on the boundary data for which $\log u_\infty$ is concave. An example is $u_0 = u_1$ and $w_0 = -w_1 \geq 0$. Indeed, if $y = \log u_\infty$, we have $y(0) = y(1)$ and $y_x(0) = -y_x(1) \geq 0$ and therefore, y is symmetric around $x = \frac{1}{2}$. Thus (see Remark 2.3) $a = e^{y_0}(y_{xx}(1) - y_{xx}(0)) = 0$ and $b = e^{y_0}y_{xx}(0) \leq 0$. This implies $(\log u_\infty)_{xx} = y_{xx} = be^{-y} \leq 0$ in $(0,1)$.

Remark 4.4. The assumption on the concavity of $\log u_\infty$ can be slightly relaxed. Indeed, we claim that the assertion of Theorem 1.3 also holds if $((\log u_\infty)_{xx})^+$ is small enough in the sense

$$4\,\frac{\max\{u_\infty(x) : 0 \leq x \leq 1\}}{\min\{u_\infty(x) : 0 \leq x \leq 1\}} \int_0^1 ((\log u_\infty)_{xx})^+ \, dx \leq 1 - \delta \tag{4.8}$$

for some $\delta > 0$, where $(x)^+ = \max\{0, x\}$. We prove this result by deriving a bound on the second integral in (4.2) in terms of the first one, employing the weighted Poincaré inequality [7, Thm. 1.4]

$$\int_0^1 u_x^2 \mu(x) dx \leq K \int_0^1 u_{xx}^2 dx$$

for all $u \in H^2(0,1)$ satisfying $u(0) = u(1)$ (which implies that $\int_0^1 u_x dx = 0$). The function μ is assumed to be nonnegative and measurable. The best constant $K > 0$ is not explicit but can be bounded by $K \leq 4 \int_0^1 \mu(x) dx$ [7, Rem. 1.10.4]. We choose $\mu(x) = (ax+b)^+ = (u_\infty (\log u_\infty)_{xx})^+$. Then the weighted Poincaré inequality and (4.4) give

$$\begin{aligned} \int_0^1 e^{z_k + y_\infty} z_{k,xx}^2 dx &\geq 4m \int_0^1 (e^{z_k/2})_{xx}^2 dx \geq \frac{4m}{K} \int_0^1 (e^{z_k/2})_x^2 \mu(x) dx \\ &= \frac{m}{K} \int_0^1 (ax+b)^+ e^{z_k} z_{k,x}^2 dx, \end{aligned}$$

where $m = \min\{u_\infty(x) : 0 \leq x \leq 1\}$. Inserting this inequality into (4.2) and using (4.3), we obtain

$$\begin{aligned} \frac{1}{\tau}\left(E^{(k)} - E^{(k-1)}\right) &\leq -\int_0^1 e^{z_k + y_\infty} z_{k,xx}^2 dx + \int_0^1 (ax+b)^+ e^{z_k} z_{k,x}^2 dx \\ &\leq \left(\frac{K}{m} - 1\right) \int_0^1 e^{z_k + y_\infty} z_{k,xx}^2 dx. \end{aligned}$$

Assumption (4.8) shows that $K/m \leq 1 - \delta$ and hence, by (4.6),

$$\frac{1}{\tau}\left(E^{(k)} - E^{(k-1)}\right) \leq -\delta \int_0^1 e^{z_k + y_\infty} z_{k,xx}^2 dx \leq -\frac{4\pi^2 \delta m}{c_2} E^{(k)}.$$

Now proceed as in the proof of Theorem 1.3. The convergence rate in the L^1 norm is given by $\lambda = 2\pi^2 \delta m / c_2$.

5. Numerical examples. In this section we show by numerical examples that the assumption of concavity of $\log u_\infty$ (or the assumption (4.8)), where u_∞ is the solution to (1.4), seems to be only technical. Equation (1.1) is solved numerically in the formulation

$$u_t = -u_{xxxx} + \left(\frac{u_x^2}{u}\right)_{xx} \quad \text{in } (0,1). \tag{5.1}$$

We use a uniform grid $(x_i, t_j) = (\triangle x \cdot i, \triangle t \cdot j)$ with spatial mesh size $\triangle x = 10^{-3}$ and time step $\triangle t = 10^{-6}$. With the approximation u_{ij} of $u(x_i, t_j)$, the fully implicit discretization reads as

$$\frac{1}{\triangle t}(u_{ij} - u_{i,j-1}) = -D^+ D^- D^+ D^- u_{ij} + D^+ D^- \left(\frac{(D^+ u_{ij})^2}{u_{ij}}\right),$$

where D^+ and D^- are the forward and backward difference operators on the spatial mesh (see [13]). The nonlinear equations are solved on each time level by Newton's method where the initial guess is chosen to be the solution of the previous time level.

For the first example we use the boundary conditions

$$u(0,t) = u_0, \quad u(1,t) = u_1, \tag{5.2}$$

$$u_x(0,t) = w_0 = 2\sqrt{u_0}(\sqrt{u_1} - \sqrt{u_0}), \quad u_x(1,t) = w_1 = 2\sqrt{u_1}(\sqrt{u_1} - \sqrt{u_0}), \tag{5.3}$$

with $u_0 \le u_1$. The advantage of these conditions is that the stationary problem (1.4) has the exact solution

$$u_\infty(x) = \left((\sqrt{u_1} - \sqrt{u_0})x + \sqrt{u_0}\right)^2, \quad x \in (0,1).$$

We choose the initial condition $u_I(x) = e^{-x}\sin(3\pi x) + 3x + 1$ and the boundary values $u_0 = 1$ and $u_1 = 4$. The numerical solution at various times is displayed in Figure 5.1. The discrete solution seems to converge to the exact solution u_∞ as $t \to \infty$. Figure 5.2 shows the exponential decay of the relative entropy

$$E_3(t) = \int_0^1 u(\cdot,t)((\log(u(\cdot,t)/u_\infty) - 1) + u_\infty)dx$$

and of the L^1 deviation $\|u(\cdot,t) - u_\infty\|_{L^1(0,1)}$. As predicted by the proof of Theorem 1.3, the decay rate of the L^1 deviation is half of the rate of the relative entropy. Notice that the function $\log u_\infty$ is concave; i.e., the assumptions of Theorem 1.3 are satisfied.

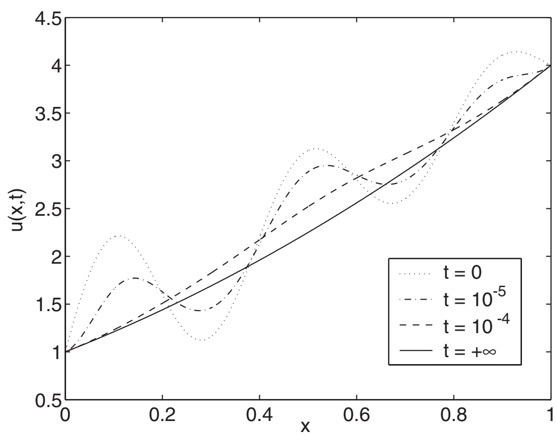


FIG. 5.1. *Numerical solution to* (5.1)–(5.3) *with* $u_0 = 1$, $u_1 = 4$, $w_0 = 2$, *and* $w_1 = 4$ *at various times.*

In the second example we show by a numerical example that the solution to (1.1) decays exponentially fast even if the function $\log u_\infty$ is convex. For this we choose the boundary conditions $u_0 = 1.5$, $u_1 = 0.8$, $w_0 = -4.6127$, and $w_1 = 2.0618$. The stationary solution u_∞ is computed numerically from the equation

$$u_\infty(\log u_\infty)_{xx} = ax + b, \quad x \in (0,1),$$

where $a = 1$ and $b = 3$. Then, $\log u_\infty$ is strictly convex in $(0,1)$ and the assumption (4.8) is not satisfied. We choose the initial function $u_I(x) = -e^{-x}\sin(2\pi x) - \frac{7}{10}x + \frac{3}{2}$. Figure 5.3 shows the discrete solution for various times. In this case, the relative entropy and the L^1 deviation are also exponentially decaying (Figure 5.4) although the condition of Theorem 1.3 is not satisfied. This suggests that the concavity hypothesis is purely technical.

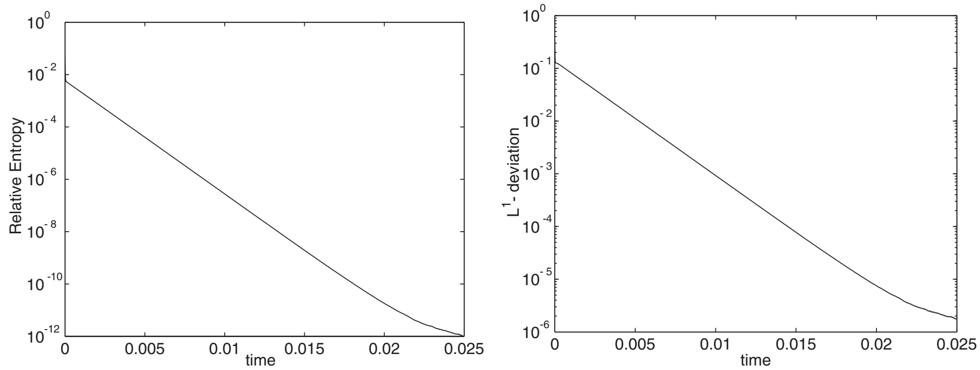


FIG. 5.2. *Logarithmic plot of the relative entropy* $E_3(t)$ *(left) and the* L^1 *deviation* $\|u(\cdot,t) - u_\infty\|_{L^1(0,1)}$ *(right) for the solution to* (5.1)–(5.3) *with* $u_0 = 1$, $u_1 = 4$.

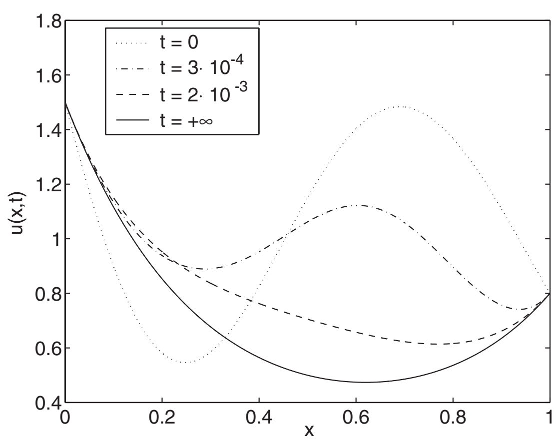


FIG. 5.3. *Numerical solution to* (5.1), (1.3) *with* $u_0 = 1.5$, $u_1 = 0.8$, $w_0 = -4.6127$, *and* $w_1 = 2.0618$ *at various times.*

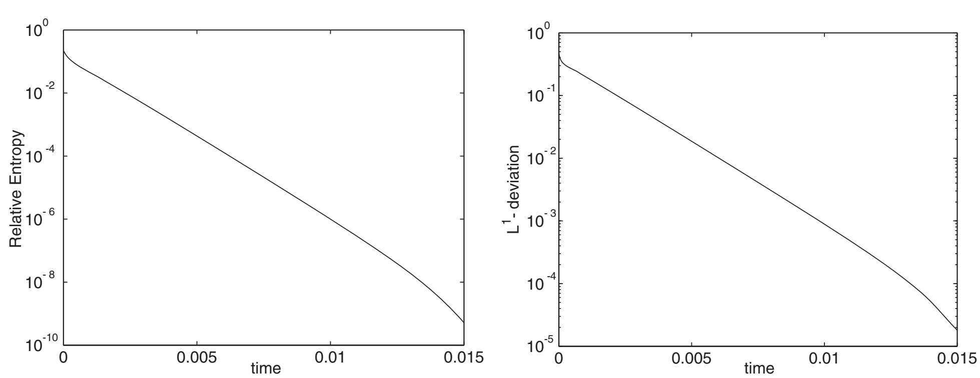


FIG. 5.4. *Logarithmic plot of the relative entropy* $E_3(t)$ *(left) and the* L^1 *deviation* $\|u(\cdot,t) - u_\infty\|_{L^1(0,1)}$ *(right) for the solution to* (5.1), (1.3) *with* $u_0 = 1.5$, $u_1 = 0.8$, $w_0 = -4.6127$, *and* $w_1 = 2.0618$.

Acknowledgment. A. Jüngel would like to thank Michel Fournié (Toulouse, France) for fruitful discussions.

REFERENCES

[1] M. Ancona, *Diffusion-drift modeling of strong inversion layers*, COMPEL, 6 (1987), pp. 11–18.

[2] J. Bebernes and D. Eberly, *Mathematical Problems from Combustion Theory*, Springer-Verlag, Berlin, 1989.

[3] M. Bertsch, R. Dal Passo, G. Garcke, and G. Grün, *The thin viscous flow equation in higher space dimensions*, Adv. Differential Equations, 3 (1998), pp. 417–440.

[4] P. Bleher, J. Lebowitz, and E. Speer, *Existence and positivity of solutions of a fourth-order nonlinear PDE describing interface fluctuations*, Comm. Pure Appl. Math., 47 (1994), pp. 923–942.

[5] M. Cáceres, J. Carrillo, and G. Toscani, *Long-time behavior for a nonlinear fourth order parabolic equation*, Trans. Amer. Math. Soc., 357 (2005), pp. 1161–1175.

[6] J. A. Carrillo, A. Jüngel, and S. Tang, *Positive entropic schemes for a nonlinear fourth-order equation*, Discrete Contin. Dyn. Syst. Ser. B, 3 (2003), pp. 1–20.

[7] S.-K. Chua and R. Wheeden, *Sharp conditions for weighted* 1*-dimensional Poincaré inequalities*, Indiana Univ. Math. J., 49 (2000), pp. 143–175.

[8] P. Degond, F. Méhats, and C. Ringhofer, *Quantum hydrodynamic models derived from the entropy principle*, Contemp. Math., 371 (2005), pp. 107–131.

[9] B. Derrida, J. Lebowitz, E. Speer, and H. Spohn, *Fluctuations of a stationary nonequilibrium interface*, Phys. Rev. Lett., 67 (1991), pp. 165–168.

[10] J. Dolbeault, I. Gentil, and A. Jüngel, *A Nonlinear Fourth-Order Parabolic Equation and Related Logarithmic Sobolev Inequalities*, preprint, Universität Mainz, Mainz, Germany, 2004.

[11] H. Fujita, *On the nonlinear equation* $\Delta u + e^u = 0$ *and* $\frac{\partial u}{\partial t} = \Delta u + e^u$, Bull. Amer. Math. Soc., 75 (1969), pp. 132–135.

[12] U. Gianazza, G. Savaré, and G. Toscani, *A Fourth-Order Nonlinear PDE as Gradient Flow of the Fisher Information in Wasserstein Spaces*, in preparation, Università di Pavia, Pavia, Italy, 2005.

[13] A. Jüngel and R. Pinnau, *Global nonnegative solutions of a nonlinear fourth-order parabolic equation for quantum systems*, SIAM J. Math. Anal., 32 (2000), pp. 760–777.

[14] A. Jüngel and R. Pinnau, *A positivity-preserving numerical scheme for a nonlinear fourth order parabolic equation*, SIAM J. Numer. Anal., 39 (2001), pp. 385–406.

[15] A. Jüngel and G. Toscani, *Exponential time decay of solutions to a nonlinear fourth-order parabolic equation*, Z. Angew. Math. Phys., 54 (2003), pp. 377–386.

[16] R. Pinnau, *A note on boundary conditions for quantum hydrodynamic equations*, Appl. Math. Lett., 12 (1999), pp. 77–82.

[17] J. Simon, *Compact sets in the space* $L^p(0,T;B)$, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1780–1799

ON THE EXISTENCE AND UNIQUENESS OF WEAK SOLUTIONS FOR A VORTICITY SEEDING MODEL*

LUIGI C. BERSELLI† AND MARCO ROMITO‡

Abstract. In this paper we study the Navier–Stokes equations with a Navier-type boundary condition that has been proposed as an alternative to common near wall models. The boundary condition we study, involving a linear relation between the tangential part of the velocity and the tangential part of the Cauchy stress-vector, is related to the vorticity seeding model introduced in the computational approach to turbulent flows. The presence of a pointwise nonvanishing normal flux may be considered as a tool to avoid the use of phenomenological near wall models in the boundary layer region. Furthermore, the analysis of the problem is suggested by recent advances in the study of large eddy simulation.

In the two-dimensional case, by using rather elementary tools, we prove existence and uniqueness of weak solutions. The asymptotic behavior of the solution, with respect to the averaging radius δ, is also studied. In particular, we prove convergence of the solutions toward the corresponding solutions of the Navier–Stokes equations with the usual no-slip boundary conditions, as the small parameter δ goes to zero.

Key words. Navier–Stokes equations, boundary models for turbulent flows, existence, uniqueness, LES models

AMS subject classifications. 35Q30, 76F75, 76B03

DOI. 10.1137/04061249X

1. Introduction.

In this paper we consider the Navier–Stokes equations and in particular the role of boundary conditions in the simulation of boundary effects in turbulent flows. We consider the Navier–Stokes equations (in nondimensional form) for viscous incompressible fluids in a bounded smooth domain $\Omega \subset \mathbb{R}^n$, $n = 2, 3$:

$$\begin{cases} \partial_t u - \dfrac{2}{Re}\nabla\cdot\mathbb{D}(u) + (u\cdot\nabla)u + \nabla p = 0 \quad \text{in } \Omega\times(0,T), \\ \nabla\cdot u = 0 \quad \text{in } \Omega\times(0,T). \end{cases} \tag{1.1}$$

We recall that $u = (u_1, \ldots, u_n)$ is the unknown velocity field, p is the hydrostatic pressure, $Re > 0$ is the Reynolds number, and $\mathbb{D}(u)$ is the deformation tensor, i.e., the symmetric part of the matrix of derivatives of u,

$$\mathbb{D}(u) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_k} + \frac{\partial u_k}{\partial x_i}\right),$$

and the Navier–Stokes equations are generally equipped with the *no-slip* boundary conditions on $\Gamma = \partial\Omega$,

$$u = 0 \quad \text{on } \Gamma\times(0,T). \tag{1.2}$$

*Received by the editors July 29, 2004; accepted for publication (in revised form) August 4, 2005; published electronically February 21, 2006.
http://www.siam.org/journals/sima/37-6/61249.html
†Dipartimento di Matematica Applicata "U. Dini," Università di Pisa, Via B. Pisano 25/b, I-56126 Pisa, Italy (berselli@dma.unipi.it).
‡Dipartimento di Matematica, Università di Firenze, Viale Morgagni 67/a, I-50134 Firenze, Italy (romito@math.unifi.it).

To introduce the problem that we shall study, we recall that while at a free surface it is natural to require continuity of the stress-tensor ($\mathbb{I}$ being the identity)

$$\mathbb{T}(u,p) = -p\,\mathbb{I} + \frac{2}{Re}\mathbb{D}(u),$$

the conditions at a solid wall are very challenging. The no-slip condition (1.2) has been justified by Stokes [37] since the contrary assumption "... *implies an infinitely greater resistance to the sliding of one portion of fluid past another than to the sliding of fluid over a solid*."

It is well known that there are situations in which the boundary condition (1.2) may not be valid. For instance, in Serrin [35, sect. 64] it is pointed out that in high altitude aerodynamics, or, more generally, when moderate pressure and low surface stresses are involved, the adherence condition is no longer true; see also the review in Truesdell [38]. In this respect several authors proposed various *slip* (generally nonlinear) conditions, modeling precise physical situations; see, for instance, Serrin [35], Beavers and Joseph [4], and Kreĭn and Laptev [23].

From the historical point of view, the slip (with friction) boundary condition proposed by Navier [30] was

$$u \cdot \underline{\mathrm{n}} = 0 \quad \text{and} \quad \beta\, u_\tau + \underline{\mathcal{T}}(u,p) = 0, \quad \beta > 0, \quad \text{on } \Gamma \times (0,T), \tag{1.3}$$

where $\underline{\mathrm{n}}$ denotes the unit normal vector to Γ, $u_\tau = u-(u\cdot\underline{\mathrm{n}})\underline{\mathrm{n}}$, while $\underline{\mathcal{T}}(u,p) = \underline{t}(u,p) - (\underline{t}(u,p)\cdot\underline{\mathrm{n}})\,\underline{\mathrm{n}}$ denotes the tangential part of the *Cauchy stress vector* $\underline{t}$ defined by

$$\underline{t}(u,p) = \underline{\mathrm{n}} \cdot \mathbb{T}(u,p) = \sum_{k=1}^{n} \mathbb{T}_{ik}(u,p)\underline{\mathrm{n}}_k.$$

Maxwell [27] analyzed the two types of boundary conditions (condition (1.3) proposed by Navier and condition (1.2) proposed by Stokes), observing that the same conditions may be derived within the kinetic theory of gases and that the parameter β should depend on the Reynolds number Re and on the mean free-path λ, satisfying the pair of consistency conditions

$$\begin{aligned} &\beta \to \infty \quad \text{as } \lambda \to 0 \text{ for } Re \text{ fixed},\\ &\beta \to 0 \quad \text{as } Re \to \infty \text{ for } \lambda \text{ fixed}. \end{aligned}$$

With the above asymptotics it possible to recover in both cases the correct no-slip boundary condition for viscous fluids and the no-penetration condition for ideal fluids. A study of the numerical problems related to the implementation of (1.3) can be found in John [20].

Recently, Fujita [12] performed the analysis with the "slip or leak with friction" boundary conditions. These conditions are of particular interest in the study of polymers, blood flow, and flow through filters. The boundary conditions studied in [12], with the techniques of variational inequalities, turn out to be particular cases of the nonlinear boundary condition proposed in [35, p. 240] and they are very strictly related to both the Navier and the no-slip boundary conditions. See also Consiglieri [9] for similar problems.

For laminar flows the Navier boundary condition (1.3) also appears in the presence of rough boundaries; see Jäger and Mikelić [18, 19] and Achdou, Pironneau, and Valentin [1]. In the modeling of turbulent flows it appears (with several variants, see

Sagaut [34]) in connection with conventional turbulence models and recently also in large eddy simulation (LES) models (see Layton [24]), as we shall introduce in the next section.

Among other nonstandard boundary conditions we recall those studied by Begue et al. [5] and the "do-nothing" Neumann conditions, very appealing for numerical studies, implemented in Heywood, Rannacher, and Turek [16].

1.1. Near wall models and turbulent flows. Our interest in nonstandard boundary conditions comes essentially from the study of turbulent flows. First, we recall that in the boundary layer theory several log-law and power-law boundaries with asymptotics are introduced, together with the fictitious boundaries, in order to model turbulent flows within a small region near the boundary. Roughly speaking, appropriate nonlinear boundary conditions are imposed on an artificial boundary that lies inside the computational domain. The boundary conditions may simulate (at least in a computational approach) the behavior of the boundary layer, and they are modeled to take into account the peculiar behavior of a fluid near the boundaries. In this respect we recall that Maxwell [27] observed " … *it is almost certain that the stratum of gas next to a solid body is in a very different state from the rest of the gas."*

Our main motivation comes from the mathematical theory of LES. In fact, the purpose of LES is to model the evolution of large coherent structures (eddies); this is done by studying the equations satisfied by a filtered velocity. Generally, the filtered velocity $\overline{u}$ is defined through a convolution

$$\overline{u}(x,t) = g_\delta(x) * u(x,t)$$

with a rapidly decreasing smoothing kernel g_δ of width δ; in several cases of practical interest g_δ is a Gaussian, i.e.,

$$g_\delta(x) = \left(\frac{\gamma}{\pi}\right)^{3/2} \frac{1}{\delta^3}\, \mathrm{e}^{-\frac{\gamma|x|^2}{\delta^2}}.$$

By definition, the value of $\overline{u}$ at a point x_0 on the boundary Γ will depend on the behavior of u in a neighborhood of width δ near that point; even if u is extended to zero for each $x \notin \overline{\Omega}$, it is clear that in general $\overline{u}(x_0) \neq 0$.

As pointed out in Galdi and Layton [15] the physical intuition may suggest that large coherent structures touching a wall do not penetrate, but instead slide along the wall and lose their energy. The boundary condition of Navier may be revisited by linking the microscale λ of the kinetic theory of gases with the radius δ of the averaging filter.

Many near wall models have been tested in the computational approach; see Sagaut [34] and Piomelli and Balaras [33]. The results are not uniformly successful, and a positive application is very often based on a fine tuning of parameters. This is why new models require at least a positive background from the physical hypotheses and a coherent mathematical analysis. In particular, a successful application of the Navier slip-with-friction boundary condition (1.3) is prevented by two main facts: (1) the presence of recirculation regions and (2) the presence of fast time-fluctuating quantities.

The first problem is motivated by the fact that in recirculation regions the local Reynolds number is very different from the main stream, and it is natural to expect that β should depend (possibly in a nonlinear way) on a local Reynolds number related to the local slip speed, i.e.,

$$\beta = \beta(\delta,\, |u_\tau|).$$

Preliminary analysis has been performed by John, Layton, and Sahin [21] and Dunca et al. [11], and an appropriate power-law choice of β seems promising to improve the estimation of reattachment points.

The limitation of Navier law (1.3) in a boundary layer theory is that it can well describe time-averaged flow profiles, while the information coming from fluctuating quantities in the wall-normal direction can play an important role in triggering separation and detachment. To try to overcome this limitation, very recently Layton [24] recognized a particular class of boundary conditions, leading to conditions similar "in spirit" to the so-called *vorticity seeding methods*. In fact, a Navier slip-with-friction boundary condition implies the generation of vorticity at the boundary, proportional to the tangential velocity. More precisely, in the case of a two-dimensional domain Ω, for each smooth function v such that $v \cdot \underline{\mathrm{n}} = 0$ on the boundary, it holds that

$$\underline{\mathrm{n}} \cdot \mathbb{D}(v) \cdot \underline{\tau} - \frac{1}{2}\operatorname{curl} v + k(v \cdot \underline{\tau}) = 0 \qquad \text{on } \Gamma,$$

where $\operatorname{curl} v = \partial v_1/\partial x_2 - \partial v_2/\partial x_1$, and $\underline{\tau}$ denotes the unit tangent vector, while k is the curvature of Γ.

In particular, in [24] the following boundary condition is proposed to simulate the boundary effects

$$u \cdot \underline{\mathrm{n}} = \delta^2 g(x,t) \qquad \text{and} \qquad \frac{L}{\delta Re} u_\tau + \underline{\mathcal{T}}(u,p) = 0, \tag{1.4}$$

where g is a *highly oscillating* function in the time variable (hopefully a random variable in numerical tests), while it may be very smooth in the space variables and should satisfy the natural compatibility condition

$$\int_\Gamma g(x,t)\, d\sigma = 0 \qquad \forall\, t \in (0,T), \tag{1.5}$$

which is required by the normal trace of a divergence-free vector field.

This way of reasoning is also similar to the introduction of stochastic fluctuations to simulate the microscale effects. A comprehensive introduction to stochastic partial differential equations in fluid mechanics and the statistical approach can be found in Monin and Yaglom [28], and one main mathematical paradigm is that an additional nonsmooth term on the right-hand side may naturally take into account the effect of the fast fluctuating quantities. This leads us to study the system

$$\partial_t u - \frac{2}{Re}\nabla \cdot \mathbb{D}(u) + (u \cdot \nabla)u + \nabla p = f + \frac{\partial g}{\partial t},$$

where g is a function that does not have a proper time derivative, but is just continuous or with other weak properties. For the above problems the study of an appropriate notion of solution, together with the corresponding statistical properties, started in Bensoussan and Temam [7] and Višik and Fursikov [39].

1.2. Setting of the problem. In what follows we shall restrict our discussion to the two-dimensional case, since the nonlinear character of the equations imposes some restriction; see Remark 2.5. Furthermore, we fix the values of both L and the Reynolds number to 2, due to the fact that we will not deal with the singular limit $Re \to \infty$. (Regarding this limit see also the recent works of Clopeau, Mikelić, and Robert [8], Lopes Filho, Nussenzveig Lopes, and Planas [26], Mucha [29], and Kelliher [22].) In our case we shall study the following boundary-initial value problem:

$$
\begin{cases}
\partial_t u - \nabla \cdot \mathbb{D}(u) + (u \cdot \nabla)u + \nabla p = f & \text{in } \Omega \times (0,T), \\
\nabla \cdot u = 0 & \text{in } \Omega \times (0,T), \\
u \cdot \underline{\mathrm{n}} = \delta^\alpha g(x,t) & \text{on } \Gamma \times (0,T), \\
u \cdot \underline{\tau} + \delta\, \underline{\mathrm{n}} \cdot \mathbb{D}(u) \cdot \underline{\tau} = 0 & \text{on } \Gamma \times (0,T), \\
u(x,0) = u_0(x) & \text{in } \Omega.
\end{cases}
\tag{1.6}
$$

Remark 1.1. The above problem with $\alpha = 2$ describes the experiment in [4], where δ represents the characteristic pore size and the system is laminar. The error estimate we derive is consistent with the first step in the homogenization procedure employed by Jäger and Mikelić [17] to obtain the law of Beavers and Joseph.

Note that a similar problem, but involving the Smagorinsky–Ladyženskaya turbulence model together with a nonlinear dependence of β on u, has been studied, for instance, by Parés [32], but in that reference the normal datum g is not allowed to depend on the time variable. In what follows our main interest will be to find weak hypotheses on $g(x,t)$ with respect to the time variable (without any essential restriction on the space regularity) that allow us to prove existence of weak solutions to the Navier–Stokes equations; see Theorem 1.2. In particular, in our analysis we will focus on two main points: (1) to show the existence of weak solutions in the sense of Leray and Hopf (since we do not want to deal with any weaker concept of solution) and (2) to use only elementary tools of functional analysis.

In other words, we want to consider solutions in a very standard sense and we also want to interact with applied people interested in this problem, while still keeping all the mathematical rigor needed to deal with the problem and a certain sharpness of the results. In the case of nonhomogeneous no-slip conditions, several results of existence and uniqueness of other classes of solutions can be found in Amann [3].

Our intent to have a nonvanishing normal datum can be heuristically understood also with the following argument: Suppose that (for simplicity in two dimensions) we have a fictitious boundary Γ_1 and we want to impose a condition on it in order to resolve numerically the equation in a smaller domain $\Omega_1 \subset \Omega$ that rules out the boundary layer (see Figure 1 below).

We have to require, by the incompressibility of the flow, that

$$
\int_{\{ABCD\}} \nabla \cdot u \, dx = \int_{\partial\{ABCD\}} u \cdot \underline{\mathrm{n}} \, d\sigma = 0
$$

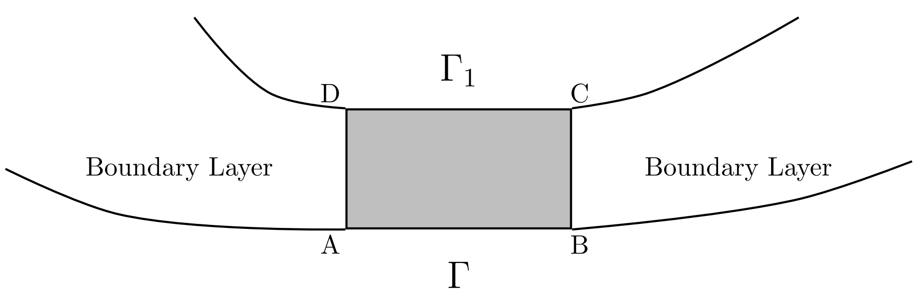


FIG. 1. *The fictitious boundary.*

for each (also curvilinear or infinitesimal) "rectangle" $\{ABCD\}$ touching the boundary Γ as in the figure. Since the behavior of the flow is not known, in general we have

$$\int_{\{CD\}} u \cdot \underline{\mathrm{n}}\, ds = -\left[\int_{\{BC\}} u \cdot \underline{\mathrm{n}}\, ds + \int_{\{DA\}} u \cdot \underline{\mathrm{n}}\, ds\right] \neq 0,$$

while the line integral over the segment $\{AB\}$ vanishes, since on the "true boundary" Γ both the Navier and no-slip conditions impose that $u \cdot \underline{\mathrm{n}} = 0$.

This may justify the introduction of a nonvanishing normal flux, also with very low regularity properties, namely, the same shared by the trace of a turbulent flow in the boundary layer region.

1.3. Main results. In this section we briefly enunciate the results we shall prove, together with their precise and rigorous statement.

In what follows Ω will denote a bounded, connected, open set in $\mathbb{R}^2$, locally situated on one side of its boundary Γ, a manifold of (at least) class $C^{1,1}$ (Lipschitz-continuous first derivatives). The existence of the unit outward normal $\underline{\mathrm{n}}$ derives by results proved in Nečas [31].

The first result we shall prove is an existence and uniqueness theorem for weak solutions of the Navier–Stokes system (1.6), with boundary conditions (1.4).

THEOREM 1.2. *Assume $g \in H^{\frac{1}{2}+\varepsilon}(0,T;H^{\frac{1}{2}}(\Gamma))$, for some $\varepsilon > 0$ satisfying the compatibility condition* (1.5)*; $f \in L^2((0,T)\times\Omega)$; and $u_0 \in L^2(\Omega)$, with $\nabla \cdot u_0 = 0$. Then there exists a unique weak solution*

$$u \in L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;H^1(\Omega)),$$

of problem (1.6)*.*

Next, we want to study the behavior of the solution to problem (1.6) as the small parameter δ converges to zero. (Other convergence results, under similar assumptions, have been also proved in [22].) In view of the considerations of the previous section, one can expect that, as the boundary layer becomes thinner and thinner, the solutions will look more and more like the classical solutions corresponding to the *no-slip* boundary condition. Indeed, this is the case, as shown by Theorem 1.3.

Let u_δ be the solution of (1.6) (we emphasize the dependence on δ in this framework) and let v be the solution to the Navier–Stokes equations with the same initial value and no-slip boundary conditions. (To be more precise, the vector field v is the solution to system (3.2).) As we shall see in section 3,

$$u_\delta = v + \mathcal{O}(\delta^{\frac{1}{3}}),$$

so that the "no-slip solution" represents the average behavior, once one neglects the effect at the boundary. The term $u_\delta - v$ can be seen as the "fluctuation term," which takes into account the nontrivial dynamics at the boundary.

THEOREM 1.3. *Assume $u_0 \in H^1(\Omega)$, with $\nabla \cdot u_0 = 0$, $g \in H^{\frac{1}{2}+\varepsilon}(0,T;H^{\frac{1}{2}}(\Gamma))$ satisfying the compatibility condition* (1.5)*, and $f \in L^2((0,T)\times\Omega)$. Then*

$$\sup_{0\le t\le T} \|u_\delta - v\|^2 + \int_0^T \left(\|\mathbb{D}(u_\delta - v)\|^2 + \frac{1}{\delta}\|(u_\delta - v)\cdot \underline{\tau}\|_\Gamma^2\right) dt = \mathcal{O}(\delta^{\frac{2}{3}}).$$

In particular, u_δ converges to v in $L^\infty(0,T;L^2(\Omega))$ and $L^2(0,T;H^1(\Omega))$.

2. A result of existence and uniqueness of weak solutions. In this section we prove Theorem 1.2. For the sake of simplicity, we consider the normal datum as

$$u \cdot \underline{\mathrm{n}} = g(x, t);$$

i.e., we drop the dependence on δ, since it is not relevant in view of the existence and uniqueness result we are going to show. In the last section we shall see how to deal with a right-hand side that scales by a power of δ.

Let us consider the evolution problem

$$\begin{cases} \partial_t u - \nabla \cdot \mathbb{D}(u) + (u \cdot \nabla)\, u + \nabla p = f & \text{in } \Omega \times (0, T), \\ \nabla \cdot u = 0 & \text{in } \Omega \times (0, T), \\ u \cdot \underline{\mathrm{n}} = g(x, t) & \text{on } \Gamma \times (0, T), \\ \delta\, \underline{\mathrm{n}} \cdot \mathbb{D}(u) \cdot \underline{\tau} + u \cdot \underline{\tau} = 0 & \text{on } \Gamma \times (0, T), \\ u(x, 0) = u_0(x) & \text{in } \Omega, \end{cases} \tag{2.1}$$

with $f \in L^2((0, T) \times \Omega)$ and $\nabla \cdot f = 0$, just to avoid technicalities, and with g not very smooth, say

$$g \in H^{1/2+\epsilon}(0, T; H^{1/2}(\Gamma)), \tag{2.2}$$

satisfying the compatibility condition (1.5).

2.1. Function spaces. We use the classical Lebesgue spaces and in particular, we will work exclusively in the Hilbert framework and thus use the space L^2. For simplicity we do not distinguish between space of scalar, vector, or either tensor valued functions, and the symbol $\|\,.\,\|$ will denote the norm in $L^2(\Omega)$. The norm in $L^2(\Gamma)$ will be denoted by $\|\,.\,\|_\Gamma$.

In what follows we shall use the customary Sobolev spaces, for which we refer to Adams [2], and the notion of "trace" over the boundary Γ of Ω. Mainly we shall use the space $H^1(\Omega)$ with norm denoted by $\|\,.\,\|_{H^1}$ and its trace space $H^{1/2}(\Gamma)$, with norm $\|\,.\,\|_{H^{\frac{1}{2}}(\Gamma)}$.

In addition, we define the spaces

$$H = \{\, u \in L^2(\Omega) \,|\, \nabla \cdot u = 0,\ u \cdot \underline{\mathrm{n}} = 0 \text{ on } \Gamma \,\},$$

and

$$V = \{\, u \in H^1(\Omega) \,|\, \nabla \cdot u = 0,\ u \cdot \underline{\mathrm{n}} = 0 \text{ on } \Gamma \,\},$$

and we endow V with the norm $\|u\|_V = \|\nabla u\|$. Moreover, we define the space of tangential vector fields as

$$H^1_\tau = \{\, u \in H^1(\Omega) \,|\, u \cdot \underline{\mathrm{n}} = 0 \text{ on } \Gamma \,\}.$$

2.1.1. Fractional derivative. In order to properly define the spaces we shall use, we also need to define fractional derivatives. The fractional derivative may be defined through singular integrals

$$D^\alpha_t U(x, t) = \frac{d}{dt} \int_0^t \frac{U(s, x)}{(t - s)^\alpha}\, ds \qquad \text{for } 0 \le \alpha < 1,$$

but for our purposes it is better to deal with a definition via the Fourier transform.

Given $\phi(x,t)$, defined for $t \in [0,T]$, with values in the Hilbert space $(\mathbb{X}, \|\,.\,\|_{\mathbb{X}})$ and integrable (in the Bochner sense), we define

$$\widetilde{\phi}(t,x) = \begin{cases} \phi(t,x) & \text{for } t \in [0,T], \\ 0 & \text{elsewhere,} \end{cases} \tag{2.3}$$

and its Fourier transform (with respect to the time variable) is

$$\widehat{\phi}(x,\xi) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widetilde{\phi}(x,t)\, \mathsf{e}^{-\mathtt{i}t\xi}\, dt,$$

so that we can define the fractional Sobolev spaces of functions having α-order derivative in L^2:

$$H^{\alpha}(\mathbb{R};\mathbb{X}) := \left\{ f \in L^2(\mathbb{R};\mathbb{X}) : \int_{\mathbb{R}} |\xi|^{2\alpha} \|\widehat{f}(\xi)\|^2_{\mathbb{X}}\, d\xi < \infty \right\}.$$

2.2. The linear stationary problem. The first step in solving (2.1) is to consider the linear stationary problem

$$\begin{cases} -\nabla \cdot \mathbb{D}(G) + \nabla \Pi = 0 & \text{in } \Omega \times (0,T), \\ \nabla \cdot G = 0 & \text{in } \Omega \times (0,T), \\ G \cdot \underline{\mathrm{n}} = g(x,t) & \text{on } \Gamma \times (0,T), \\ \delta\, \underline{\mathrm{n}} \cdot \mathbb{D}(G) \cdot \underline{\tau} + G \cdot \underline{\tau} = 0 & \text{on } \Gamma \times (0,T), \end{cases} \tag{2.4}$$

where the time variable is now just a parameter.

THEOREM 2.1. *Let the following be given: $g \in H^{1/2+\epsilon}(0,T;H^{1/2}(\Gamma))$, satisfying the compatibility condition* (1.5). *Then, there exists a unique G solution of* (2.4) *such that*

$$G(x,t) \in H^{1/2+\epsilon}(0,T;H^1(\Omega)). \tag{2.5}$$

Moreover, there is a constant C_0, depending only on Ω, such that

$$\|\nabla G\| + \|\Pi\| \leq C_0(1+\delta^{-\frac{1}{2}})\|g\|_{H^{\frac{1}{2}}(\Gamma)}. \tag{2.6}$$

Proof. See Solonnikov and Ščadilov [36] and Beirão da Veiga [6]. In fact, for each t it is possible to solve a linear stationary Stokes problem (with the appropriate boundary conditions) that has a unique solution belonging to $H^1(\Omega)$. The regularity in the time variable is inherited by the function G.

We give a formal (but completely justified) argument for the estimate (2.6), following the approach to the existence in L^2-spaces introduced in Beirão da Veiga [6] to find appropriate estimates on G. We consider the bilinear form

$$B(u,\phi) = \int_{\Omega} \mathbb{D}(u)\mathbb{D}(\phi)\, dx,$$

and the functions (G,Π) that solve (2.4) must satisfy

$$B(G,\phi) - \int_{\Omega} \Pi\, \nabla \cdot \phi\, dx + \frac{1}{\delta} \int_{\Gamma} G \cdot \phi\, d\sigma = 0 \qquad \forall\, \phi \in H^1_{\tau}.$$

In order to deal with the inhomogeneous problem, we introduce a vector field G_1 such that

$$\begin{cases} \nabla \cdot G_1 = 0 & \text{in } \Omega, \\ G_1 \cdot \underline{\mathrm{n}} = g & \text{on } \Gamma, \\ \|G_1\|_{H^1} \leq C\|g\|_{H^{\frac{1}{2}}(\Gamma)}. \end{cases}$$

The construction of such a vector field is rather standard and can be found, for instance, in Galdi [14].

By defining $G = G_1 + G_2$, the function G_2 must satisfy

$$B(G_2, \phi) - \int_\Omega \Pi\, \nabla \cdot \phi\, dx + \frac{1}{\delta}\int_\Gamma G_2 \cdot \phi\, d\sigma = -B(G_1, \phi) - \frac{1}{\delta}\int_\Gamma G_1 \cdot \phi\, d\sigma$$

for each ϕ tangential to the boundary. If $\phi = G_2$, we get (since $G_2 \cdot \underline{\mathrm{n}} = 0$ and $\nabla \cdot G_2 = 0$)

$$\|\nabla G_2\|^2 + \frac{1}{\delta}\|G_2\|_\Gamma^2 \leq \|\nabla G_2\|\, \|\nabla G_1\| + \frac{1}{\delta}\|G_1\|_\Gamma \|G_2\|_\Gamma$$

and, consequently,

$$\frac{1}{2}\|\nabla G_2\|^2 + \frac{1}{2\delta}\|G_2\|_\Gamma^2 \leq \frac{1}{2}\|\nabla G_1\|^2 + \frac{1}{2\delta}\|G_1\|_\Gamma^2.$$

This finally implies that

$$\|\nabla G_2\|^2 \leq C\left(1 + \frac{1}{\delta}\right)\|g\|^2_{H^{\frac{1}{2}}(\Gamma)},$$

where the constant C depends on Ω but is independent of δ. The estimate on the pressure can be obtained by approximation, by studying a slightly different equation; see [6] for details. □

Indeed, notice that, for our aims, it is enough to have

$$G(x,t) \in H^{1/2+\epsilon}(0,T; L^2(\Omega)) \cap L^2(0,T; H^1(\Omega)), \tag{2.7}$$

but currently we do not know the minimal assumption on g in order to have the above regularity. Regarding the usual no-slip boundary conditions, see the result proved by Fursikov, Gunzburger, and Hou [13].

2.3. The linear evolution problem. The next step for the analysis of the nonlinear evolution problem (2.1) is the following linear evolution problem:

$$\begin{cases} \partial_t z - \nabla \cdot \mathbb{D}(z) + \nabla q = 0 & \text{in } \Omega \times (0,T), \\ \nabla \cdot z = 0 & \text{in } \Omega \times (0,T), \\ z \cdot \underline{\mathrm{n}} = g & \text{on } \Gamma \times (0,T), \\ \delta\, \underline{\mathrm{n}} \cdot \mathbb{D}(z) \cdot \underline{\tau} + z \cdot \underline{\tau} = 0 & \text{on } \Gamma \times (0,T), \\ z(x,0) = G(x,0) & \text{in } \Omega. \end{cases} \tag{2.8}$$

We shall treat the nonlinear problem as a perturbation of such a linear system. Let us introduce the new unknowns

$$Z(x,t) = z(x,t) - G(x,t) \quad \text{and} \quad Q(x,t) = q(x,t) - \Pi(x,t)$$

so that we are reduced to a homogeneous problem for the new unknowns (Z, Q):

$$
\begin{cases}
\partial_t Z - \nabla \cdot \mathbb{D}(Z) + \nabla Q = -\partial_t G & \text{in } \Omega \times (0,T), \\
\nabla \cdot Z = 0 & \text{in } \Omega \times (0,T), \\
Z \cdot \underline{\mathrm{n}} = 0 & \text{on } \Gamma \times (0,T), \\
\delta\, \underline{\mathrm{n}} \cdot \mathbb{D}(Z) \cdot \underline{\tau} + Z \cdot \underline{\tau} = 0 & \text{on } \Gamma \times (0,T), \\
Z(x,0) = 0 & \text{in } \Omega.
\end{cases} \tag{2.9}
$$

The above problem is not completely standard, since the right-hand side does not satisfy the usual properties. For instance, one can note that $\partial_t G$ does not belong to the domain of the Stokes operator, since $\partial_t G \cdot \underline{\mathrm{n}} \neq 0$. This is the main difficulty: The low regularity of this term can be treated in a more standard way, while the above fact is responsible for a different approach.

THEOREM 2.2. *Assume that* (G, Π) *is a solution to system* (2.4), *with* G *satisfying the regularity property* (2.7). *Then there exists a unique solution* (z, q) *(where* q *is unique up to an additive function not depending on the space variable) to system* (2.8) *such that*

$$
z \in L^\infty(0,T; L^2(\Omega)) \cap L^2(0,T; H^1(\Omega)) \cap H^{\frac{1}{2}-\varepsilon}(0,T; L^2(\Omega)). \tag{2.10}
$$

Moreover,

$$
\begin{aligned}
\sup_{0 \le t \le T} \|z(t)\|^2 + \int_0^T \left(\|\mathbb{D}(z)(t)\|^2 + \frac{1}{\delta} \|z(t) \cdot \underline{\tau}\|_\Gamma^2 \right) dt + \|z\|^2_{H^{\frac{1}{2}-\varepsilon}(0,T;L^2(\Omega))} \\
\le C \left(\|G\|^2_{H^{\frac{1}{2}+\varepsilon}(0,T;L^2(\Omega))} + \int_0^T \|\mathbb{D}(G)\|^2 \, dt \right).
\end{aligned} \tag{2.11}
$$

Proof. By virtue of Theorem 2.1, it is enough to prove the same claim of this theorem on the solution (Z, Q) of problem (2.9). Since we just know that $\partial_t G \in H^{-\frac{1}{2}+\varepsilon}(0,T; L^2(\Omega))$, we introduce a sequence $G^N \in H^1(\mathbb{R}; L^2(\Omega))$ of approximate functions such that

(a) $G^N|_{[0,T]} \longrightarrow G$ in $H^{\frac{1}{2}+\varepsilon}(0,T; L^2(\Omega))$ as $N \to \infty$,

(b) $\|\partial_t G^N\|_{L^2(0,T;L^2(\Omega))} = N$.

The way to do this extension is rather standard: First, we can define $\overline{G} : \mathbb{R} \to L^2(\Omega)$ with an extension by reflection. Then, we consider a sequence ρ_N of mollifiers and the function G_N will be the restriction on $[0,T]$ of the function $\rho_N * \overline{G}$.

The proof is based on the Faedo–Galerkin procedure. By Clopeau, Mikelić, and Robert [8] (but also the recent abstract results in [6]), we know that there exists a basis $\{\phi_n\}_{n \in \mathbb{N}}$ of functions in $H^3(\Omega)$ of the space V (and also of H), such that

$$
\delta\, \underline{\mathrm{n}} \cdot \mathbb{D}(\phi_n) \cdot \underline{\tau} + \phi_n \cdot \underline{\tau} = 0.
$$

Now, let $Z_n^N(t,x) = \sum_{k=1}^n \zeta_{n,k}^N(t) \phi_k(x)$ be the solution of the following (finite-dimensional) linear system of ordinary differential equations (ODEs):

$$
\begin{cases}
\dfrac{d}{dt} \displaystyle\int_\Omega Z_n^N \cdot \phi_k + \int_\Omega \mathbb{D}(Z_n^N) \cdot \mathbb{D}(\phi_k) + \frac{1}{\delta} \int_\Gamma (Z_n^N \cdot \underline{\tau})\, (\phi_k \cdot \underline{\tau})\, d\sigma = -\frac{d}{dt} \int_\Omega G^N \cdot \phi_k, \\
\displaystyle\int_\Omega Z_n^N(x,0) \cdot \phi_k \, dx = 0
\end{cases}
$$

for $t \in (0,T)$ and $k = 1,\dots,n$. Notice that the divergence-free constraint and the boundary conditions on Z_n^N are automatically verified. By using a standard argument it is easy to prove that such a system of ODEs has a unique solution $Z_n^N \in L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;H^1(\Omega))$. Indeed, by multiplying each equation by the corresponding term $\zeta_{n,k}^N$, summing over k, and integrating by parts over Ω, one easily obtains the following estimate:

$$\sup_{0\le t\le T}\|Z_n^N(t)\|^2 + \int_0^T\left(\|\mathbb{D}(Z_n^N)\|^2 + \frac{1}{\delta}\int_0^T\|Z_n^N\|_\Gamma^2\right)dt \le C\|G^N\|^2_{H^1(0,T;L^2(\Omega))},$$

with a constant C, depending only on Ω. Unfortunately, such an estimate, beside being uniform in n, is not uniform in N due to property (b) of the approximate sequence $\{G^N\}_{N\in\mathbb{N}}$. Hence, we need other a priori estimates on the solutions Z_n^N of the finite-dimensional problem. We continue working on the Z_n^N, since such functions are smooth enough for all computations that will be performed.

We again multiply the equations by the terms $\zeta_{n,k}^N$, sum over k, and integrate by parts, but now we estimate the right-hand side in the following way:

(2.12)
$$\begin{aligned}\sup_{0\le t\le T}\|Z_n^N(t)\|^2 + \int_0^T\left(\|\mathbb{D}(Z_n^N)\|^2 + \frac{1}{\delta}\|Z_n^N\|_\Gamma^2\right)dt &\le \left|\int_0^T\int_\Omega \partial_t G^n\cdot Z_n^N\right| dx\,dt\\ &\le \|\partial_t G^N\|_{H^{-\frac12+\varepsilon}(0,T;L^2(\Omega))}\|Z_n^N\|_{H^{\frac12-\varepsilon}(0,T;L^2(\Omega))}\\ &\le \|G^N\|_{H^{\frac12+\varepsilon}(0,T;L^2(\Omega))}\|Z_n^N\|_{H^{\frac12-\varepsilon}(0,T;L^2(\Omega))},\end{aligned}$$

so that we need only show a uniform estimate (with respect to both n and N) of Z_n^N in the space $H^{\frac12-\varepsilon}(0,T;L^2(\Omega))$. We shall use the Fourier transform characterization of the norm of fractional Sobolev spaces (see Adams [2]) to get such an estimate. Each $\widetilde{Z}_n^N$ (such functions have been defined in (2.3)) is a solution of the following equation:

$$\begin{aligned}&\frac{d}{dt}\int_\Omega \widetilde{Z}_n^N\cdot\phi_k + \int_\Omega \mathbb{D}(\widetilde{Z}_n^N)\cdot\mathbb{D}(\phi_k) + \frac{1}{\delta}\int_\Gamma \widetilde{Z}_n^N\cdot\phi_k\,d\sigma\\ &\quad= -\frac{d}{dt}\int_\Omega \widetilde{G}^N\cdot\phi_k + \delta(t)\int_\Omega G^N(0)\cdot\phi_k - \delta(t-T)\int_\Omega (Z_n^N(T)+G^N(T))\cdot\phi_k\end{aligned}$$

for each $k = 1,\dots,n$, in the sense of distributions with respect to the time variable. Here $\delta(\cdot)$ is the usual Dirac's *delta function*. In the frequency Fourier variable ξ, the above equation reads as follows:

$$\begin{aligned}&-\mathtt{i}\xi\int_\Omega \widehat{Z}_n^N\cdot\phi_k + \int_\Omega \mathbb{D}(\widehat{Z}_n^N)\cdot\mathbb{D}(\phi_k) + \frac{1}{\delta}\int_\Gamma \widehat{Z}_n^N\cdot\phi_k\,d\sigma\\ &\qquad= \mathtt{i}\xi\int_\Omega \widehat{G}^N\cdot\phi_k + \int_\Omega G^N(0)\cdot\phi_k - \mathtt{e}^{-\mathtt{i}\xi T}\int_\Omega (Z_n^N(T)+G^N(T))\cdot\phi_k;\end{aligned}$$

see, for instance, Lions [25], where this tool is used to prove estimates on the fractional derivative of the solution. Note that in [25], and in all works involving fractional derivatives for the Navier–Stokes equations, the starting point is the existence of a weak solution, on which it is possible to prove additional estimates. In our case the existence of a weak solution derives from the fractional derivative estimates and at present it does not seem possible to prove the usual existence results.

Consequently, we get

$$-\,\mathtt{i}\xi\|\widehat{Z}_n^N(\xi)\|^2+\|\mathbb{D}(\widehat{Z}_n^N)(\xi)\|^2+\frac{1}{\delta}\|\widehat{Z}_n^N(\xi)\|_\Gamma^2$$
$$=\mathtt{i}\xi\int_\Omega \widehat{G}^N\cdot\overline{\widehat{Z}_n^N}+\int_\Omega G^N(0)\cdot\overline{\widehat{Z}_n^N}-\mathtt{e}^{-\mathtt{i}\xi T}\int_\Omega (Z_n^N(T)+G^N(T))\cdot\overline{\widehat{Z}_n^N}.$$

Take the imaginary part and multiply both sides of the previous formula by $|\xi|^{2\alpha-1}$, with $\alpha<\frac{1}{2}$ so that, by using Young's inequality, one gets

$$|\xi|^{2\alpha}\|\widehat{Z}_n^N(\xi)\|^2\le C|\xi|^{2\alpha}\|\widehat{G}^N\|^2+C|\xi|^{2\alpha-2}(\|G^N(T)\|+\|Z_n^N(T)\|+\|G^N(0)\|)^2.$$

In order to estimate the integral $\int_{\mathbb{R}}|\xi|^{2\alpha}\|\widehat{Z}_n^N(\xi)\|^2\,d\xi$, we split it into two parts: by the above estimate,

$$\int_{|\xi|>1}|\xi|^{2\alpha}\|\widehat{Z}_n^N(\xi)\|^2$$
$$\le C\int_{\mathbb{R}}|\xi|^{2\alpha}\|\widehat{G}^N\|^2+C(\|G^N(T)\|+\|Z_n^N(T)\|+\|G^N(0)\|)^2\int_{|\xi|>1}|\xi|^{2\alpha-2};$$

the first term on the right-hand side is controlled by $C\|G^N\|^2_{H^{\frac{1}{2}+\varepsilon}(0,T;L^2(\Omega))}$, while

$$\|Z_n^N(T)\|^2\le C\|G^N\|_{H^{\frac{1}{2}+\varepsilon}(0,T;L^2(\Omega))}\|Z_n^N\|_{H^{\frac{1}{2}-\varepsilon}(0,T;L^2(\Omega))},$$

by virtue of (2.12); $\|G^N(0)\|$ is bounded by $\|G(0)\|$; and finally, by using the Morrey inequality (see Adams [2]), which implies $H^{1/2+\varepsilon}(0,T)\subset C([0,T])$, we get

$$\|G^N(T)\|\le\|G^N\|_{H^{\frac{1}{2}+\varepsilon}(0,T;L^2(\Omega))}.$$

The second part is estimated as follows, by using Parseval's theorem, Poincaré's inequality, and estimate (2.12):

$$\begin{aligned}\int_{|\xi|\le 1}|\xi|^{2\alpha}\|\widehat{Z}_n^N\|^2\,d\xi&\le\int_{\mathbb{R}}\|\widehat{Z}_n^N\|^2\,d\xi=\int_0^T\|Z_n^N(t)\|^2\,dt\\&\le C\int_0^T\|\mathbb{D}(Z_n^N)\|^2dt\\&\le C\|G^N\|_{H^{\frac{1}{2}+\varepsilon}(0,T;L^2(\Omega))}\|Z_n^N\|_{H^{\frac{1}{2}-\varepsilon}(0,T;L^2(\Omega))}.\end{aligned}$$

In conclusion, by collecting all of the above estimates we finally get that, for each $\varepsilon\in(0,\frac{1}{2})$, there exists a constant C, depending only on Ω and ε, such that

$$\|Z_n^N\|_{H^{\frac{1}{2}-\varepsilon}(0,T;L^2(\Omega))}\le C\|G^N\|_{H^{\frac{1}{2}+\varepsilon}(0,T;L^2(\Omega))},\tag{2.13}$$

which, together with (2.12), shows that Z_n^N is bounded, uniformly in n and N, in the spaces $H^{\frac{1}{2}-\varepsilon}(0,T;L^2(\Omega))$, $L^\infty(0,T;L^2(\Omega))$ and $L^2(0,T;H^1(\Omega))$.

Hence, it is possible to extract a (diagonal) subsequence converging weakly in $L^2(0,T;H^1(\Omega))$, weakly-$*$ in $L^\infty(0,T;L^2(\Omega))$, and strongly in $L^2((0,T)\times\Omega)$ to the unique solution Z of problem (2.9). Indeed, $Z\in H^{\frac{1}{2}-\varepsilon}(0,T;L^2(\Omega))$, that is, the topological dual space of $H^{-\frac{1}{2}+\varepsilon}(0,T;L^2(\Omega))$, the space to which $\partial_t G$ belongs. Furthermore, by passing to the limit and using the semicontinuity of the norms, it follows that Z satisfies the claim stated at the beginning of this proof. ◻

Remark 2.3. The assumption $g \in H^{\frac{1}{2}+\varepsilon}(0,T;H^{\frac{1}{2}}(\Gamma))$, which in turn gives $G \in H^{\frac{1}{2}+\varepsilon}(0,T;H^1(\Omega))$, seems to be rather technical for the presence of ϵ. If $g \in H^{\frac{1}{2}}(0,+\infty;H^{\frac{1}{2}}(\Gamma))$, it follows that $G \in H^{\frac{1}{2}}(0,+\infty;H^1(\Omega))$ and Theorem 2.2 holds accordingly.

Indeed, the main point is estimate (2.12), in which the right-hand side becomes $\|G^N\|_{H^{\frac{1}{2}}}\|Z_n^N\|_{H^{\frac{1}{2}}}$ and, following the lines of the proof presented above, the estimate on the Fourier transform gives that Z_n^N is bounded, uniformly in n and N, in $H^{\frac{1}{2}}(0,+\infty;L^2(\Omega))$. Notice that in the *critical* case $H^{\frac{1}{2}}$, we work on the whole time interval $[0,+\infty)$ to avoid the boundary terms $\|G^N(0)\|$ and $\|G^N(T)\|$, which cannot be estimated by using the Morrey inequality.

We also note that this small relaxation on the assumptions on G requires us to add the hypothesis $G \in L^\infty(0,T;L^2(\Omega))$. Otherwise the function z will not itself belong to $L^\infty(0,T;L^2(\Omega))$ and this fact is crucial to proving the corresponding bound for weak solutions to the full nonlinear Navier–Stokes problem.

2.4. The nonlinear problem. In this section we finally prove Theorem 1.2. Again, we make use of an auxiliary problem; namely, we introduce the new variables

$$U = u - z \qquad \text{and} \qquad P = p - q,$$

where (z,q) is the solution to the linear evolution problem (2.8), and the pair (U,P) solves the following problem:

$$\begin{cases} \partial_t U - \nabla\cdot\mathbb{D}(U) + [(U+z)\cdot\nabla](U+z) + \nabla P = f & \text{in } \Omega\times(0,T),\\ \nabla\cdot U = 0 & \text{in } \Omega\times(0,T),\\ U\cdot\underline{\mathrm{n}} = 0 & \text{on } \Gamma\times(0,T),\\ \delta\,\underline{\mathrm{n}}\cdot\mathbb{D}(U)\cdot\underline{\tau} + U\cdot\underline{\tau} = 0 & \text{on } \Gamma\times(0,T),\\ U(x,0) = u_0(x) - G(x,0) & \text{in } \Omega. \end{cases} \tag{2.14}$$

By virtue of Theorem 2.2, the existence Theorem 1.2 for the nonlinear problem is a straightforward consequence of the following proposition.

PROPOSITION 2.4. *Assume that* (G,Π) *is a solution to system* (2.4), *with* $G \in H^{\frac{1}{2}+\varepsilon}(0,T;L^2(\Omega)) \cap L^2(0,T;H^1(\Omega))$. *Then, there exists a unique*

$$U \in L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;H^1(\Omega)) \tag{2.15}$$

solution to problem (2.14). *Moreover, the following estimate holds true:*

$$\begin{aligned} \sup_{0\le s\le t}\|U(s)\|^2 + \int_0^t \left(\|\mathbb{D}(U)\|^2 + \frac{1}{\delta}\|U\|_\Gamma^2\right) ds \\ \le \|u_0 - G(\cdot,0)\|^2 \mathrm{e}^{A(t)} + C\int_0^t (\|f\|^2 + \|\nabla z(s)\|^2\|z(s)\|)\mathrm{e}^{A(t)-A(s)}\, ds, \end{aligned} \tag{2.16}$$

where

$$A(t) = Ct + C\left(1 + \|z\|^2_{L^\infty(0,T;L^2(\Omega))}\right)\int_0^t \|\nabla z\|^2\, ds$$

and C *is a constant depending only on* Ω.

Proof. The proof is rather standard and proceeds via a Faedo–Galerkin approximation, as in the proof of Theorem 2.2. We show only an a priori estimate, whose computations are formal but completely meaningful at the level of the Faedo–Galerkin approximate functions. Multiply (2.14) by U and integrate by parts to get

$$\frac{1}{2}\frac{d}{dt}\|U\|^2+\|\mathbb{D}(U)\|^2+\frac{1}{\delta}\|U\|_\Gamma^2=\int_\Omega U\cdot[(U+z)\cdot\nabla](U+z)+\int_\Omega f\cdot U.$$

The estimate of the integral involving f is straightforward, since it is bounded by $\|f\|^2+\|U\|^2$. We estimate the nonlinear term in the right-hand side by using the Gagliardo–Nirenberg inequality

$$\|u\|_{L^4}\leq C\|u\|^{1/2}\|\nabla u\|^{1/2}\qquad\forall\, u\in H^1(\Omega). \tag{2.17}$$

Note that such an inequality is a little bit more general than the so-called *Ladyžhenskaya inequality*, since here the functions are not vanishing on the boundary of Ω and the constant C depends on Ω.

We first observe that since $\nabla\cdot U=0$ and $U\cdot\underline{\mathrm{n}}=0$, then

$$\int_\Omega U\cdot(U\cdot\nabla)U=0\qquad\text{and}\qquad\int_\Omega U\cdot(U\cdot\nabla)z=-\int_\Omega z\cdot(U\cdot\nabla)U,$$

so that, by using repeatedly the Gagliardo–Nirenberg inequality given above, Hölder's inequality, and Young's inequality, we get

$$\begin{aligned}\left|\int_\Omega U\cdot[(U+z)\cdot\nabla](U+z)\right| &\leq 2\|U\|_{L^4}\|z\|_{L^4}\|\nabla U\|+\|U\|_{L^4}\|z\|_{L^4}\|\nabla z\|\\ &\leq\frac{1}{2}\|\nabla U\|^2+C\|z\|_{L^4}^4\|U\|^2+C\|z\|_{L^4}^{\frac{4}{3}}\|\nabla z\|^{\frac{4}{3}}\|U\|^{\frac{2}{3}}\\ &\leq\frac{1}{2}\|\nabla U\|^2+C\|z\|^2\|\nabla z\|^2\|U\|^2+C\|z\|^{\frac{2}{3}}\|\nabla z\|^2\|U\|^{\frac{2}{3}}\\ &\leq\frac{1}{2}\|\nabla U\|^2+C(1+\|z\|^2)\|\nabla z\|^2\|U\|^2+C\|\nabla z\|^2\|z\|.\end{aligned}$$

Since, by Theorem 2.2, $z\in L^\infty(0,T;L^2(\Omega))\cap L^2(0,T;H^1(\Omega))$, both terms $(1+\|z\|^2)\|\nabla z\|^2$ and $\|\nabla z\|^2\|z\|$ are integrable in time, and, by Gronwall's lemma, we can deduce that U is bounded in $L^\infty(0,T;L^2(\Omega))$ and $L^2(0,T;H^1(\Omega))$. Moreover, formula (2.16) also follows.

Finally, uniqueness of the solution follows from similar arguments. Indeed, if $\widetilde{U}$ is the difference between two solutions U_1 and U_2, one easily gets

$$\frac{1}{2}\frac{d}{dt}\|\widetilde{U}\|^2\leq(\|U_2\|_{L^4}^4+\|z\|_{L^4}^4+\|\nabla z\|^2)\|\widetilde{U}\|^2$$

and, since $\widetilde{U}(0)=0$, from Gronwall's lemma it follows that $\widetilde{U}\equiv 0$. ☐

Remark 2.5. In the proof of the result of this section we used in a fundamental way estimate (2.17). In the three-dimensional case this inequality is no longer true. Instead, it holds that

$$\|u\|_{L^4}\leq C\|u\|^{1/4}\|\nabla u\|^{3/4}\qquad\forall\, u\in H^1(\Omega),$$

which can be used to prove just local existence of weak solutions. The global result proved in the two-dimensional case depends in an essential manner on the stronger estimate, and this is the critical difference between the two cases.

3. The proof of Theorem 1.3. In this section we study the limit of the solution of the vorticity seeding model as $\delta \to 0$. We prove the convergence of the solutions of (3.1) to the solutions of the nonstationary Navier–Stokes system with the no-slip boundary condition (3.2). This supports the idea of using a nonstandard boundary condition on the new boundary $\Gamma_1 \subset \Omega$, such that the region between Γ and Γ_1 is very narrow. When the width of this region (presumably the boundary layer) shrinks to zero, then the solution of the Navier–Stokes equations with the usual no-slip boundary condition is recovered.

3.1. Comparison of solutions with different boundary data. Let the following be given: $g \in H^{\frac{1}{2}+\varepsilon}(0,T;H^{\frac{1}{2}}(\Gamma))$, satisfying the compatibility condition (1.5); $f \in L^2((0,T)\times\Omega)$; and $u_0 \in H^1(\Omega)$, with $\nabla\cdot u_0 = 0$. Without loss of generality, we can assume that $u_0 \equiv 0$. Denote by G_δ the solution of the linear stationary problem (2.4), with boundary condition $G_\delta \cdot \underline{\mathrm{n}} = \delta\, g$. Consider the solution (u_δ, p_δ) to the system

$$
\begin{cases}
\partial_t u_\delta - \nabla\cdot\mathbb{D}(u_\delta) + u_\delta\cdot\nabla u_\delta + \nabla p_\delta = f & \text{in } \Omega\times(0,T),\\
\nabla\cdot u_\delta = 0 & \text{in } \Omega\times(0,T),\\
u_\delta\cdot\underline{\mathrm{n}} = \delta\, g(x,t) & \text{on } \Gamma\times(0,T),\\
\delta\,\underline{\mathrm{n}}\cdot\mathbb{D}(u_\delta)\cdot\underline{\tau} + u_\delta\cdot\underline{\tau} = 0 & \text{on } \Gamma\times(0,T),\\
u_\delta(x,0) = u_0(x) + G(x,0) & \text{in } \Omega.
\end{cases}
\tag{3.1}
$$

We want to show the convergence of the vector valued function u_δ to the solution v of the Navier–Stokes equations with zero Dirichlet data:

$$
\begin{cases}
\partial_t v - \nabla\cdot\mathbb{D}(v) + v\cdot\nabla v + \nabla\pi = f & \text{in } \Omega\times(0,T),\\
\nabla\cdot v = 0 & \text{in } \Omega\times(0,T),\\
v = 0 & \text{on } \Gamma\times(0,T),\\
v(x,0) = u_0 & \text{in } \Omega.
\end{cases}
\tag{3.2}
$$

To this end, we also introduce the solution z_δ to the linear evolution problem (2.8), with boundary condition $z_\delta\cdot\underline{\mathrm{n}} = \delta\, g$, and we set

$$
U_\delta = u_\delta - z_\delta \qquad \text{and} \qquad w_\delta = u_\delta - z_\delta - v.
$$

The function w_δ satisfies the following "homogeneous" system

$$
\begin{cases}
\partial_t w_\delta - \nabla\cdot\mathbb{D}(w_\delta) + R(w_\delta, z_\delta, v, U_\delta) + \nabla r = 0 & \text{in } \Omega\times(0,T),\\
\nabla\cdot w_\delta = 0 & \text{in } \Omega\times(0,T),\\
w_\delta\cdot\underline{\mathrm{n}} = 0 & \text{on } \Gamma\times(0,T),\\
\delta\,\underline{\mathrm{n}}\cdot\mathbb{D}(w_\delta)\cdot\underline{\tau} + w_\delta\cdot\underline{\tau} = -\delta\,\underline{\mathrm{n}}\cdot\mathbb{D}(v)\cdot\underline{\tau} & \text{on } \Gamma\times(0,T),\\
w_\delta(x,0) = 0 & \text{in } \Omega,
\end{cases}
\tag{3.3}
$$

where

$$
R(w_\delta, z_\delta, v, U_\delta) = (U_\delta\cdot\nabla)w_\delta + (w_\delta\cdot\nabla)v + (U_\delta\cdot\nabla)z_\delta + (z_\delta\cdot\nabla)U_\delta + (z_\delta\cdot\nabla)z_\delta.
$$

We multiply the first equation in (3.3) by w_δ and integrate by parts to get

$$
\frac{1}{2}\frac{d}{dt}\|w_\delta\|^2 + \|\mathbb{D}(w_\delta)\|^2 + \frac{1}{\delta}\|w_\delta\|_\Gamma^2 = -\int_\Gamma (\underline{\mathrm{n}}\cdot\mathbb{D}(v)\cdot\underline{\tau})(w_\delta\cdot\underline{\tau})\,d\sigma - \int_\Omega w_\delta\cdot R(w_\delta, z_\delta, v, U_\delta)\,dx.
$$

The boundary integral in the right-hand side may be increased in the following way.

$$\left|\int_\Gamma (\underline{\mathrm{n}} \cdot \mathbb{D}(v) \cdot \underline{\tau})(w_\delta \cdot \underline{\tau})\, d\sigma\right| \le \frac{\delta}{2}\|v\|^2_{H^2} + \frac{1}{2\delta}\|w_\delta\|^2_\Gamma. \tag{3.4}$$

Then we analyze the second integral. By integration by parts over Ω (note that $U_\delta \cdot \underline{\mathrm{n}} = 0$ on Γ) and by using Hölder's inequality, the Gagliardo–Nirenberg inequality, and Young's inequality, we obtain

$$\begin{aligned}
&\left|\int_\Omega w_\delta \cdot R(w_\delta, z_\delta, v, U_\delta)\, dx\right| \\
&\le \frac{1}{2}\|\mathbb{D}(w_\delta)\|^2 + C\big[\|\nabla v\|^2 + \|\nabla U_\delta\|^{\frac{2}{3}}\|\nabla z_\delta\|^{\frac{4}{3}} + \|\nabla U_\delta\|^{\frac{4}{3}}\|\nabla z_\delta\|^{\frac{2}{3}} + \|\nabla z_\delta\|^2\big]\|w_\delta\|^2 \\
&\quad + C\big[\|\nabla U_\delta\|^{\frac{2}{3}}\|\nabla z_\delta\|^{\frac{4}{3}}\|U_\delta\| + \|\nabla U_\delta\|^{\frac{4}{3}}\|\nabla z_\delta\|^{\frac{2}{3}}\|z_\delta\| + \|\nabla z_\delta\|^2\|z_\delta\|\big].
\end{aligned}$$

Indeed, for example,

$$\begin{aligned}
\left|\int_\Omega w_\delta \cdot (z_\delta \cdot \nabla) U_\delta\right| &\le \|w_\delta\|_{L^4}\|z_\delta\|_{L^4}\|\nabla U\| \\
&\le C\|w_\delta\|^{\frac{1}{2}}\|\mathbb{D}(w_\delta)\|^{\frac{1}{2}}\|z_\delta\|^{\frac{1}{2}}\|\mathbb{D}(z_\delta)\|^{\frac{1}{2}}\|\nabla U_\delta\| \\
&\le \frac{1}{8}\|\mathbb{D}(w_\delta)\|^2 + C\|\nabla U_\delta\|^{\frac{4}{3}}\|\nabla z_\delta\|^{\frac{2}{3}}\|w_\delta\|^{\frac{2}{3}}\|z_\delta\|^{\frac{2}{3}} \\
&\le \frac{1}{8}\|\mathbb{D}(w_\delta)\|^2 + C\|\nabla U_\delta\|^{\frac{4}{3}}\|\nabla z_\delta\|^{\frac{2}{3}}(\|w_\delta\|^2 + \|z_\delta\|).
\end{aligned}$$

For simplicity we set

$$\begin{aligned}
\psi(t) &:= C\big[\|\nabla U_\delta\|^{\frac{2}{3}}\|\nabla z_\delta\|^{\frac{4}{3}}\|U_\delta\| + \|\nabla U_\delta\|^{\frac{4}{3}}\|\nabla z_\delta\|^{\frac{2}{3}}\|z_\delta\| + \|\nabla z_\delta\|^2\|z_\delta\|\big], \\
\phi(t) &:= \big[\|\nabla v\|^2 + \|\nabla U_\delta\|^{\frac{2}{3}}\|\nabla z_\delta\|^{\frac{4}{3}} + \|\nabla U_\delta\|^{\frac{4}{3}}\|\nabla z_\delta\|^{\frac{2}{3}} + \|\nabla z_\delta\|^2\big], \\
\chi(t) &:= \|v\|^2_{H^2},
\end{aligned}$$

so that, by collecting all of the estimates we obtained before, we end up with the following differential inequality:

$$\frac{d}{dt}\|w_\delta\|^2 + \|\mathbb{D}(w_\delta)\|^2 + \frac{1}{\delta}\|w_\delta\|^2_\Gamma \le \delta\chi(t) + \psi(t) + \phi(t)\|w_\delta\|^2. \tag{3.5}$$

3.1.1. Estimate of and in terms of . First, we note that the functions ψ and ϕ belong to $L^1(0,T)$, by virtue of the regularity properties (2.10) and (2.15) of z_δ and U_δ, respectively. The function $\chi(t)$ belongs to $L^1(0,T)$ as well, since

$$\|v\|_{L^2(0,T;H^2)} \le C\big[\|v(x,0)\|_{H^1} + \|f\|_{L^2((0,T)\times\Omega)}\big];$$

see, for instance, Constantin and Foiaş [10]. And also at this point is crucial to consider the two-dimensional problem, since such an estimate is available just for small times in the three-dimensional case.

It is important now to sharply check the correct behavior of the functions $\phi(t)$ and $\psi(t)$ in terms of δ. Before going further, we collect in the following lemma the results of section 2 that we shall need.

LEMMA 3.1. *Under the assumptions of Theorem* 1.3*, the following estimates hold:*

(i) $\sup_{0\le t\le T}\|z_\delta\|^2+\int_0^T\|\mathbb{D}(z_\delta)\|^2\,dt\le C_1\delta\|g\|^2_{H^{\frac12+\varepsilon}(0,T;H^{\frac12}(\Gamma))}$,

(ii) $\sup_{0\le t\le T}\|U_\delta\|^2+\int_0^T\Big(\|\mathbb{D}(U_\delta)\|^2+\frac{1}{\delta}\|U_\delta\|^2_\Gamma\Big)\,dt\le C_2$,

for all $0<\delta\le 1$*, where the constant* C_1 *depends only on* Ω*, while the constant* C_2 *depends only on* Ω*,* T*,* $\|f\|_{L^2}$*, and* $\|g\|_{H^{\frac12+\varepsilon}}$*.*

Proof. Property (i) is a consequence of (2.11). Indeed, by (2.6) and Poincaré's inequality,

$$\|G_\delta\|\le C\|\mathbb{D}(G_\delta)\|\le C\sqrt{\delta}\|g\|,$$

so that $\|G_\delta\|_{H^{\frac12+\varepsilon}(0,T;L^2(\Omega))}\le C\sqrt{\delta}\|g\|_{H^{\frac12+\varepsilon}(0,T;L^2(\Omega))}$.

Again, from (2.6),

$$\int_0^T\|\mathbb{D}(G_\delta)\|^2\,dt\le C\delta\|g\|^2_{H^{\frac12+\varepsilon}(0,T;L^2(\Omega))},$$

and in conclusion (i) holds true.

As it concerns (ii), we have that, by Poincaré's inequality, (2.6), and Sobolev embeddings,

$$\|G_\delta(\cdot,0)\|\le C\|\mathbb{D}(G_\delta)(\cdot,0)\|\le C\sqrt{\delta}\|g(\cdot,0)\|\le C\sqrt{\delta}\|g\|_{H^{\frac12+\varepsilon}(0,T;L^2)}.$$

Moreover, since $\delta\le 1$ and from the above estimates it follows that

$$A(t)\le A(T)\le CT+C\Big(1+\|g\|^4_{H^{\frac12+\varepsilon}(0,T;L^2(\Gamma))}\Big),$$

that is, $A(t)$ is uniformly bounded by a constant independent of δ, and that

$$\begin{aligned}\int_0^t(\|f\|^2+\|\nabla z_\delta\|^2\|z_\delta\|)\,\mathsf{e}^{A(t)-A(s)}\,ds&\le\mathsf{e}^{A(T)}\left(\int_0^T\|f\|^2+\|z_\delta\|_{L^\infty(0,T;L^2)}\|\mathbb{D}(z_\delta)\|^2\right)\\&\le\Big(\|f\|^2_{L^2((0,T)\times\Omega)}+C\|g\|^3_{H^{\frac12+\varepsilon}(0,T;L^2(\Gamma))}\Big)\mathsf{e}^{A(T)}\end{aligned}$$

as well, so that, by (2.16), (ii) also follows. ∎

The first consequence of the above lemma is that

$$\int_0^t\phi(s)\,ds\le\int_0^T\|\nabla v\|^2\,dt+C(\Omega,T,f,g),\qquad 0\le t\le T,$$

with a bound uniform in δ, for δ small. Moreover, for $t\in[0,T]$

$$\begin{aligned}\int_0^t\psi(s)\,ds&\le C\left(\int_0^T\|\nabla U_\delta\|^2\right)^{\frac13}\left(\int_0^T\|\nabla z_\delta\|^2\right)^{\frac23}\|U_\delta\|_{L^\infty(0,T;L^2)}\\&\quad+C\left(\int_0^T\|\nabla U_\delta\|^2\right)^{\frac23}\left(\int_0^T\|\nabla z_\delta\|^2\right)^{\frac13}\|z_\delta\|_{L^\infty(L^2)}\\&\quad+C\|z_\delta\|_{L^\infty(L^2)}\int_0^T\|\nabla z_\delta\|^2\\&\le C(\delta^{\frac23}+\delta^{\frac56}+\delta^{\frac32})\\&\le C\delta^{\frac23},\end{aligned}$$

so that, by (3.5) and Gronwall's inequality, it follows that

$$\begin{aligned}\sup_{0\le t\le T}\|w_\delta\|^2+\int_0^T\left(\|\mathbb{D}(w_\delta)\|^2+\frac{1}{\delta}\|w_\delta\|_\Gamma^2\right)dt &\le \int_0^T[\delta\chi(s)+\psi(s)]\mathrm{e}^{\int_s^T\phi(r)\,dr}\,ds\\ &\le \mathrm{e}^{\int_0^T\phi(s)\,ds}\int_0^T[\delta\chi(s)+\psi(s)]\,ds\\ &\le C\delta^{\frac{2}{3}}\end{aligned}$$

and finally

$$\sup_{0\le t\le T}\|w_\delta\|^2+\int_0^T\left(\|\mathbb{D}(w_\delta)\|^2+\frac{1}{\delta}\|w_\delta\|_\Gamma^2\right)dt=\mathcal{O}(\delta^{\frac{2}{3}}).$$

Since the corresponding norms of z_δ are of order $\delta^{\frac{1}{2}}$, by the previous lemma, we finally deduce that $u_\delta-v$ is of the order $\delta^{\frac{1}{3}}$, and Theorem 1.3 is proved.

Remark 3.2. The tangential trace of the function w_δ converges a little better, since from the above estimate we can deduce

$$\|w_\delta\|_{L^2(\Gamma\times(0,T))}=\mathcal{O}(\delta^{5/6}).$$

Remark 3.3. The above proof shows that there is a crucial loss in the estimates due to: (1) the boundary effect in the estimate (2.6) and (2) the nonlinearity (recall the contribution of U_δ in the estimate of $\psi(t)$). We observe that the original model was supplemented by the boundary condition

$$u\cdot\underline{\mathrm{n}}=\delta^2 g,$$

but we used in fact $u\cdot\underline{\mathrm{n}}=\delta\, g$. By running through the proof of Theorem 1.3 one gets, using the boundary condition

$$u\cdot\underline{\mathrm{n}}=\delta^\alpha g,$$

that $z_\delta=\mathcal{O}(\delta^{\alpha-\frac{1}{2}})$, so that $\int_0^T\psi(t)\,dt=\mathcal{O}(\delta^{\frac{4}{3}(\alpha-\frac{1}{2})})$ and finally $w_\delta=\mathcal{O}(\delta^{\frac{2}{3}(\alpha-\frac{1}{2})})$. In particular, for $\alpha=2$, which is the value corresponding to models in [4, 24], the behavior of z_δ matches the loss in the estimates due to the boundary term (3.4); see section 1.2. Hence, for larger values of α, the convergence remains slow, because of this term. Supposedly, the loss in the convergence rate caused by this term is an intrinsic feature of the problem, while those losses caused by the nonlinearity may have technical reasons.

REFERENCES

[1] Y. Achdou, O. Pironneau, and F. Valentin, *Effective boundary conditions for laminar flows over periodic rough boundaries*, J. Comput. Phys., 147 (1998), pp. 187–218.
[2] R. A. Adams, *Sobolev Spaces*, Pure and Applied Mathematics, Vol. 65, Academic Press, New York, London, 1975.
[3] H. Amann, *On the strong solvability of the Navier–Stokes equations*, J. Math. Fluid Mech., 2 (2000), pp. 16–98.
[4] G. Beavers and D. D. Joseph, *Boundary conditions at a naturally impermeable wall*, J. Fluid Mech., 30 (1967), pp. 197–207.
[5] C. Bègue, C. Conca, F. Murat, and O. Pironneau, *Les équations de Stokes et de Navier–Stokes avec des conditions aux limites sur la pression*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. IX (Paris, 1985–1986), Pitman Res. Notes Math. Ser. 181, Longman Scientific and Technical, Harlow, UK, 1988, pp. 179–264.

[6] H. Beirão da Veiga, *Regularity for Stokes and generalized Stokes systems under nonhomogeneous slip type boundary conditions*, Adv. Differential Equations, 9 (2004), pp. 1079–1114.

[7] A. Bensoussan and R. Temam, *Équations stochastiques du type Navier–Stokes*, J. Funct. Anal., 13 (1973), pp. 195–222.

[8] T. Clopeau, A. Mikelić, and R. Robert, *On the vanishing viscosity limit for the* 2D *incompressible Navier–Stokes equations with the friction type boundary conditions*, Nonlinearity, 11 (1998), pp. 1625–1636.

[9] L. Consiglieri, *A nonlocal friction problem for a class of non-Newtonian flows*, Port. Math. (N.S.), 60 (2003), pp. 237–252.

[10] P. Constantin and C. Foiaş, *Navier–Stokes Equations*, University of Chicago Press, Chicago, 1988.

[11] A. Dunca, V. John, W. J. Layton, and N. Sahin, *Numerical analysis of large eddy simulation*, in Proceedings of the Third AFOSR International Conference on DNS and LES, L. S. C. Liu and T. Beutner, eds., Greyden Press, Columbus, OH, 2001, pp. 359–364.

[12] H. Fujita, *A coherent analysis of Stokes flows under boundary conditions of friction type*, in Scientific and Engineering Computations for the 21st century—Methodologies and Applications (Shizuoka, 2001), J. Comput. Appl. Math., 149 (2002), pp. 57–69.

[13] A. Fursikov, M. Gunzburger, and L. Hou, *Trace theorems for three-dimensional, time-dependent solenoidal vector fields and their applications*, Trans. Amer. Math. Soc., 354 (2002), pp. 1079–1116.

[14] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier–Stokes Equations*. Vol. I. *Linearized Steady Problems*, Springer Tracts Nat. Philos. 38, Springer-Verlag, New York, 1994.

[15] G. P. Galdi and W. J. Layton, *Approximation of the larger eddies in fluid motions.* II. *A model for space-filtered flow*, Math. Models Methods Appl. Sci., 10 (2000), pp. 343–350.

[16] J. G. Heywood, R. Rannacher, and S. Turek, *Artificial boundaries and flux and pressure conditions for the incompressible Navier–Stokes equations*, Internat. J. Numer. Methods Fluids, 22 (1996), pp. 325–352.

[17] W. Jäger and A. Mikelić, *On the interface boundary condition of Beavers, Joseph, and Saffman*, SIAM J. Appl. Math., 60 (2000), pp. 1111–1127.

[18] W. Jäger and A. Mikelić, *On the roughness-induced effective boundary conditions for an incompressible viscous flow*, J. Differential Equations, 170 (2001), pp. 96–122.

[19] W. Jäger and A. Mikelić, *Couette flows over a rough boundary and drag reduction*, Comm. Math. Phys., 232 (2003), pp. 429–455.

[20] V. John, *Slip with friction and penetration with resistance boundary conditions for the Navier–Stokes equations—Numerical tests and aspects of the implementation*, J. Comput. Appl. Math., 147 (2002), pp. 287–300.

[21] V. John, W. J. Layton, and N. Sahin, *Derivation and analysis of near wall models for channel and recirculating flows*, Comput. Math. Appl., 28 (2004), pp. 1135–1151.

[22] J. Kelliher, *Navier–Stokes equations with Navier boundary conditions for a bounded domain in the plane*, SIAM J. Math. Anal., to appear.

[23] S. G. Kreĭn and G. I. Laptev, *On the problem of the motion of a viscous fluid in an open vessel*, Funkcional. Anal. i Priložen., 2 (1968), pp. 40–50.

[24] W. J. Layton, *Advanced models for large eddy simulation*, in Proceedings of the 32nd Computational Fluid Dynamics—Multiscale Methods, VKI Lectures Series Monographs, vol. VKI LS 2002-06, von Kármán Institute for Fluid Dynamics, Rhode-Saint-Genèse, Belgium, 2002.

[25] J.-L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Gauthier–Villars, Paris, 1969.

[26] M. C. Lopes Filho, H. J. Nussenzveig Lopes, and G. Planas, *On the inviscid limit for two-dimensional incompressible flow with Navier friction condition*, SIAM J. Math. Anal., 36 (2005), pp. 1130–1141.

[27] J. C. Maxell, *On stresses in rarefied gases arising from inequalities of temperature*, Phil. Trans. Roy. Soc., 170 (1879), pp. 249–256.

[28] A. S. Monin and A. M. Yaglom, *Statistical Fluid Mechanics: Mechanics of Turbulence*, Vol. 1, MIT Press, Cambridge, MA, 1975.

[29] P. B. Mucha, *On the inviscid limit of the Navier–Stokes equations for flows with large flux*, Nonlinearity, 16 (2003), pp. 1715–1732.

[30] C. L. M. H. Navier, *Mémoire sur les Lois du Mouvement des Fluides*, Mém. Acad. Sci. Inst. de France (2), 6 (1823), pp. 389–440.

[31] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson et Cie, Éditeurs, Paris, 1967.

[32] C. Parés, *Existence, uniqueness, and regularity of solution of the equations of a turbulence model for incompressible fluids*, Appl. Anal., 43 (1992), pp. 245–296.

[33] U. Piomelli and E. Balaras, *Wall-layer models for large-eddy simulations*, in Annual Review of Fluid Mechanics, Ann. Rev. Fluid Mech. 34, Annual Reviews, Palo Alto, CA, 2002, pp. 349–374.

[34] P. Sagaut, *Large Eddy Simulation for Incompressible Flows*, 2nd ed., Sci. Comput., Springer-Verlag, Berlin, 2002.

[35] J. Serrin, *Mathematical principles of classical fluid mechanics*, in Handbuch der Physik, Bd. 8/1, Strömungsmechanik I, Springer-Verlag, Berlin, 1959, pp. 125–263.

[36] V. A. Solonnikov and V. E. Ščadilov, *A certain boundary value problem for the stationary system of Navier–Stokes equations*, in Boundary Value Problems of Mathematical Physics 8, Trudy Mat. Inst. Steklov., 125 (1973), pp. 196–210, 235.

[37] G. G. Stokes, *On the theories of the internal friction of fluid in motion, and of the equilibrium and motion of elastic solids*, Trans. Cambridge Phil. Soc., 8 (1845), pp. 287–305.

[38] C. Truesdell, *The mechanical foundations of elasticity and fluid dynamics*, J. Rational Mech. Anal., 1 (1952), pp. 125–171, 173–300.

[39] M. I. Višik and A. V. Fursikov, *Mathematical Problems of Statistical Hydromechanics*, Kluwer, Dordrecht, The Netherlands, 1980.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1800–1827

ON THE GEOMETRY OF OPTIMAL WINDOWS, WITH SPECIAL FOCUS ON THE SQUARE*

ALMUT BURCHARD† AND JOCHEN DENZLER‡

Abstract. For the Laplace operator with mixed (Dirichlet and Neumann) boundary conditions, the dependence of the principal eigenvalue on the placement of the Dirichlet part is investigated. An optimal window is a Dirichlet part of the boundary that minimizes the principal eigenvalue among all competitors of the same area.

In the special case of a square, we provide both numerical evidence and rigorous partial results for the conjecture that optimal windows in a square are segments centered at either a corner or the midpoint of a side. In particular, we prove that the principal eigenvalue decreases as a window is shifted from a side-centered position towards the corner. An optimal window contained in two sides of the square is connected and contains a corner in its interior. Optimal windows whose length does not exceed the length of one side break the symmetry of the square.

We also construct a star-shaped domain whose optimal window(s) must be disconnected. Finally we give, for general domains in $\mathbb{R}^d$, continuity results for the eigenvalue as a function of the window, and examples of discontinuity when crucial hypotheses are violated. We also give a variation formula that relates the eigenvalue to the singularities of the eigenfunction (stress intensity coefficient) near the boundary of the window.

Methods are based on the variational problem and include rearrangement, Dirichlet–Neumann bracketing, capacity estimates, and deformation under a flow.

Key words. optimal eigenvalue, Laplace operator, mixed boundary conditions, shape optimization, capacity, singular coefficient, rearrangement

AMS subject classifications. 49R50, 35J05, 35R05, 31C40

DOI. 10.1137/S0036141004444184

1. Introduction.

1.1. Overview over the results. Consider the first eigenvalue of the Laplace operator in a fixed domain $\Omega \subset \mathbb{R}^d$ (say, bounded Lipschitz)

$$-\Delta u = \lambda u \,, \quad u \geq 0 \quad \text{in } \Omega \tag{1.1}$$

with Dirichlet boundary conditions on some subset $D \subset \Omega$ and Neumann on the complement of D, i.e.,

$$u|_D = 0, \quad \partial_\nu u|_{\partial\Omega\setminus D} = 0. \tag{1.2}$$

Technical questions of how these boundary conditions should be interpreted will be discussed below. We will call $\lambda = \lambda(D)$ the principal eigenvalue of the Laplacian under the *window boundary conditions on D*. The problem of optimal windows asks for minimization of this eigenvalue for prescribed surface area of the window.

As explained in [8], one may think of Ω as representing a room with perfectly heat-conducting windows at D and insulating walls along $\partial\Omega \setminus D$. The principal eigenvalue $\lambda(D)$ gives the rate of exponential decay of any initial temperature distribution due

*Received by the editors May 31, 2004; accepted for publication (in revised form) July 29, 2005; published electronically February 21, 2006.

http://www.siam.org/journals/sima/37-6/44418.html

†Department of Mathematics, University of Virginia, Charlottesville, VA 22904 (burchard@virginia.edu). This author's research was partially supported by NSF grant DMS-0308040.

‡Department of Mathematics, University of Tennessee, Knoxville, TN 37996 (denzler@math.utk.edu).

to heat diffusion through the window as time becomes large, while the corresponding eigenfunction gives the asymptotic temperature profile. An optimal window minimizes long-term heat loss among all windows of a given size.

It has been shown in [9] that such optimal windows exist and that in the case of a ball of any dimension the optimal window is a spherical cap of the appropriate area. Similar results have been obtained independently by Cox and Uhlig [6], who treat windows as a singular limiting case of Robin boundary conditions.

Here we are concerned with the question of what can be said about the geometry of optimal windows when Ω is not a ball. We suspect that an optimal window in a convex domain Ω should be connected, have some basic regularity properties, and lie in a region of $\partial\Omega$ with large mean curvature. This is certainly not the case for more general domains, as we show by constructing an example of a star-shaped domain with a disconnected optimal window. Heuristic evidence concerning the location of optimal windows has been discussed in [8], and is also corroborated by results of Harrell, Kröger, and Kurata [13] on a different, but related, problem.

As a model case for a convex domain, we study a square: The determination of the shape of optimal windows is already nontrivial in this case. Here we conjecture that the optimal window is a segment, centered either at the midpoint of a side or at a corner depending on the prescribed boundary measure (length), and that there are no other optimal windows, up to sets of measure zero. This conjecture is supported by a number of rigorous partial results as well as numerical evidence (see Figure 2.1). In particular, we prove that the eigenvalue decreases as a short segment is moved from a side-centered position to a position adjacent to a corner and that this monotonicity extends at least for some distance as the window is moved around the corner. We show the first part of this result by means of a Dirichlet–Neumann bracketing argument (section 3.1). The second part is proved by means of an Euler–Lagrange-type variational formula, which we derive for any domain of sufficient regularity in arbitrary dimension (section 6).

Furthermore, we show that some segment containing a corner in its interior is optimal among all windows lying on only two sides of the square (adjacent or not). The proof relies on discrete rearrangement arguments that are specific to the square (with some obvious, but maybe not too interesting, generalization to a cube or hypercube).

Numerical evidence shows that for segments whose length exceeds the sidelength by a certain small amount, up to slightly more than two sidelengths, the corner-centered position ceases to be optimal, with the side-centered position being better (section 2). This can be understood heuristically in terms of the fact that optimal windows prefer to use corners, as was already discussed in terms of a model problem in [8]. However, sacrificing connectedness to distribute the window evenly around the corners is not advantageous and results in windows inferior to either the side-centered segment or the corner-centered segment (section 3.3). For windows up to one sidelength, we can prove this analytically, for larger ones numerically. On the other hand, our analytic results prove that the (nonoptimal) window with four congruent corner-centered components is still better than any other window that has the full symmetry of the square.

Our study of the square also serves as a building block for an example of a star-shaped domain where any optimal window is disconnected (section 4). A related example was discussed heuristically in Figure 3 of [8].

The variation formula mentioned above is derived in section 6 for windows in general domains of any dimension. Its upshot is that the rate of change of the eigenvalue as a function of the window is determined by certain singular coefficients of the eigenfunction that show up at each interface of window and wall on the boundary. In a neighborhood of such interface points, the eigenfunction (albeit in $W^{1,2}$) cannot be expected to be $W^{2,2}$.

In special geometries the singularities have been studied by Grisvard (e.g., [12]). In the situation of a simple interface of a wall and a window segment on a side of the square, the typically expected singular behavior of the eigenfunction is like $c\,\mathrm{Im}\sqrt{z}$ in a neighborhood of 0, where the number c is the singular coefficient. We give a simple lower estimate for this coefficient, based on a maximum principle, to ensure that it does not vanish. In contrast, in a corner of the square, the singular coefficient vanishes. These two facts make it possible for the eigenvalue to be lowered by moving a segment a bit around the corner. For segments up to one sidelength, a better control of the singular coefficients appearing in the variational formula (which depend on global properties of the eigenfunction) should extend this monotonicity all the way until a minimum is achieved when the segment is centered at a corner, but a proof of this extended monotonicity has eluded us so far.

Contributions from geometric singularities (corners, ridges, conical points) have been studied by many authors, and in vast generality, e.g., Maz'ya and Plamenevskii [21]. Surveys are [16] and [23]. We are using only the very simplest case here.

The continuous dependence of the eigenvalue under shifts of the window and other reasonable modifications of its geometry is an intuitively plausible but nontrivial result of relevance. For deformations of windows that can be achieved by the flow of a vector field, our Euler–Lagrange argument proves even differentiability. However, in the absence of good a priori information on the window geometry, such flow type modifications are rather weak; this is why we include some continuity results for other modifications, in general Lipschitz domains (section 5). In this context, it is crucial to consider, in addition to the formulation of the eigenvalue problem (EVP) adopted in [9], [8], a more sophisticated definition that takes into account fine properties of eigenfunctions. It is easy to see that the results in [9], [8] carry over. We will argue this point specifically for the existence of optimal windows in section 1.3. Both definitions coincide for *optimal* windows, as well as for windows of sufficient regularity, in particular for open windows.

1.2. Basic facts, context, and notation; variational formulation. Let us introduce some notation. The symbol Ω will generally denote a bounded Lipschitz domain in $\mathbb{R}^d$, and the window D will be a measurable subset of $\partial\Omega$. As surface measure on $\partial\Omega$, we use $d-1$-dimensional Hausdorff measure, denoted here by σ. A point in $\overline{D} \cap \overline{\partial\Omega \setminus D}$ will be called an *interface point*.

The Laplacian with the window boundary conditions (1.2) will be denoted by Δ_D. The word "eigenvalue" without adjective or ordinal will always denote the *lowest* eigenvalue, which is simple. This eigenvalue will be denoted by $\lambda(D)$.

Define the *optimal eigenvalue* for windows of a given surface measure by

$$\lambda_*(\ell) \;=\; \inf\{\lambda(D) \;\mid\; D \subset \partial\Omega,\ \sigma(D) \;=\; \ell\}\,. \tag{1.3}$$

A set $D \subset \partial\Omega$ will be called an *optimal window* if $\lambda(D) = \lambda_*(\sigma(D))$, that is, if

$$\lambda(D) \;=\; \inf\{\lambda(D') \;\mid\; D' \subset \partial\Omega,\ \sigma(D') \;=\; \sigma(D)\}\,. \tag{1.4}$$

The eigenvalue $\lambda(D)$ in (1.1) and (1.3) can be defined by the Courant–Hilbert variational problem (CHVP)

$$\lambda(D) = \min\left\{\int_\Omega |\nabla u|^2\,dx \;\middle|\; u \in W^{1,2}(\Omega),\ \int_\Omega u^2\,dx = 1,\ u|_D = 0\right\}. \tag{1.5}$$

In [8] and [9], the restriction $u|_D$ of a function $u \in W^{1,2}(\Omega)$ was to be understood as the *trace* in $L^2(\partial\Omega)$. By general Sobolev space theory ([1, Theorem 5.4] or [10, section 4.3, Theorem 1]), the trace of a $W^{1,2}$-function is guaranteed to be an $L^2(\partial\Omega)$ function. Note that the condition $u|_D = 0$, in the L^2-sense, allows the boundary condition to be violated on a set of $d-1$-dimensional measure zero. We refer to this definition of λ as the *coarse* formulation of the eigenvalue problem (1.1) and the CHVP (1.5).

The Neumann boundary conditions on $\partial\Omega \setminus D$ in (1.1) arise as natural boundary conditions for the variational problem in (1.5). The minimizing function u is a normalized eigenfunction corresponding to $\lambda(D)$ and can be chosen to be nonnegative. It agrees a.e. with an analytic function in the interior of Ω, but is not guaranteed to be continuous up to the boundary $\partial\Omega$, unless some assumptions are made on the geometry of D.

Clearly, the principal eigenvalue $\lambda_1(D)$ increases under inclusion of windows,

$$D_1 \subset D_2 \implies \lambda(D_1) \subset \lambda(D_2),$$

since the minimizing function in the CHVP for $\lambda(D_1)$ is an admissible test function for $\lambda(D_2)$.

1.3. Fine variational formulation. As mentioned above, there is another meaningful definition of the boundary conditions (1.2) and the corresponding variational problem (1.5). Since $W^{1,2}$-functions can actually be determined *quasi-everywhere*, that is, up to a set of zero *capacity*, one can insist that the Dirichlet boundary conditions in (1.1) and in the CHVP (1.5) hold up to a set of zero capacity. Since every set of capacity zero has $d-1$-dimensional measure zero, but not vice versa, this is a stronger condition. It corresponds to choosing a smaller domain for the quadratic form associated with Δ_D. We will refer to this definition of $\lambda(D)$ as the *fine* formulation of (1.1) or (1.5). When necessary, we distinguish the two definitions by superscripts, writing $\lambda^c(D)$ and $\lambda^f(D)$ for the coarse and fine eigenvalues, respectively. In general,

$$\lambda^f(D) \;\geq\; \lambda^c(D), \tag{1.6}$$

and it is easy to construct examples where the inequality is strict: any (fractal) window with Hausdorff dimension between $d-2$ and $d-1$ has measure 0 and nonvanishing capacity [10, section 4.7.2],[19, section 2.1.7], hence coarse eigenvalue 0, but positive fine eigenvalue. The notion of sets of capacity 0 is well defined even in two dimensions, where capacity can only be defined subject to some arbitrary choice. Such subtleties can also be avoided by replacing the window D in $\Omega \subset \mathbb{R}^2$ with the equivalent window $D \times [0,1]$ in $\Omega \times]0,1[\subset \mathbb{R}^3$.

In order to relate coarse and fine eigenvalues more precisely, we represent an element u of a Sobolev space by a function defined everywhere, which will be called the preferred representative. For any given Lipschitz domain Ω and any neighborhood V of $\overline{\Omega}$, there is a linear bounded operator $\mathcal{E} : W^{1,2}(\Omega) \to \mathring{W}^{1,2}(V) \hookrightarrow W^{1,2}(\mathbb{R}^d)$ that extends Sobolev functions in Ω to the entire space as outlined in [10, section 4.4], i.e., $\mathcal{E}u|_\Omega = u$. Then we choose the *preferred representative* as

$$\tilde{u}(x) \;:=\; \limsup_{r\to 0} ⨍_{B_r(x)} \mathcal{E}u(y)\,dy \qquad \text{for} \qquad x \in \overline{\Omega}. \tag{1.7}$$

The lim sup is in fact a limit, except on a set of zero capacity, and $\tilde{u}$ is quasi-continuous (as defined in [19, section 2.17]).

The extension operator $\mathcal{E}$ is not unique but depends on a choice of locally flattening coordinate charts, and we make such a choice once and for all for each given Ω. The preferred representative on $\overline{\Omega}$ depends on the extension operator, but any two choices will only differ on a set of capacity zero. See [19, Theorem 2.55 with following remark] or [10, section 4.8, Theorem 1]. The restriction of $\tilde{u}$ to $\partial\Omega$ represents the trace of u.

THEOREM 1.1. *For the CHVP* (1.5) *in the fine formulation, there exists a minimizer which is uniquely determined quasi-everywhere up to choice of a sign.*

Proof. This is a slight modification of the classical argument for the existence of a minimizer for (1.5) in the (coarse) Sobolev sense. Let u_j be a minimizing sequence of quasi-continuous functions (in $W^{1,2}(\mathbb{R}^d)$, by extension) satisfying the boundary conditions in the fine sense. Extracting a subsequence (again denoted by u_j), we may assume weak convergence in $W^{1,2}(\Omega)$, strong convergence in $L^2(\Omega)$, and (by compactness of the trace map) strong convergence in $L^2(\partial\Omega)$, to a limit function u_*. We have to show that u_* inherits the fine boundary conditions from $\{u_j\}$. To this end, we replace the sequence $\{u_j\}$ by a sequence $\{\bar{u}_j\}$ of convex combinations that converges *strongly* in $W^{1,2}$, according to Mazur's theorem (see, e.g., [18, section 2.13]). The normalized sequence $\hat{u}_j := \bar{u}_j/\|\bar{u}_j\|_{L^2(\Omega)}$ still converges strongly in $W^{1,2}(\Omega)$ because $\|\bar{u}_j\|_{L^2(\Omega)} \to 1$. The $\hat{u}_j$ inherit the fine boundary conditions from u_j and therefore form a sequence of legitimate competitors in the CHVP.

Now by convexity we obtain

$$\int |\nabla u_*|^2 = \lim \int |\nabla \hat{u}_j|^2 = \lim \int |\nabla \bar{u}_j|^2 \leq \lim \int |\nabla u_j|^2 = \inf \int |\nabla u|^2.$$

We must show that u_* inherits the fine boundary conditions from the $\hat{u}_j$. This follows from the arguments in section 2.1.3 of [19], which we sketch briefly, for the sake of being self-contained:

(1) If a sequence of C_0^∞ functions v_j (with uniformly bounded support) converges strongly in $W^{1,2}(\Omega)$ to some u_*, then it holds for some subsequence (again called v_j):

$$\forall \varepsilon > 0\ \exists V_\varepsilon \text{ open}: \operatorname{cap}(V_\varepsilon) < \varepsilon,\ \|v_j - u_*\|_{C^0(\overline{\Omega}\setminus V_\varepsilon)} \to 0$$

(2) Every $W^{1,2}(\Omega)$ function u can be approximated in $W^{1,2}(\Omega)$ norm by C_0^∞ functions v_k (with uniformly bounded support), such that

$$\forall \varepsilon > 0\ \exists W_\varepsilon \text{ open}: \operatorname{cap}(W_\varepsilon) < \varepsilon,\ \|v_k - u\|_{C^0(\overline{\Omega}\setminus W_\varepsilon)} \to 0.$$

(The existence of a quasi-continuous representative is actually a consequence of this.)

Now there are open sets V_j with $\operatorname{cap}(V_j) < 2^{-j}$ such that $\hat{u}_j$ is continuous on $\overline{\Omega} \setminus V_j$ and vanishes on $D \setminus V_j$, and there are smooth approximants $\hat{v}_j$ such that $\|\hat{v}_j - \hat{u}_j\|_{W^{1,2}(\Omega)} < 2^{-j}$ and $\|\hat{v}_j - \hat{u}_j\|_{C^0(\overline{\Omega}\setminus V_j \setminus W_j)} < 2^{-j}$ for appropriate open sets W_j with $\operatorname{cap}(W_j) < 2^{-j}$. Therefore, for every j_0, the sequence $\hat{v}_j$ converges uniformly to u_* outside the set $\overline{V}_{j_0} := \bigcup_{j \geq j_0}(V_j \cup W_j)$, whose capacity is at most 2^{2-j_0}. Hence u_* vanishes on $D \setminus \overline{V}_{j_0}$, for every j_0.

Uniqueness and positivity follow from the strong maximum principle as in the classical argument. ∎

The coarse and fine formulations of *optimal* windows and their eigenfunctions essentially agree.

PROPOSITION 1.2. *Let $\lambda_*^c(\ell)$ and $\lambda_*^f(\ell)$ be optimal eigenvalues for windows of size ℓ, as defined by* (1.3) *in the coarse and fine sense, respectively. Then*

$$\lambda_*^f(\ell) \;=\; \lambda_*^c(\ell), \qquad (0 \leq \ell \leq \lambda_{\mathrm{Dir}}). \tag{1.8}$$

Furthermore, D is an optimal window with respect to the coarse definition if and only if it differs by a set of $d-1$-dimensional measure zero from an optimal window for the fine formulation.

Proof. Clearly, from (1.6) we have

$$\lambda_*^f(\ell) \geq \lambda_*^c(\ell).$$

To see the converse inequality, take an optimal window D for the coarse formulation, i.e., $\lambda^c(D) = \lambda_*^c(\ell)$, and let u^c be a minimizer of the corresponding CHVP (1.5). Let $\tilde{u}^c$ be the preferred representative of u^c, as defined above, and set

$$D' := \{x \in \partial\Omega \mid \tilde{u}^c(x) = 0\}.$$

We refer to this procedure as *refining* the window D. By definition, $\sigma(D') \geq \ell$ and $\lambda^c(D') = \lambda_*(\ell)$. Since $\tilde{u}^c|_{D'}$ vanishes identically, it is an admissible candidate for the CHVP (1.5) for $\lambda^f(D')$. It follows that

$$\lambda_*^f(\ell) \;\leq\; \lambda^f(D') = \lambda^c(D) = \lambda_*^c(\ell). \qquad \square$$

Note that we always have $\lambda^f(D') \leq \lambda^c(D)$ since $\tilde{u}^c$ vanishes on D'. Whenever $\lambda^f(D) > \lambda^c(D)$ occurs, this is due to $D \setminus D'$ having positive capacity, which makes $\tilde{u}^c$ ineligible for the fine CHVP. Whenever de Giorgi's continuity argument applies at each point of D, i.e., when u^c has a representative that is continuous on $\Omega \cup D$, then $\tilde{u}^c$ is admissible for the fine CHVP and thus $\lambda^f(D) = \lambda^c(D)$. This holds in particular if D is open, notwithstanding possible discontinuities of u^c at interface points. A de Giorgi argument can also be used to show continuity of u^c, provided the window has positive density at every interface point $p \in \overline{D} \cap \overline{\partial\Omega \setminus D}$. We suspect that eigenfunctions for optimal windows should be continuous up to the boundary, but this is an unresolved question.

2. Numerical results for the square. We have used the MATLAB pdetool to calculate, by means of finite elements, the lowest eigenvalue for various window configurations. The calculation was done with a sequence of at least three subsequent mesh refinements so that numerical convergence within the precision of the graphics could be checked by inspection. In the accompanying Figure 2.1, we show the eigenvalue as a function of the length of the window, for five different simple geometric configurations.

As outlined in section 1.1, we conjecture that the configurations giving the lowest eigenvalue in Figure 2.1 (namely either a side-centered or a corner-centered segment, depending on the length) are in fact the optimal configurations. As a rule of thumb, the better of the two choices of symmetric and connected windows is the one that contains more corners. Exceptions to this rule occur near integer multiples of a sidelength.

Figure 2.1 also displays a feature of the first variation formula: When the interface points are in the corner (which implies vanishing of the singular coefficients), the derivative of the eigenvalue vanishes. These are just the explicitly calculable cases marked in the figure. Our numerics do not resolve the modulus of continuity at length 0. However, for each of the curves printed, it can be seen analytically that $c_1/\ln(1/\delta) \leq \lambda(\delta) \leq c_2/\ln(1/\delta)$ (with δ the total length of the window): The lower bound follows from Theorem 5 in [8] (slightly modified for two dimensions, as pointed out there); the upper bound is an immediate consequence of our capacity estimate in Proposition 5.2 and its proof. Sharp asymptotics for a different, but closely related, problem can be found in Chapter 9 of [20].

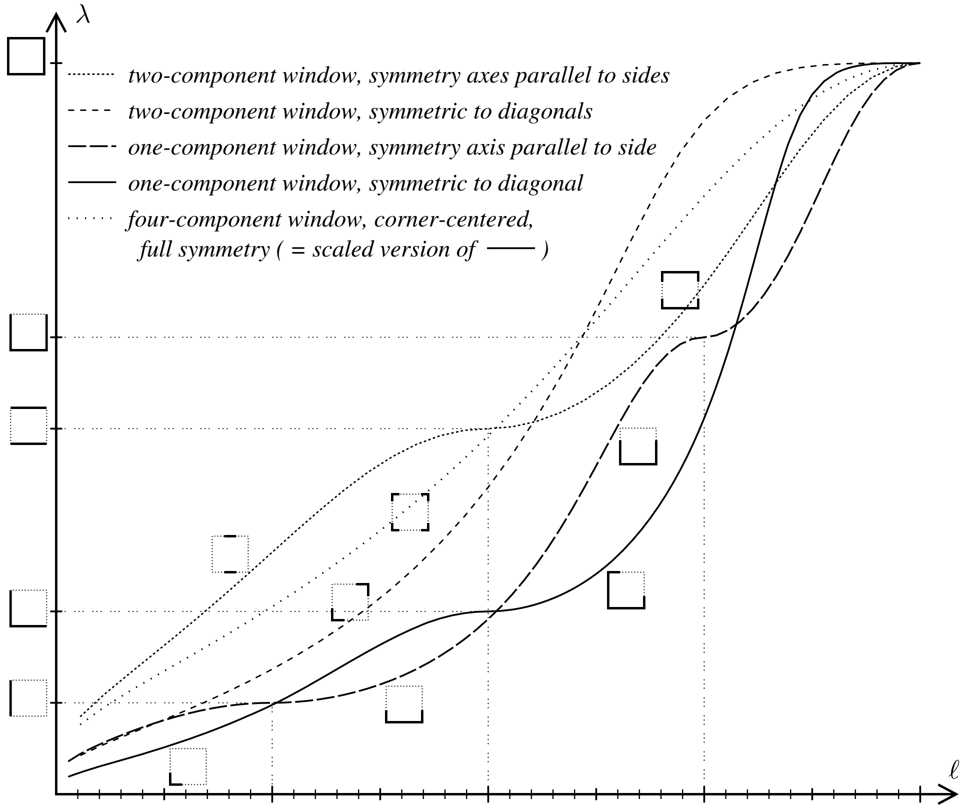


Fig. 2.1. *Five sections through the space of windows, each parametrized by length. The horizontal axis measures the length, in units of the perimeter of the square, the vertical axis gives the eigenvalue relative to the full Dirichlet eigenvalue. The labels mark those window configurations that can be calculated explicitly by separation of variables. In these pictograms, bold lines denote Dirichlet BCs.*

We next study the dependence of the eigenvalue on the position of the window. We cannot expect the eigenvalue to depend monotonically on the shift parameter for all lengths, because the side-centered and the corner-centered configuration yield the same eigenvalue for three particular lengths, close to 1.02, 2.04, 3.15 sidelengths, as seen in Figure 2.1. In Figure 2.2, we shift windows of a given length from a side-centered to a corner-centered position. However, we observe in each case that local minima only occur in symmetric positions, supporting our conjecture.

The derivative of the eigenvalue with respect to the shift parameter is proportional to the difference of the squares of the singular coefficients at the endpoints of the segment (section 6). In the symmetric configurations, this derivative vanishes by symmetry. When both endpoints lie in a corner, where the singular coefficients vanish, the derivative appears to vanish to a higher order, indicating further cancellations. This can be plausibly observed for lengths 1 and 3 in the side-centered configuration, and for length 2 in the corner-centered configuration.

Figure 2.3 shows the effect of tearing apart a connected window into two pieces. Note the competition between corner positions and connectedness as geometric features favoring low eigenvalues.

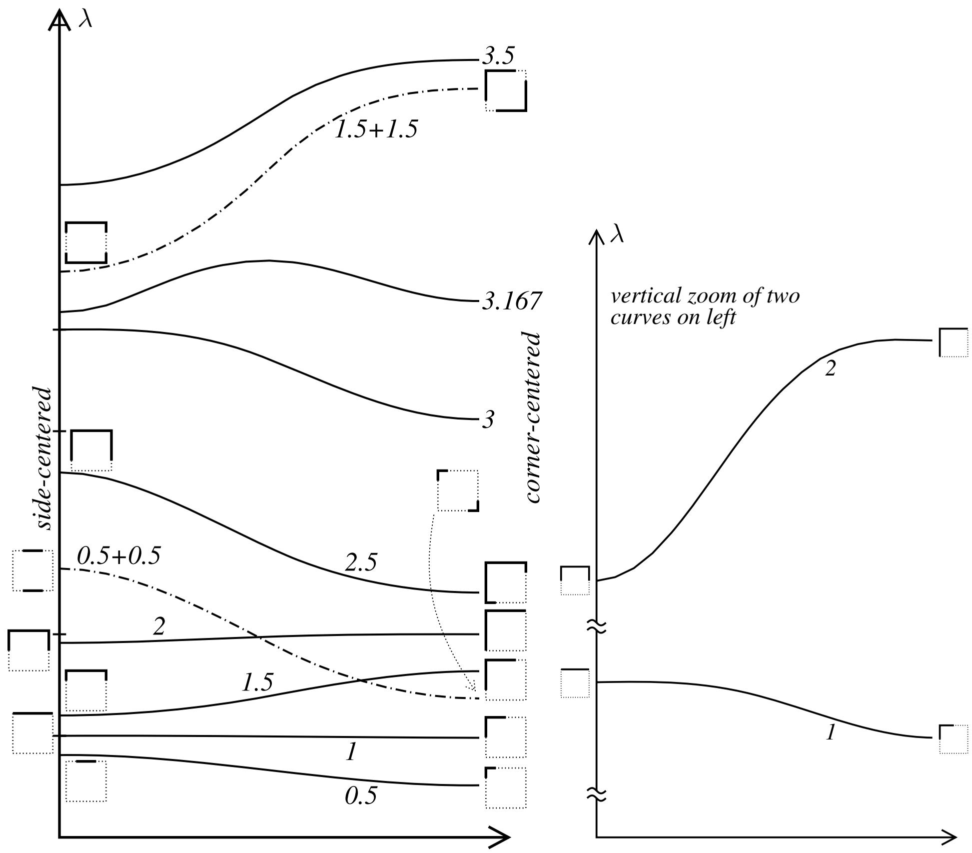


Fig. 2.2. *The eigenvalue for connected windows of nine different lengths and two-component windows of two different lengths, as a function of a shift parameter. The horizontal axis indicates the distance to a side-centered position.*

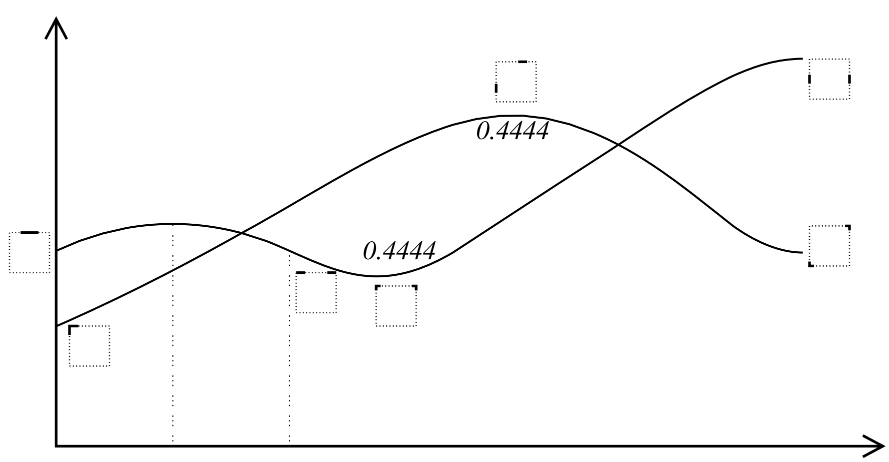


Fig. 2.3. *Examples of competition between connectedness and corner position for small windows. Dotted lines: The first hump for the side-centered window is actually axially symmetric, due to the argument used in section* 3.1.

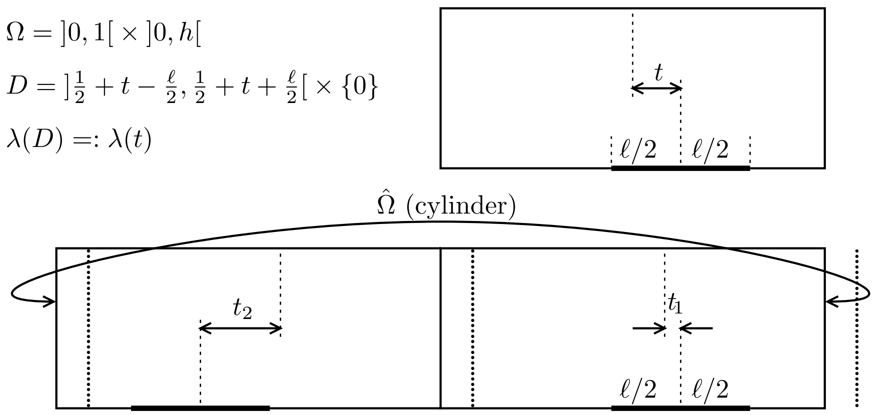


FIG. 3.1. *Shifting windows in a rectangle after doubling it to obtain a cylinder.*

3. Rigorous results for the square. In this section, we collect some inequalities and monotonicity results that are specific to the square.

3.1. Monotonicity of shifting (rectangle). For the geometric situation, see the top of Figure 3.1.

THEOREM 3.1. *Let Ω be a rectangle. The principal eigenvalue of a connected window D that is contained entirely in one side of $\partial\Omega$ is a continuous, strictly decreasing function of the distance of D from the side-centered position.*

Proof. By scaling, rotating, and translating, we may assume that $\Omega =]0,1[\times]0,h[$, and that the window is contained in the bottom side of the rectangle; see Figure 3.1. For $0 < \ell < 1$ and $|t| \le (1-\ell)/2$, let $D(t) =]\frac{1}{2}+t-\frac{\ell}{2}, \frac{1}{2}+t+\frac{\ell}{2}[\times \{0\}$ be the window of length ℓ that has been shifted by t from the side-centered position, and denote the corresponding eigenvalue by $\lambda(t)$. By symmetry, λ is an even function of t. Continuity of $\lambda(t)$ follows most easily from Proposition 5.2. To prove the last assertion, we will show that

$$\lambda\big(\tfrac{1}{2}(t_1+t_2)\big) \; > \; \min\{\lambda(t_1), \lambda(t_2)\} \tag{3.1}$$

holds for any pair $t_1 < t_2$. Setting $t_1 = -t$, $t_2 = t$ in (3.1) and using that $\lambda(-t) = \lambda(t)$ shows that λ takes its global maximum at $t = 0$. Setting $t_{1/2} = t \mp \varepsilon$ in (3.1) shows that $\lambda(t)$ cannot assume a local minimum on the open interval $]0,(1-\ell)/2[$. We conclude that $\lambda(t)$ is strictly decreasing on $[0,(1-\ell)/2]$, as claimed.

In order to prove claim (3.1), we combine a doubling trick with a special case of Dirichlet–Neumann bracketing (see [22, section XIII.15]). Fix $t_1 \le t_2$ with $|t_1|, |t_2| \le (1-\ell)/2$, set $t = (t_1+t_2)/2$, and let u be the positive normalized eigenfunction for the rectangle $\Omega =]0,1[\times]0,h[$ with the window $D(t)$.

Consider the CHVP on the cylinder $\hat{\Omega} = (\mathbb{R}/2\mathbb{Z}) \times]0,h[$, with window $\hat{D}(t) = D(t) \cup (-D(t))$, which is obtained by gluing a copy of Ω with window $D(t)$ to its mirror image along the vertical edges. Since the minimizing function $\hat{u}$ is automatically symmetric by simplicity of the principal eigenvalue in a connected domain, it follows that this CHVP on $\hat{\Omega}$ with window $\hat{D}(t)$ is equivalent to the CHVP on Ω with window $D(t)$. In particular,

$$\hat{u}(x,y) \;=\; \frac{1}{\sqrt{2}} u(|x|,y), \quad (-1 \le x \le 1,\; 0 \le y \le h),$$

and the corresponding principal eigenvalue coincides with $\lambda(t)$.

On the other hand, with the understanding that x coordinates are interpreted modulo 2, the cylinder $\hat{\Omega}$ contains the disjoint union of the rectangles $\Omega_1 =](t_2 - t_1)/2, 1 + (t_2 - t_1)/2[\times]0, h[$ and $\Omega_2 =]-1 + (t_2 - t_1)/2, (t_2 - t_1)/2[\times]0, h[$, which are copies of Ω with windows $D(t_1)$ and $D(t_2)$, respectively. By restricting $\hat{u}$ to $\Omega_1 \cup \Omega_2$ we obtain a test function for $\Omega_1 \cup \Omega_2$ with window $D(t_1) \cup D(t_2)$. Since the principal eigenvalue for a disjoint union of domains is the smaller of the two eigenvalues, it follows that

$$\lambda(t) \geq \min\{\lambda(t_1), \lambda(t_2)\}.$$

The functions $u_1 = \hat{u}|_{\Omega_1}$ and $u_2 = \hat{u}|_{\Omega_2}$ cannot be eigenfunctions for $\lambda(t_1)$ and $\lambda(t_2)$, because the gradient of $\hat{u}$ vanishes at those boundary points of Ω_1 or Ω_2 that were corners of Ω, in violation of the Hopf boundary point lemma for u_1 and u_2. This completes the proof of (3.1). $\square$

The doubling argument used in the proof shows that two connected window segments of length ℓ each, placed symmetrically with distance $2s$ from the center of a rectangle of sidelength 2, yield the same eigenvalue as two such windows placed symmetrically with distance $2(1 - \ell - s)$ apart. This can be observed in the curve for the side-centered configurations in Figure 2.3.

3.2. Optimality among windows on one or two sides. The monotonicity argument in the previous subsection implies in particular that among all connected windows contained in *one* side of a rectangle, the one touching a corner produces the minimal eigenvalue. We next consider windows contained in two sides of a square.

THEOREM 3.2. *Among all windows that lie on only two sides of the square (adjacent or not), the optimal window is connected and contains a corner of the square in its interior.*

The proof is based on rearrangement techniques, which have been widely used for geometric inequalities (see [15], [18] for a general reference). Here we will use two rearrangements adapted to the square: the increasing rearrangement and polarization.

For a nonnegative measurable function u on a rectangle, we define the *increasing rearrangement in the x-direction*, $\mathcal{R}u$, by replacing the restriction of u to each line $y = const$ with the unique nondecreasing left-continuous function which is equimeasurable with $u(\cdot, y)$; see Figure 3.2. By Fubini's theorem, $\mathcal{R}u$ is equimeasurable with u.

LEMMA 3.3. *Let u be a nonnegative $W^{1,2}$-function on a rectangle $\Omega =]0,1[\times]0,h[$, and let $\mathcal{R}u$ be its increasing rearrangement in the x-direction. If the trace of u vanishes σ-a.e. on a window*

$$D = \Big(\{0\} \times D_l\Big) \cup \Big(\{1\} \times D_r\Big) \cup \Big(D_b \times \{0\}\Big) \cup \Big(D_t \times \{h\}\Big),$$

then $\mathcal{R}u$ vanishes σ-a.e. on the window $\mathcal{R}D$ defined by

$$(\mathcal{R}D)_l = D_l \cup D_r, \quad (\mathcal{R}D)_r = \emptyset, \quad (\mathcal{R}D)_b =]0, \sigma(D_b)[, \quad (\mathcal{R}D)_t =]0, \sigma(D_t)[. \tag{3.2}$$

In general,

$$\sigma(D) \geq \sigma(\mathcal{R}D),$$

with equality certainly when $D_r = \emptyset$. The corresponding principal eigenvalues satisfy

$$\lambda(D) \geq \lambda(\mathcal{R}D),$$

with equality only when $\mathcal{R}D$ agrees σ-a.e. with either D or its mirror image.

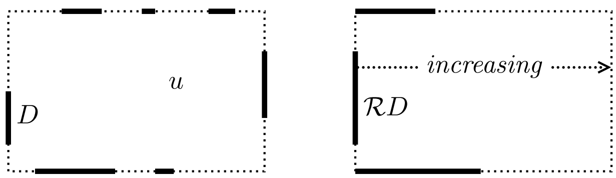


FIG. 3.2. *The effect of the symmetric increasing rearrangement on a window with components on all four sides of a rectangle.*

Proof. If u is continuous up to the boundary of the rectangle and vanishes on D, then its rearrangement $\mathcal{R}u$ is also continuous and vanishes on the window $\mathcal{R}D$. To see that the trace of $\mathcal{R}u$ vanishes on $\mathcal{R}D$ for any nonnegative function u in $W^{1,2}$ vanishing on D, we note that the increasing rearrangement is closely related with Steiner symmetrization. In fact, if we extend both u and $\mathcal{R}u$ by reflection across the line $x = 1$ to functions $\hat{u}$ and $\hat{\mathcal{R}}u$ on the doubled rectangle $\hat{\Omega} =]0, 2[\times]0, h[$, then $\hat{\mathcal{R}}u$ is just the Steiner symmetrization of $\hat{u}$. Since Steiner symmetrization is continuous on $W^{1,2}$ according to [4], $\mathcal{R}$ is continuous as well, and the first claim follows by a density argument.

For the second claim we use that $\mathcal{R}$ preserves the L^2-norm but reduces the norm of the gradient. In particular, $\mathcal{R}$ can only decrease the Rayleigh quotient. Choosing u to be the principal eigenfunction of the CHVP corresponding to the window D, we see that

$$\lambda(D) = \frac{\int |\nabla u|^2 \, dx}{\int |u|^2 \, dx} \geq \frac{\int |\nabla \mathcal{R}u|^2 \, dx}{\int |\mathcal{R}u|^2 \, dx} \geq \lambda(\mathcal{R}D). \tag{3.3}$$

By analyticity, the partial derivative $\partial_x u$ vanishes only on a set of zero measure. It follows from a theorem of Brothers and Ziemer [3] that the rearrangement inequality in (3.3) is strict unless u is already either increasing or decreasing in x on each line $y = const.$ ∎

The second rearrangement exploits the symmetry of the square under reflections at the diagonals. Let $\Omega =]0, 1[\times]0, 1[$ be the unit square, and let $\tau(x, y) = (y, x)$ denote the reflection at the diagonal joining the lower left with the upper right-hand corner. For any function u on Ω, the *polarization* $\mathcal{P}u$ of u with respect τ is given by

$$\mathcal{P}u(x, y) \;=\; \begin{cases} \max\{u(x, y), u(\tau(x, y))\} & \text{if } y \geq x, \\ \min\{u(x, y), u(\tau(x, y))\} & \text{if } y \leq x. \end{cases}$$

For a comprehensive account of polarization we refer to [2]. We have the following lemma.

LEMMA 3.4. *Let u be a nonnegative $W^{1,2}$-function on the unit square $\Omega =\,]0, 1[\times]0, 1[$, and let $\mathcal{P}u$ be its polarization, as defined above. If (the trace of) u vanishes σ-a.e. on a window*

$$D \;=\; \Big(\{0\} \times D_l\Big) \cup \Big(\{1\} \times D_r\Big) \cup \Big(D_b \times \{0\}\Big) \cup \Big(D_t \times \{1\}\Big),$$

then $\mathcal{P}u$ vanishes σ-a.e. on the window $\mathcal{P}D$ with

$$(\mathcal{P}D)_l = D_l \cap D_b, \quad (\mathcal{P}D)_r = D_r \cup D_t, \quad (\mathcal{P}D)_b = D_l \cup D_b, \quad (\mathcal{P}D)_t = D_r \cap D_t.$$

In general,

$$\sigma(D) = \sigma(\mathcal{P}D),$$

and the principal eigenvalues satisfy

$$\lambda(D) \geq \lambda(\mathcal{P}D),$$

with equality only if $\mathcal{P}D$ agrees σ-a.e. with either D or $\tau(D)$.

Proof. The form of $\mathcal{P}D$ is immediate from the definition of $\mathcal{P}$. To see the second claim, choose u to be the principal eigenfunction corresponding to the window D. Since $\mathcal{P}u$ is equimeasurable with u and $|\nabla \mathcal{P}u|$ is equimeasurable with $|\nabla u|$ by definition of the polarization, we have

$$\lambda(D) = \frac{\int |\nabla u|^2\, dx}{\int |u|^2\, dx} = \frac{\int |\nabla \mathcal{P}u|^2\, dx}{\int |\mathcal{P}u|^2\, dx} \geq \lambda(\mathcal{P}D).$$

Unless $\mathcal{P}u$ agrees with either u or $u \circ \tau$, it cannot be real analytic, and hence is not the eigenfunction corresponding to $\lambda(\mathcal{P}D)$. We conclude that then the last inequality is strict. ∎

Proof of Theorem 3.2. Within the class of windows contained in two sides of the square, there clearly exists an optimal one. By Lemma 3.4, a window consisting of two nonempty parts contained in two opposite sides of the square cannot be optimal, since it can be improved by polarization.

If D is contained in two adjacent sides (say, left and bottom) of the square, Lemma 3.3 implies that replacing D with $\mathcal{R}D$ strictly reduces the principal eigenvalue, unless the bottom part of the window is connected and contains a corner. Note that in this case, $\mathcal{R}D$ has the same length as D. Repeating this argument for the vertical direction, we see that also the part of D on the left-hand side must be connected and contain the lower left corner.

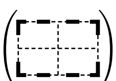
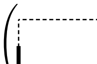
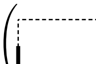
It remains to show that a corner must lie in the *interior* of the window. If the length of D happens to equal the length of one side of the square, we refer to the numerical result, which shows that the corner-centered position improves over the one-side position. Otherwise, we refer to Corollary 6.3 below to show that moving the segment a short distance around the corner improves the eigenvalue. ∎

3.3. Nonoptimality of $\mathbb{Z}_2 \times \mathbb{Z}_2$-symmetric windows. We have the following theorem.

THEOREM 3.5. *In a rectangle, any window of sufficiently small length that has the full symmetry group of a rectangle is not optimal. In particular, a symmetric window whose length does not exceed the length of one side is not optimal in a square.*

As mentioned before, numerical results for the square indicate that the length restriction is not needed.

Proof. In self-explanatory pictogram notation, we reason that

(3.4) $\lambda\Bigg($  $\Bigg) = \lambda\Bigg($  $\Bigg) = 4\lambda\Bigg($  $\Bigg) \geq 4\lambda\Bigg($  $\Bigg)$,

exploiting symmetry, scaling, and the rearrangement of Lemma 3.3 in turn. The last inequality is strict unless the window consists of four L-shaped windows in the corners to begin with, showing that an optimal window having full symmetry must be of that form. In (3.4), we have gained a factor 4, but lost half of the window length. We now double the window using Lemma 6.1.

Assume that the rectangle has the form $\Omega =]0, a[\times]0, b[$ and the L-shaped window (called D_L) has lengths $q_x a$ and $q_y b$ on the horizontal and vertical parts, respectively.

An admissible test function for the CHVP for an L-shaped window with sidelengths $2q_xa$ and $2q_yb$ is given by $u\circ\psi$, where $\psi(x,y)=(h(x),k(y))$ with h,k piecewise linear such that $h(0)=k(0)=0$, $h(a)=a$, $k(b)=b$, and $h(2q_xa)=q_xa$, $k(2q_yb)=q_yb$. It is easy to see that $\psi:\Omega\to\Omega$ is bi-Lipschitz. The largest value for the spectral radius of $(D\psi)(D\psi)^T/\det D\psi$ is $2(1-q)/(1-2q)$ with $q=\max\{q_x,q_y\}$, and the largest value for $\det D\psi$ is $(1-q_x)/(1-2q_x)\times(1-q_y)/(1-2q_y)$. By Lemma 6.1, the window $\psi^{-1}(D_L)$ is an improvement over the original window, whenever $2(1-q)^3/(1-2q)^3\le 4$, which happens for $q<0.17$ and translates to a smallness condition on the window size, depending on the sidelengths of the rectangle.

In the square $]0,1[\times]0,1[$, an optimal window which is symmetric under reflections at the vertical and horizontal axes must be symmetric under reflection in the diagonals as well, since otherwise a better window is obtained by polarization; this gives $q_x=q_y=:q$. We can now get a better quantitative estimate in (3.4) for the square. Define a bi-Lipschitz map by setting

$$\psi:(x,y)\mapsto\begin{cases}\left(x,1-\frac{1-q}{1-2q}(1-y)\right) & \text{if } y\ge 1-(1-2q)(1-x) \qquad \text{(I)},\\ \left(x,\frac{1}{2}(x+y)\right) & \text{if } x\le y\le 1-(1-2q)(1-x) \quad \text{(II)}\end{cases}$$

above the diagonal, and an analogous formula below the diagonal. The spectral radius of $(D\psi)(D\psi)^T/\det D\psi$ is $(1-q)/(1-2q)$ in domain (I) and $\frac{1}{2}(3+\sqrt{5})$ in (II). The Jacobian $\det D\psi$ is largest in (I), namely $(1-q)/(1-2q)$. Lemma 6.1 implies that the window can be doubled with a factor ≤ 4 in the eigenvalue, provided $\sigma(D)=4q\le 4(5-\sqrt{5})/(13-\sqrt{5})\approx 1.027$. ∎

COROLLARY 3.6. *The result of Theorem* 3.5 *holds, for sufficiently small windows in a rectangle* $]-a,a[\times]-b,b[$, *under the weaker assumption that either* (a) *there is equal window area in each of the four quadrants, or* (b) *the window is symmetric under the* $180°$ *rotation* $(x,y)\mapsto(-x,-y)$.

Proof. For (a), the first step in (3.4) can be replaced with an inequality, where that quarter is selected that contributes the smallest Rayleigh quotient. For (b), note that the symmetry is inherited by the eigenfunction, and we have $u(0,y)=u(0,-y)$. Thus we can define $\hat u\in W^{1,2}$ by $\hat u(x,y)=u(x,y)$ for $x\le 0$ and $\hat u(x,y)=u(x,-y)$ for $x\ge 0$. $\hat u$ represents another window $\hat D$ with the same area as D, has the same Rayleigh quotient, and is not the optimizer yet, unless $\hat D=D$; this reduces the corollary to the theorem again. ∎

4. A star-shaped domain with disconnected optimal window. Here we prove the properties of the following example.

Example 4.1. There exists a star-shaped Lipschitz domain Ω in $\mathbb{R}^2$ and a length ℓ such that a connected window of length ℓ in Ω cannot be optimal.

Proof. In a one-parameter family of domains Ω_ε, we calculate an upper bound for the eigenvalue of a certain window D_2 with two components. Then we establish a larger lower bound for the eigenvalue of any connected window D. These estimates, based on Dirichlet–Neumann bracketing, work for sufficiently small ε, and can be made quantitative.

Ω_ε is the union of a "torso" rectangle T_ε and a pair of "handles" H_ε, $-H_\varepsilon$:

$$T_\varepsilon:=\,]-1,1[\times]-1-\varepsilon,1+\varepsilon[,\quad H_\varepsilon:=[1,9-\varepsilon[\times]-\varepsilon,\varepsilon[. \tag{4.1}$$

See the top left part of Figure 4.1. We choose $D_2:=(\partial\Omega_\varepsilon)\setminus\overline{T}_\varepsilon$, with $\sigma(D_2)=32$. The remaining boundary $W:=(\partial\Omega_\varepsilon)\setminus D_2$ has measure $\sigma(W)=8$.

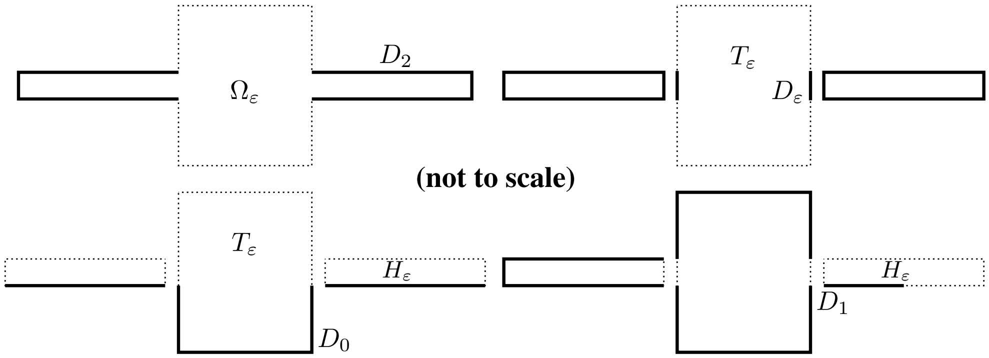


FIG. 4.1. *Top left: A star-shaped Lipschitz domain whose optimal window(s) of a certain length ℓ cannot be connected. Top right: Upper bound for eigenvalue of disconnected window. Bottom: Lower bounds for connected windows.*

For comparison, disconnect the handles from the torso by means of extra Dirichlet boundary $D_\varepsilon = \{\pm 1\} \times [-\varepsilon, \varepsilon]$, as in the top right of Figure 4.1. With fewer competitors in the CHVP (1.5), we get an upper bound. In self-explanatory notation, we conclude

$$\lambda(\Omega_\varepsilon, D_2) < \min\Big\{\lambda_{\mathrm{Dir}}(H_\varepsilon), \lambda(T_\varepsilon, D_\varepsilon)\Big\} = \lambda(T_\varepsilon, D_\varepsilon).$$

By testing the EVP for T_ε with $\sin\frac{\pi}{2}(|y|-\varepsilon)_+$, one can see that the evaluation of the minimum is valid for all $\varepsilon \leq 1$.

For any connected window D of length 32, it can easily be seen that, except for reflection symmetry, either $D \supset D_0$ or $D \supset D_1$, where

$$\begin{aligned} D_0 &= \{(x, y) \in \partial\Omega_\varepsilon \mid y \leq -\varepsilon\}, \\ D_1 &= \{(x, y) \in \partial\Omega_\varepsilon \mid x \leq 1\} \cup [1, 5] \times \{-\varepsilon\}. \end{aligned}$$

To get lower bounds for $\lambda(\Omega_\varepsilon, D_0)$ and $\lambda(\Omega_\varepsilon, D_1)$, disconnect the handles from the torso by means of extra Neumann boundary $\{\pm 1\} \times [-\varepsilon, \varepsilon]$. With slight abuse of notation, we write $\lambda(T_\varepsilon, D_i)$ for $\lambda(T_\varepsilon, D_i \cap \partial T_\varepsilon)$, and similarly for H_ε. We have either

$$\lambda(\Omega_\varepsilon, D) > \min\Big\{\lambda(H_\varepsilon, D_0), \lambda(T_\varepsilon, D_0)\Big\} = \lambda(T_\varepsilon, D_0) > (\tfrac{\pi}{4+4\varepsilon})^2 \tag{4.2}$$

or

$$\lambda(\Omega_\varepsilon, D) > \min\Big\{\lambda(H_\varepsilon, D_1), \lambda(T_\varepsilon, D_1)\Big\} = \lambda(H_\varepsilon, D_1). \tag{4.3}$$

The evaluation of the minimum in (4.2), for any ε, relies on a test function that vanishes for $y \leq -\varepsilon$. The evaluation of the minimum in (4.3) is valid for all $\varepsilon < \frac{3}{2}$, since then, using comparison functions $\cos(\pi y/(2+2\varepsilon))$ and $\sin\pi(x-5)_+/2(4-\varepsilon)$,

$$\lambda(T_\varepsilon, D_1) > (\pi/(2+2\varepsilon))^2 \geq (\pi/(2(4-\varepsilon)))^2 > \lambda(H_\varepsilon, D_1).$$

For $\varepsilon < \frac{2}{3}$, we can also conclude that

$$\lambda(H_\varepsilon, D_1) < (\pi/(8-2\varepsilon))^2 < (\pi/(4+4\varepsilon))^2 < \lambda(T_\varepsilon, D_0).$$

It therefore only remains to prove the middle inequality in

$$\lambda(\Omega_\varepsilon, D) > \lambda(H_\varepsilon, D_1) > \lambda(T_\varepsilon, D_\varepsilon) > \lambda(\Omega_\varepsilon, D_2).$$

But as $\varepsilon \to 0$, one has $\lambda(T_\varepsilon, D_\varepsilon) \to 0$, whereas $\lambda(H_\varepsilon, D_1) \to (\pi/8)^2$. This plausible singular domain limit can be proved in a straightforward way by writing the quadratic form $\int_{H_\varepsilon}(u_x^2 + u_y^2)\, dx\, dy$ as a quadratic form $\int(\varepsilon\chi(\xi)^{-1}u_\xi^2 + \varepsilon^{-1}\chi(\xi)u_\eta^2)\, d\xi\, d\eta$ on L^2 with measure $\varepsilon\chi(\xi)d\xi\, d\eta$ in a fixed reference domain $]0,8[\times]-1,1[$. Here $\chi(\xi) = 1$ for $\xi < 4$, and $\chi(\xi) = 1 - \varepsilon/4$ for $\xi > 4$. If we carry out the limit $\varepsilon \to 0$ in the CHVP with the appropriate eigenfunctions, we have the uniform upper bound $(\pi/(8-2\varepsilon))^2$ for the eigenvalue, as mentioned before. This controls the $W^{1,2}$ norm in the fixed domain, and actually enforces $u_\eta \to 0$. The limiting function will indeed not depend on the η coordinate and solve the 1-dimensional eigenvalue problem $-u_{\xi\xi} = \lambda u$ on $[4,8] \ni \xi$, with $u(4) = 0$, $u_\xi(8) = 0$. ∎

5. Some continuity results. In this section, we study how the eigenvalue changes if a window of a particular size is added at a particular location. The basic philosophy is that windows can be added more cheaply at locations where the eigenfunction was already small before the addition. In the second subsection, we discuss related continuity properties of the corresponding eigenfunctions.

5.1. Continuity of eigenvalues. Our first result is an estimate for the increase of the principal eigenvalue if a set of small capacity is added to a given window.

LEMMA 5.1. *Let $D_2 \supset D_1$ and let u_1 be the normalized eigenfunction for D_1. Let G be a domain containing $D_2 \setminus D_1$; in case the dimension $d = 2$, assume additionally that G is bounded. Then*

$$\lambda(D_2) - \lambda(D_1) \le \frac{\lambda(D_1)\,\mathrm{vol}\,(G\cap\Omega) + \mathrm{cap}(D_2 \setminus D_1, G)}{1 - (\sup_{G\cap\Omega} u_1)^2\,\mathrm{vol}\,(G\cap\Omega)} \big(\sup_{G\cap\Omega} u_1\big)^2, \tag{5.1}$$

where cap *is the capacity defined in* [19, section 2.2.1], *namely*

$$\mathrm{cap}(D_2 \setminus D_1, G) := \inf\left\{ \int_G |\nabla v|^2 \;\middle|\; v = 1 \text{ in a nbhd of } D_2 \setminus D_1\,;\ v \in C_0^\infty(G) \right\}.$$

Proof. Let $M := \sup_{G\cap\Omega} u_1$. As explained in [9] near equation (3.2), it follows from de Giorgi's argument (see formula (5.12) in Chapter 2 of Ladyzhenskaya and Ural'tseva [17]) that $\sup_\Omega u_1$ is finite, and can even be chosen to depend only on Ω, not on D_1. To obtain a test function for the CHVP which determines $\lambda(D_2)$, we modify u_1 in G: In $\overline{G}\cap\Omega$, let $u_2 := \min\{u_1, M(1-v)\}$, where v is one of the functions that approximate the capacity of $D_2 \setminus D_1$; outside (if any), let $u_2 = u_1$. Since $u_2 = u_1$ on $\Omega\cap\partial G$, this does not introduce discontinuities, and u_2 is an admissible test function for $\lambda(D_2)$.

Clearly

$$\int_\Omega u_2^2(x) \ge \int_{\Omega\setminus G} u_1^2 \ge 1 - M^2\,\mathrm{vol}\,(G\cap\Omega)$$

and

$$\int_\Omega |\nabla u_2|^2 \le \int_\Omega |\nabla u_1|^2 + M^2 \int_G |\nabla v|^2 \to \int_\Omega |\nabla u_1|^2 + M^2\,\mathrm{cap}(D_2 \setminus D_1, G)$$

as v runs through a minimizing sequence for the capacity functional. We conclude (5.1) immediately. ∎

In applications of the lemma, G should be a small neighborhood of $D_2 \setminus D_1$, so that in the numerator on the right-hand side of (5.1), the capacity term dominates the volume term. It can be used to establish continuity of the eigenvalue under deformations of sufficiently regular windows. The following simplified estimate suffices to show the continuous dependence of the eigenvalue on the length and position of a segment in a square.

PROPOSITION 5.2. *For a given bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, there exists a nonnegative continuous function η with $\eta(0) = 0$ so that for any pair of windows $D_1 \subset D_2 \subset \partial\Omega$,*

$$\lambda(D_2) \;\leq\; \lambda(D_1) + \eta\left(\operatorname{diam}(D_2 \setminus D_1)\right),$$

where the η is a continuous function with $\eta(0) = 0$ which depends only on Ω but not on D_1 and D_2. The result applies to the coarse as well as to the fine definition of the eigenvalue.

Proof. We assume $D_2 \setminus D_1 \subset B_\delta(x_0)$ where $\delta := \operatorname{diam}(D_2 \setminus D_1)$ and use the Green's function as a legitimate limiting case for v in the capacity functional; namely, for dimension $d = 2$, let $G = B_R(x_0)$ and $1 - v := \ln_+(|x - x_0|/\delta)/\ln(R/\delta)$, with, say, $R = \sqrt{\delta}$ when $\delta < 1$. For $d \geq 3$, we can take $G = B_R(x_0)$ with $R := \delta^{(d-2)/d}$ and let $1 - v := (\delta^{-(d-2)} - |x - x_0|^{-(d-2)})_+/(\delta^{-(d-2)} - R^{-(d-2)})$. For simplicity, we can take $M := \sup_\Omega u_1$ as an upper bound for $\sup_{G\cap\Omega} u_1$ and obtain the claim with

$$\eta(\delta) := \begin{cases} \pi M^2 \dfrac{\delta\lambda_{\mathrm{Dir}} + 2/\ln(\delta^{-1/2})}{1 - \pi M^2 \delta} & \text{for } d = 2, \\ \dfrac{(d-1)^2 \omega_d M^2 \delta^{d-2}}{1 - \omega_d M^2 \delta^{d-2}} & \text{for } d > 2, \end{cases} \tag{5.2}$$

where ω_d is the volume of the unit ball. $\square$

It should be noted that the modulus of continuity of the eigenvalue cannot be expressed in terms of $\sigma(D_2 \setminus D_1)$ alone. This is due to the fact [8, Theorem 8] that for any ε there exists a window of measure $< \varepsilon$ with eigenvalue $> \lambda_{\mathrm{Dir}} - \varepsilon$. This observation also implies, in view of the a priori estimate for $\|u\|_\infty$ and Hölder's inequality, that an estimate in terms of $\|u_1\|_p$ is not possible for any $p < \infty$.

THEOREM 5.3. *The optimal eigenvalue λ_* depends continuously on the prescribed boundary measure of the window.*

Proof. We will prove that in dimensions $d > 2$, the function $\ell \mapsto \lambda_*(\ell)$ is Hölder continuous with exponent $(d-2)/(d-1)$, for $\ell < \sigma(\partial\Omega)$. In $d = 2$ dimensions, we will obtain a logarithmic estimate for the modulus of continuity.

Fix $\ell_1 < \sigma(\partial\Omega)$ and let D_1 be an optimal window with $\sigma(D_1) = \ell_1$. It follows from Proposition 5.2 that

$$\lambda\big(D_1 \cup (B_\delta(x_0) \cap \partial\Omega)\big) - \lambda(D_1) < \eta(\delta)$$

for any choice of $x_0 \in \partial\Omega$ and $\delta > 0$. We want to choose x_0 in such a way that $\sigma((B_\delta(x_0) \cap \partial\Omega) \setminus D_1)$ is bounded away from zero. To do this, we use Fubini's theorem to estimate

$$\begin{aligned} \fint_{\partial\Omega} \sigma((B_\delta(x) \cap \partial\Omega) \setminus D_1) d\sigma(x) &= \frac{1}{\sigma(\partial\Omega)} \int_{\partial\Omega} \int_{\partial\Omega\setminus D_1} \mathbf{1}_{|x-y|<\delta}\, d\sigma(y) d\sigma(x) \\ &= \frac{1}{\sigma(\partial\Omega)} \int_{\partial\Omega\setminus D_1} \sigma(B_\delta(y))\, d\sigma(y) \\ &\geq \frac{\sigma(\partial\Omega) - \sigma(D_1)}{\sigma(\partial\Omega)} \inf_{y\in\partial\Omega} \sigma(B_\delta(y)). \end{aligned}$$

Since Ω is a bounded Lipschitz domain, there exists a constant c, depending only on Ω, such that $\sigma(B_\delta(x_0)) \geq c\delta^{d-1}$. We conclude that for any value of δ there exists a point $x_0 \in \partial\Omega$ such that

$$\sigma\big(B_\delta(x_0) \setminus D_1\big) \geq \left(1 - \frac{\ell_1}{\sigma(\partial\Omega)}\right) c\delta^{d-1}.$$

For $\ell_2 > \ell_1$, set

$$\delta = \left(\frac{\ell_2 - \ell_1}{c(1 - \ell_1/\sigma(\partial\Omega))}\right)^{1/(d-1)},$$

and let $D_2 = D_1 \cup (B_\delta(x_0) \cap \partial\Omega)$. Since $\sigma(D_2) \geq \ell_2$, it follows that

$$\lambda_*(\ell_2) - \lambda_*(\ell_1) \leq \lambda(D_2) - \lambda(D_1) \leq \eta(\delta).$$

The claim now follows from the expression for η given in Proposition 5.2. ◻

The punchline of Theorem 5.3 is that we get a uniform modulus of continuity without extra regularity assumptions on the boundary. For smoother $\partial\Omega$, stronger results could be obtained using the tools of section 6. We conjecture (but have not pursued) that the window D_2 in Example 4.1 is actually optimal, and that the modulus of continuity at that length in Example 4.1 is precisely $O(\delta^{2/3})$. This intuition is based on the $r^{1/3}$ singularity of the eigenfunction at the re-entrant corner, the role of singularities revealed in section 6, and the estimate from Lemma 5.1. A Lipschitz estimate for $\ell \mapsto \lambda_*(\ell)$ should not be expected without further assumptions on $\partial\Omega$, but smoothness (a.e.) of $\partial\Omega$ will improve upon Theorem 5.3.

The following simple lemma estimates the change of the eigenvalue under increase of a window solely in terms of the eigenfunction.

Lemma 5.4. *Let $\Omega \subset \mathbb{R}^d$, and consider two windows $D_1 \subset D_2 \subset \partial\Omega$. Let u_1 be the normalized eigenfunction corresponding to $\lambda(D_1)$. Then*

$$\lambda(D_2) - \lambda(D_1) \leq \lambda(D_1)\frac{\sqrt{\operatorname{vol}(\Omega)}\,\sup_{D_2\setminus D_1} u_1}{1 - \sqrt{\operatorname{vol}(\Omega)}\,\sup_{D_2\setminus D_1} u_1}. \tag{5.3}$$

Proof. Let $\varepsilon := \sup_{D_2\setminus D_1} u_1$. Then $v_\varepsilon = (u_1 - \varepsilon)_+$ is an admissible test function for both the CHVPs defining $\lambda(D_2)$ and $\lambda(D_1)$. We compute

$$\|\nabla v_\varepsilon\|_2^2 = \int \nabla u_1 \cdot \nabla(u_1 - \varepsilon)_+ = \lambda(D_1)\int u_1(u_1 - \varepsilon)_+, \tag{5.4}$$

where we have used the weak form $\int \nabla u \nabla\varphi = \lambda \int u\varphi$ of the eigenvalue equation $\Delta u = -\lambda u$, with $\varphi := v_\varepsilon$. It follows that

$$\begin{aligned}
\lambda(D_2) - \lambda(D_1) &\leq \frac{\int |\nabla v_\varepsilon|^2}{\int v_\varepsilon^2} - \lambda(D_1)\\
&\leq \lambda(D_1)\frac{\int \varepsilon(u_1 - \varepsilon)_+}{\int (u_1 - \varepsilon)_+^2}\\
&\leq \lambda(D_1)\frac{\varepsilon}{\|(u_1 - \varepsilon)_+\|_2}\left(\operatorname{vol}(\Omega)\right)^{1/2}.
\end{aligned}$$

The triangle inequality $\|(u_1 - \varepsilon)_+\|_2 \ \geq\ 1 - \varepsilon\left(\operatorname{vol}(\Omega)\right)^{1/2}$ now yields the claim. ◻

For a given window $D \subset \partial\Omega$, denote by

$$D_\delta := \Big(\bigcup_{x\in D} B_\delta(x)\Big) \cap \partial\Omega \quad (\delta > 0), \quad D_0 := D \tag{5.5}$$

the relative δ-neighborhood of D in $\partial\Omega$. Continuity of the eigenfunction up to the boundary is sufficient for continuity of the eigenvalue function $\delta \mapsto \lambda(D_\delta)$.

THEOREM 5.5. *Let u be an eigenfunction for window boundary conditions on D, and assume that the preferred representative $\tilde u$ vanishes everywhere on D.*

(a) *If $\tilde u$ is upper semicontinuous on $\overline{\Omega}$, then $\lambda(\cdot)$ is outer regular at D in the sense that for every $\varepsilon > 0$ there exists a relatively open subset $U \subset \partial\Omega$ containing D, with the property that*

$$\lambda(U) \le \lambda(D) + \varepsilon.$$

(b) *If u is continuous up to the boundary of Ω, then the map $\delta \mapsto \lambda(D_\delta)$ is right continuous at $\delta = 0$.*

We will show below (Theorem 5.7) that the hypothesis of part (a) is satisfied for $C^{1,\alpha}$ domains in $\mathbb{R}^2$, and at flat pieces of the boundary in any dimension. We conjecture that upper semicontinuity may hold at least for smooth domains in any dimension.

Concerning part (b), continuity up to the boundary can be shown for the eigenfunction by a careful analysis of de Giorgi's argument, under the assumption that the window D has positive Lebesgue density at every interface point $x_0 \in \overline{D} \cap \overline{\partial\Omega \setminus D}$. We conjecture, but cannot prove, that eigenfunctions for optimal windows are continuous up to the boundary. Below, we show by an example that continuity of the eigenfunction is not necessary for continuity of $\delta \mapsto \lambda(D_\delta)$.

Proof of Theorem 5.5. If the preferred representative $\tilde u$ is upper semicontinuous on the closure of Ω, then the set

$$U = \{x \in \partial\Omega \ \mid \ u(x) < \eta\}$$

is a (relatively) open set containing D. By Lemma 5.4, we have that

$$\lambda(D) \le \lambda(U) \le \frac{\lambda(D)}{1 - \eta\mu(\Omega)^{1/2}} < \lambda(D) + \varepsilon$$

if $\eta = \eta(\varepsilon)$ is chosen sufficiently small (e.g., $\eta := \varepsilon(\lambda_{\mathrm{Dir}}(\Omega)\,\mathrm{vol}\,(\Omega))^{-1}$). This proves outer regularity. If $\tilde u$ is continuous, then D is compact, and hence there exists a $\delta > 0$ so that $D_\delta \subset U$, which proves the second claim. $\square$

Note that assuming that $\tilde u$ vanishes everywhere on D amounts to replacing D with its refinement and selecting the fine eigenvalue. Since coarse and fine eigenvalues agree for the open windows D_δ, continuity of λ^c certainly fails at any window D for which $\lambda^c(D) < \lambda^f(D)$. Cantor sets of zero measure but positive capacity provide examples of such windows.

However, $\delta \mapsto \lambda^f(D_\delta)$ cannot be continuous in general either. For an open-dense window D of small measure, we clearly have $\lambda(D_\delta) = \lambda_{\mathrm{Dir}}$ for all $\delta > 0$. However, we claim that $\lambda_D < \lambda_{\mathrm{Dir}}$. To see this, note that u_D cannot agree with u_{Dir}, since eigenfunctions do not take on"extra" Dirichlet boundary conditions, as was shown near Figure 1 in [9]. Since u_{Dir} is an admissible candidate for the CHVP for λ_D, it follows from the uniqueness of the minimizer that $\lambda_{\mathrm{Dir}} > \lambda_D$. This also provides an example of a window whose eigenfunction is discontinuous on a set of large measure on the boundary.

Example 5.6. There exists an open window D with discontinuous eigenfunction, such that still $\delta \mapsto \lambda(D_\delta)$ is right continuous.

Proof. In a planar domain, parametrize a portion of the boundary by arclength and refer to segments on the boundary as intervals in this parameter. We will construct two decreasing sequences $x_n \searrow 0$ and $\delta_n \searrow 0$ and let $I_n :=]x_n - \delta_n, x_n + \delta_n[$. The sequences x_n and δ_n will be specified later. The window will be $D := \bigcup_{n=1}^\infty I_n$, and we will also define $D_N := \bigcup_{n=1}^N I_n$, with the eigenvalues and normalized eigenfunctions λ, λ_N, u, u_N, respectively. If N is the first index such that $x_N < \delta$, then

$$D_\delta \setminus D \subset]-\delta, \delta[\cup \bigcup_{n=1}^{N} [x_n + \delta_n, x_n + \delta_n + \delta[\cup \bigcup_{n=1}^{N-1}]x_n - \delta_n - \delta, x_n - \delta_n].$$

It follows from Proposition 5.2 that $\lambda(D_\delta) - \lambda(D) < (2N+1)\eta(\delta) < (2N+1)\eta(x_{N-1})$. Choosing the sequence (x_n) such that $(2N+1)\eta(x_{N-1}) \to 0$ as $N \to \infty$ ensures the right continuity of $\delta \mapsto \lambda(D_\delta)$.

With (x_n) thus fixed, we introduce the compact set $K := \{0\} \cup \{y_n \mid n \in \mathbb{N}\}$, where $y_n = (x_n + x_{n+1})/2$ and construct the sequence (δ_n) inductively. Let $\delta_1 = (x_1 - y_1)/2$.

Since D_1 has positive Lebesgue density at all interface points, it follows from de Giorgi's argument that the corresponding eigenfunction u_1 is Hölder continuous up to the boundary. Let $a := \inf_K u_1 > 0$ and define $a_n := (1/2 + 1/2^n)a$. We will choose δ_N in such a way that $\inf_K u_N \geq a_N$. Assume $\delta_1, \ldots, \delta_{N-1}$ have been constructed. The interval I_N, and thus D_N and u_N, will depend on the choice of δ_N. But as $\delta_N \to 0$, the local de Giorgi estimates near K remain uniform, because the L^∞ estimate for u_N does not depend on the window and the interface stays away from K. Then $u_N^{(\delta_N)}$ converges weakly in $W^{1,2}(\Omega)$, strongly in $L^2(\Omega)$, and strongly in $L^2(\partial\Omega)$ by the usual compactness arguments. It also converges strongly in $W^{1,2}(\Omega)$ to u_{N-1} since $\lambda_N^{(\delta_N)} \to \lambda_{N-1}$; the convergence is uniform in a neighborhood of K by the equicontinuity obtained from de Giorgi. Since $u_{N-1} \geq a_{N-1}$ on the compact set K, we can achieve $u_N \geq a_{N-1} - \varepsilon$ for any $\varepsilon > 0$ by making δ_N small; in particular we can achieve $u_N \geq a_N$.

It is now easy to show that u is discontinuous at 0. Indeed, as $N \to \infty$, $u_N \to u$ in the Sobolev spaces mentioned above. Again, the convergence is uniform in a neighborhood of each single y_n. Therefore, $u(y_n) \geq a/2$ for each n, whereas $u(x_n) = 0$. Hence u is discontinuous at 0. □

We finally refer to Lemma 6.1, which gives continuity estimates under distortion of a window by means of a bi-Lipschitz homeomorphism. Due to the similarity of proofs, we conveyed it to section 6.

5.2. On upper semicontinuity of eigenfunctions. Here, we will prove semicontinuity of eigenfunctions as a consequence of a subharmonicity argument.

THEOREM 5.7. *If $\Omega \subset \mathbb{R}^2$ has a $C^{1,\alpha}$ boundary, then for any measurable window $D \subset \partial\Omega$, the eigenfunction u has an upper semicontinuous preferred representative $\tilde{u}$. If $\Omega \subset \mathbb{R}^d$ with $d > 2$, then $\tilde{u}$ is upper semicontinuous at any boundary point where the boundary is locally part of a hyperplane.*

Proof. Let u be the solution of the CHVP (1.5) for D, the eigenvalue being $\lambda(D)$. Fix $x_0 \in \partial\Omega$. We will show that if the $\partial\Omega$ coincides with a hyperplane in some neighborhood of x_0, then the limit

$$\tilde{u}(x) := \lim_{r \to 0} \fint_{B_r(x) \cap \Omega} u(y)\, dy \tag{5.6}$$

exists for all points in this neighborhood and defines an upper semicontinuous function. This limit agrees with the preferred representative defined in (1.7). In the special case of two dimensions, the conclusion holds assuming only that $\partial\Omega$ is of regularity $C^{1,\alpha}$ near x_0. We note that in the interior of Ω, u is always smooth.

The basic idea is as follows: When the boundary is locally part of a hyperplane, extend u by *even* reflection, regardless of the type of boundary conditions. The nonnegative function u, thus extended, has only such discontinuities as are possible for a subharmonic distribution, and this fact is shown by means of the test function $(u - t\varphi)_+$ in the CHVP, where φ is smooth nonnegative. Subharmonicity implies upper semicontinuity according to Theorem 9.3 in [18]. For curved boundary in 2D, the Riemann mapping theorem locally provides an analogue of the reflection.

Consider first the case where there exists a neighborhood V of x_0 such that $\partial\Omega \cap V$ is contained in a hyperplane. We may assume that the hyperplane is given by $x_d = 0$, that Ω lies above the hyperplane, and that V is symmetric under the reflection $(x', x_d) \mapsto (x', -x_d)$. Let φ be a smooth nonnegative function with support in V. Since $(u - t\varphi)_+$ is a legitimate candidate for the CHVP when $t \geq 0$, we have

$$\frac{A(t)}{B(t)} \geq \frac{A(0)}{B(0)}, \tag{5.7}$$

where

$$A(t) := \int_\Omega |\nabla(u - t\varphi)_+|^2, \quad B(t) := \int_\Omega |(u - t\varphi)_+|^2.$$

We calculate from the weak Euler equations

$$\begin{aligned} A(t) &= \int_\Omega \nabla(u - t\varphi)_+ \nabla u - t \int_\Omega \nabla(u - t\varphi)_+ \nabla\varphi \\ &= \lambda \int_\Omega (u - t\varphi)_+ u - t \int_{u > t\varphi} \nabla u \nabla\varphi + t^2 \int_{u > t\varphi} |\nabla\varphi|^2, \end{aligned} \tag{5.8}$$

and expand

$$B(t) = \int_\Omega u\,(u - t\varphi)_+ - t \int_\Omega \varphi\,(u - t\varphi)_+. \tag{5.9}$$

Inserting (5.8) and (5.9) into (5.7) and using that $A(0) = \lambda$ and $B(0) = 1$, we obtain the following for $t > 0$:

$$\begin{aligned} 0 &\leq t^{-1}(A(t)B(0) - A(0)B(t)) \\ &= \int_{u > t\varphi} \Big[-\nabla u \nabla\varphi + \lambda u\varphi \Big] + t \int_{u > t\varphi} \Big[|\nabla\varphi|^2 - \lambda\varphi^2 \Big]. \end{aligned}$$

Since all integrals over sets $u > t\varphi$ converge to integrals over Ω by Lebesgue's dominated convergence theorem, we obtain for $t \to 0+$ that

$$0 \leq \int_{V \cap \Omega} \Big[-\nabla u \nabla\varphi + \lambda u\varphi \Big]. \tag{5.10}$$

We now extend u by even reflection $u(x', -x_d) := u(x', x_d)$ and use (5.10) for the likewise reflected test function φ. Adding the reflected and the original (5.10), we obtain

$$0 \leq \int_V \Big[-\nabla u \nabla\varphi + \lambda u\varphi \Big] = \int_V \Big[u\Delta\varphi + \lambda u\varphi \Big], \tag{5.11}$$

where we have used that φ is C^2 and supported in V.

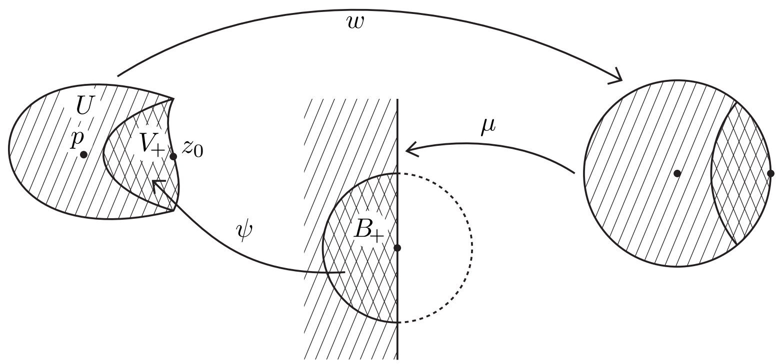


FIG. 5.1. *The Riemann mappings used in the proof of Theorem* 5.7.

We have shown that $\Delta u + \lambda u$ is nonnegative in the sense of distributions. If $v := u + \frac{M\lambda}{2d}|x|^2$, where $M := \|u\|_\infty < \infty$, then $\Delta v \geq 0$ in the sense of distributions. By [18, Theorem 9.3], v is subharmonic; that is,

$$v(x) \;\leq\; \fint_{B_r} v \tag{5.12}$$

for almost every $x \in V$, provided $B_r(x) \subset V$. Furthermore, the preferred representative $\tilde{v}$ of v is upper semicontinuous and satisfies the subharmonicity condition (5.12) for all x and r so that $B_r(x) \subset V$. Since $\tilde{u}$ differs from $\tilde{v}$ by a continuous function, it is upper semicontinuous as well. This settles the case where $\partial\Omega \cap V$ is contained in a hyperplane, in some neighborhood of x_0.

In the case where $\Omega \subset \mathbb{R}^2$ we use complex notation. Let V be a neighborhood of $z_0 \in \partial\Omega$ such that $\partial\Omega \cap V$ is of class $C^{1,\alpha}$, and let V_+ be the intersection of V with Ω. Replacing V by a subset, we may assume that there exists a conformal map ψ from a semidisc B_+ to V_+ such that the diameter of the semidisc maps onto $V \cap \partial\Omega$. The function $\bar{u} = u \circ \psi$ on the semidisc satisfies $\Delta\bar{u} = |\psi'|^2\,(\Delta u) \circ \psi$. Our argument will rely on the boundedness of $|\psi'|$ (shown below). By reflection, we can extend $\bar{u}$ into the full disc B. The extended function $\bar{u}$ is still in $W^{1,2}(B_+)$ since $\psi' \in L^\infty$; as before, the extended function remains in $W^{1,2}(B)$. From (5.10), we conclude, using the conformal invariance of the Dirichlet integral, that

$$0 \leq \int_{B_+} \left[-\nabla(u \circ \psi)\nabla(\varphi \circ \psi) + \lambda|\psi'|^2(u \circ \psi)(\varphi \circ \psi)\right]$$

for all $0 \leq \varphi \in W^{1,2}(V_+)$ that vanish on $\Omega \cap \partial V_+$, particularly for all $\varphi := \bar{\varphi} \circ \psi^{-1}$ with $\bar{\varphi} \in C_0^2(B)$. As with (5.11), we can now conclude that $v := \bar{u} + M\lambda 2d \sup|\psi'|^2$ is subharmonic and finish the argument as before.

We still need to explain why $|\psi'|$ remains bounded near $\partial\Omega$; this is where the $C^{1,\alpha}$ regularity of the boundary enters. Refer to Figure 5.1. Choose U to be the intersection of a neighborhood of $z_0 \in \partial\Omega$ with Ω such that U is simply connected. Choose a point $p \in U$. The Green's function of U can be obtained in the form $\ln|z-p| + \xi(z)$ with ξ harmonic subject to boundary values $-\ln|z-p|$. Near z_0, this harmonic function ξ is $C^{1,\alpha}$ up to the boundary, because the boundary has this regularity there. This result follows from the Schauder estimates given in [11], namely their Theorem 5.1 in connection with Lemma 2.1. If η is a conjugate harmonic to ξ

(namely $\eta_y = \xi_x$, $\eta_x = -\xi_y$), then $w : z \mapsto (z-p)\exp[\xi(z)+i\eta(z)]$ is a conformal map of U onto a disc. (For more details, see [5, section I.7].) The mapping w inherits the $C^{1,\alpha}$ regularity from ξ. With a conformal mapping μ from the disc onto a half plane, we select an appropriate semidisc B_+ from this half plane and let $\psi := (\mu \circ w)^{-1}|_{B_+}$ with $V_+ := \psi(B_+) \subset U$. ∎

It is worth noting that a C^1 boundary is not sufficient for the bounded derivatives of a Riemann map, as can be seen from the map $w(z) = z \ln z$ and its inverse, which map neighborhoods of 0 in the half planes $\operatorname{Re} z > 0$ or $\operatorname{Re} w > 0$, respectively, onto domains bounded by C^1 curves.

6. First variation, and the role of singular coefficients in optimality. In this section, we study how the principal eigenvalue of the Laplacian with window boundary conditions changes under deformations of the window. The first lemma contains some estimates for distortions by bi-Lipschitz maps.

LEMMA 6.1. *Let $\psi : \overline{\Omega}_1 \to \overline{\Omega}_2$ be a bi-Lipschitz map. Then for any window D in $\overline{\Omega}_2$, it holds that*

$$\lambda(\psi^{-1}(D)) \leq \lambda(D) \sup_{\Omega_1} \rho\Big((D\psi)(D\psi)^T(\det D\psi)^{-1}\Big) \sup_{\Omega_1}(\det D\psi),$$

where ρ denotes the spectral radius. In terms of the distortion ratios

$$a(x) := \limsup_{y\to x} \frac{|\psi(y)-\psi(x)|}{|y-x|}, \quad b(x) := 1/\liminf_{y\to x} \frac{|\psi(y)-\psi(x)|}{|y-x|},$$

we have the simpler (but weaker) estimates

$$\frac{\lambda(\psi^{-1}(D))}{\lambda(D)} \leq \sup(ab^{d-1})\sup(a^{d-1}/b) \leq (\sup a)^{d+1}(\sup b)^{d-1}.$$

Proof. For any two differentiable functions h_1, h_2 on Ω and any diffeomorphism ψ, we have the transformation formulas

$$\int_\Omega h_1(y)\, h_2(y)\, dy = \int_{\psi^{-1}(\Omega)} (h_1\circ\psi)(x)\, (h_2\circ\psi)(x)\, \det D\psi(x)\, dx \tag{6.1}$$

and

$$\int_\Omega \nabla h_1(y)\cdot\nabla h_2(y)\, dy = \int_{\psi^{-1}(\Omega)} \nabla_x(h_1\circ\psi)(x)^T M(x)\nabla_x(h_2\circ\psi)(x)\, dx, \tag{6.2}$$

where the matrix M is given by

$$M(x) = D\psi(x)^{-1}\, D\psi(x)^{-T}\, \det D\psi(x). \tag{6.3}$$

Let u be the nonnegative normalized eigenfunction for window $D \subset \partial\Omega$, and take $u\circ\psi$ as a test function in the CHVP for $\psi^{-1}(D)$. The first claim follows from (6.1)–(6.3) by setting $h_1 = h_2 = u$ and using that the smallest eigenvalue of $M(x)$ is the reciprocal of the spectral radius of $M(x)^{-1}$. The distortion ratio estimates follow for $\psi \in C^1$ from $\rho(D\psi(x)^T D\psi(x)) \leq a(x)^2$ and $a(x)^2/b(x)^{2(d-1)} \leq \det(D\psi(x)^T D\psi(x)) \leq a(x)^{2(d-1)}/b(x)^2$, as calculated in an eigenbasis of this symmetric matrix. Both estimates extend to bi-Lipschitz maps by approximation. ∎

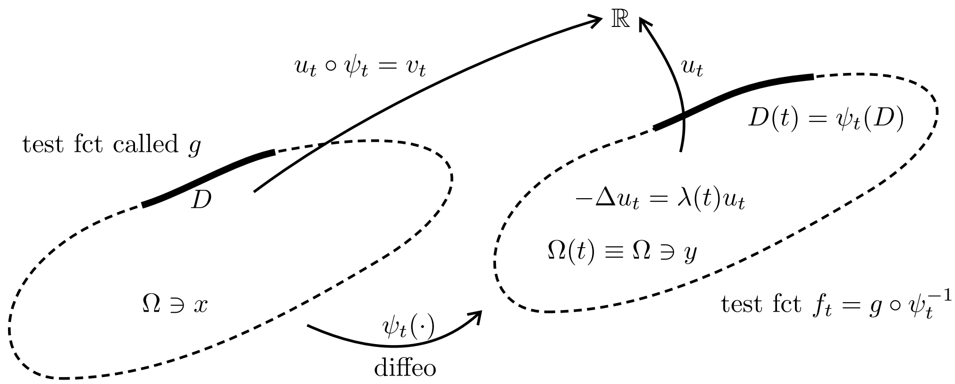


FIG. 6.1. *The mappings in the proof of Theorem* 6.2.

Our main result in this section describes the change of the principal eigenvalue under a diffeomorphism generated by a flow.

THEOREM 6.2. *Let Ω be a Lipschitz domain in $\mathbb{R}^d$, D a window, u its normalized eigenfunction, and X a vector field of regularity $C^1(\overline{\Omega})$ that is "parallel" to the boundary in the sense that Ω is the union of an increasing sequence of smoothly bounded subdomains Ω_δ, with $\delta \searrow 0$, such that X is tangential on $\partial\Omega_\delta$ for δ sufficiently small.*

Let ψ_t be the flow of X. Consider the dependence of the first eigenvalue λ as D changes under the flow. Then it holds that

$$\frac{d}{dt}\lambda(\psi_t(D))\Big|_{t=0} = -2 \lim_{\delta\to 0} \int_{\partial\Omega_\delta} \partial_\nu u \, L_X u, \tag{6.4}$$

where $L_X u$ denotes the directional derivative of u in direction X.

Remark. The assumptions guarantee that X is tangential to the boundary of Ω at smooth boundary points and X may need to vanish in boundary points where the boundary is not C^1.

Proof. Let $\psi_t : x \mapsto \psi_t(x) = y$, and $\Omega \to \Omega$ be the bi-Lipschitz homeomorphism arising from the vector field X, i.e., $\frac{d}{dt}\psi_t(x)|_{t=0} = X(\psi_t(x))$, $\psi_0(x) = x$. Refer to Figure 6.1. Since $X \in C^1$, ψ is a C^1-diffeomorphism in the interior of Ω and satisfies a Lipschitz estimate up to the boundary.

Let $u_t(\cdot)$ and $\lambda(t)$ be the eigenfunctions and eigenvalue for $D(t) := \psi_t(D)$, and let g be a test function on Ω whose trace vanishes on D. The variation of geometry will be expressed as a variation of the operator by referring all windows back to the coordinates x.

We will denote the pullback of the eigenfunction u_t to Ω with window boundary conditions on D as $u_t \circ \psi_t =: v_t$. Similarly $f_t := g \circ \psi_t^{-1}$, the pushforward of the test function g. The weak eigenvalue equation for $u_t(\cdot)$ is

$$\int_{\Omega(t)} \nabla_y u_t(y) \cdot \nabla_y f_t(y)\, dy = \lambda(t) \int_{\Omega(t)} u_t(y) f_t(y)\, dy,$$

where, in our case, $\Omega(t) \equiv \Omega$, g vanishes on D, and f_t vanishes on $D(t)$.

We now use (6.1)–(6.3) with $\psi = \psi_t$, $h_1 = u_t$, $h_2 = f_t$ and expand to first order in t. From $\frac{d}{dt}\psi_t(x) = X(\psi_t(x))$, $\psi_0(x) = x$, we obtain

$$\begin{aligned}
\psi_t(x) &= x + tX(x) + o(t) \\
D\psi_t(x)^j{}_i &= \delta_i^j + t\frac{\partial X^j}{\partial x^i} + o(t) \\
\left(D\psi_t(x)^{-1}\right)^j{}_i &= \delta_i^j - t\frac{\partial X^j}{\partial x^i} + o(t) \\
\det D\psi_t(x) &= 1 + t \operatorname{div} X + o(t).
\end{aligned}$$

The estimates for the remainder terms are uniform in $x \in \Omega$. Inserting the first and last estimate into (6.1) with $\psi = \psi_t$, $h_1 = u_t$, $h_2 = f_t$ yields

$$u_t f_t \, dy = \left(v_t g(1 + t \operatorname{div} X) + o(t)\right) dx,$$

where the $o(t)$ term represents an L^1 function. Similarly, we obtain from (6.2)

$$\begin{aligned}
\nabla_y u_t(y) \cdot \nabla_y f_t(y)\, dy = \Bigg\{ &\nabla_x v_t(x) \cdot \nabla_x g(x) \\
&+ t\left((\operatorname{div} X)\nabla_x v_t \cdot \nabla_x g - \left(\frac{\partial g}{\partial x^i}\frac{\partial v_t}{\partial x^j} + \frac{\partial v_t}{\partial x^i}\frac{\partial g}{\partial x^j}\right)\frac{\partial X^j}{\partial x^i}\right) + o(t)\Bigg\}\, dx,
\end{aligned}$$

where the $o(t)$ term again represents an L^1 function. We have used the Einstein summation convention to express the sum over i and j.

If we truncate the bilinear forms by dropping the $o(t)$ terms, it is immediate that the eigenvalue will only change by $o(t)$. Since the truncated operators depend analytically on the perturbation parameter t, we may use results from Chapter VII of Kato [14] to estimate the eigenvalue up to errors of order $o(t)$. Kato's Theorem VII.4.2 and his discussion in Chapter VII §6, sections 2–5 ascertain, via spectral projections and for any finite set of isolated eigenvalues, that the perturbation theory works as in finite-dimensional spaces. In particular, a simple eigenvalue and its corresponding eigenfunction of the truncated operators depend analytically on t. We may therefore write down expansions $v_t = v_0 + tv_1 + O(t^2)$ of the eigenfunction for the truncated problem, and $\lambda(t) = \lambda_0 + t\lambda_1 + o(t)$ of the eigenvalue (for the truncated as well as for the full problem), and compare like powers of t.

Order t^0 yields

$$\int_\Omega \nabla v_0 \cdot \nabla g \, dx = \lambda_0 \int_\Omega v_0 g \, dx,$$

which is just the weak Euler equation for v_0. Order t^1 yields

$$\begin{aligned}
\lambda_1 \int & v_0 g \, dx + \lambda_0 \int \{v_1 g + (\operatorname{div} X) v_0 g\}\, dx \\
&= \int \left\{ \nabla v_1 \cdot \nabla g + (\operatorname{div} X)\nabla v_0 \cdot \nabla g - \frac{\partial X^j}{\partial x^i}\left(\frac{\partial g}{\partial x^i}\frac{\partial v_0}{\partial x^j} + \frac{\partial v_0}{\partial x^i}\frac{\partial g}{\partial x^j}\right)\right\} dx.
\end{aligned}$$

These equations are valid for integration over any subdomain of Ω. We will integrate over Ω_δ, where Ω_δ runs through an increasing sequence of smoothly bounded domains compactly contained in Ω such that X is tangent to the boundary of Ω_δ. We write

$$\int\!\!\!\!\!\!\bullet := \int_{\Omega_\delta} \quad \text{and} \quad \oint := \int_{\partial\Omega_\delta}$$

for volume and surface integrals, respectively. Using $g = v_0$ as a test function, we obtain in first order

$$\lambda_1 \oint v_0^2 = \oint (\nabla v_1 \cdot \nabla v_0 - \lambda_0 v_0 v_1) + \oint \left\{ (\operatorname{div} X)\big(|\nabla v_0|^2 - \lambda_0 v_0^2\big) - 2\frac{\partial X^j}{\partial x^i}\frac{\partial v_0}{\partial x^i}\frac{\partial v_0}{\partial x^j} \right\}.$$

Since v_1 lies in $W^{1,2}(\Omega)$ and satisfies window boundary conditions for D, it is a valid test function in the Euler–Lagrange equation for v_0, and we conclude that the first integral vanishes as $\delta \to 0$. For the second integral, we use the identity

$$\begin{aligned}\frac{\partial X^j}{\partial x^i}\frac{\partial v_0}{\partial x^i}\frac{\partial v_0}{\partial x^j} &= \frac{\partial}{\partial x^i}\left(\frac{\partial v_0}{\partial x^i} L_X v_0\right) - X^j \frac{\partial}{\partial x^i}\left(\frac{\partial v_0}{\partial x^i}\frac{\partial v_0}{\partial x^j}\right) \\ &= \operatorname{div}(L_X v_0\, \nabla v_0) + \frac{1}{2} L_X(\lambda_0 v_0^2 - |\nabla v_0|^2)\end{aligned}$$

and Gauss' divergence theorem to compute

$$\begin{aligned}\oint &\left\{ (\operatorname{div} X)\big(|\nabla v_0|^2 - \lambda_0 v_0^2\big) - 2\frac{\partial X^j}{\partial x^i}\frac{\partial v_0}{\partial x^i}\frac{\partial v_0}{\partial x^j} \right\} \\ &\qquad = \oint \operatorname{div}\Big(\big(|\nabla v_0|^2 - \lambda_0 v_0^2\big) X\Big) - 2\operatorname{div}(\nabla v_0\, L_X v_0) \\ &\qquad = \oint (|\nabla v_0|^2 - \lambda_0 v_0^2)\, X \cdot \nu - 2\partial_\nu v_0\, L_X v_0.\end{aligned}$$

The first term under the integral vanishes since X is tangential to the boundary of Ω_δ by assumption, and the claim follows as $\delta \to 0$. ◻

We note that, at least formally, the integrand on the right-hand side of (6.4) vanishes on both the Dirichlet and the Neumann parts of the boundary of Ω. The evaluation of the limit of the integral as $\delta \to 0$ is far from trivial in higher dimensions, but reasonably straightforward in two dimensions with nice window geometry. It amounts to the evaluation of certain singular coefficients at interface points between the Neumann and Dirichlet parts of $\partial\Omega$. It has been shown that in polygonal domains, in the neighborhood of a corner, solutions of elliptic boundary problems lie locally in the direct sum of $W^{2,2}$ with a singular space, and in two dimensions, this singular space is one-dimensional. See, eg., Grisvard [12], in particular his Theorem 2.4.3. Indeed, functions in the singular space behave like the explicit harmonic functions $\operatorname{Re}(cz^\alpha)$ with α appropriate for the boundary conditions. In this context, it is understood that an interface point between Dirichlet and Neumann data is a corner even if (in particular if!) the geometric boundary is smooth there. As noted, corners that can be made to disappear by means of the reflection principle (like the geometric corners of a rectangle) do not have a singular space. The singular coefficients (also called stress intensity coefficients) must be calculated (numerically) in practical situations. They depend on global information. For a wider background concerning singular contributions, see [7], [12], [16], [21], [23], and much other work by these authors and references given there.

In particular, the variational equation gives rise to the following corollary.

COROLLARY 6.3. *Consider a segment on the boundary of a rectangle, such that one endpoint of the segment is a corner of the rectangle, whereas the other endpoint is a point that is not a corner. Such a segment is not an optimal window, but can be improved infinitesimally by shifting in the direction that brings the corner point inside the window*

Proof. In self-explanatory notation, we refer to the windows as intervals; let $[a,b]$ be an interval with corner point b and noncorner point a. We will show that (with some positive constants m, M)

$$\lambda([a+\varepsilon,b+\varepsilon]) \le \lambda([a+\varepsilon,b]) + M\varepsilon^2 \qquad \text{and} \qquad \lambda([a+\varepsilon,b]) \le \lambda([a,b]) - m\varepsilon\,.$$

From this the claim is immediate.

The first estimate (local near b) follows from Lemma 5.1, with G a ball of radius 2ε centered at the corner b. The eigenfunction is smooth near b, because reflection in the Neumann boundary removes the singularity: $|u| = O(\varepsilon)$ in G, and the estimate is uniform with respect to small changes at the other end a. The capacity term is bounded as $\varepsilon \to 0$, based on a radial test function $\ln_+(|x-b_1|/\varepsilon)/\ln 2$ as in the proof of Proposition 5.2.

The second estimate (local near a) follows from an evaluation of the singular boundary integral $\int_{\partial\Omega} L_X u \partial_\nu u$. In the particular case of an interface point on a straight line, the local behavior of a solution u is $u = c\sqrt{r}\sin(\varphi/2) + v$ with $v \in W^{2,2}$.

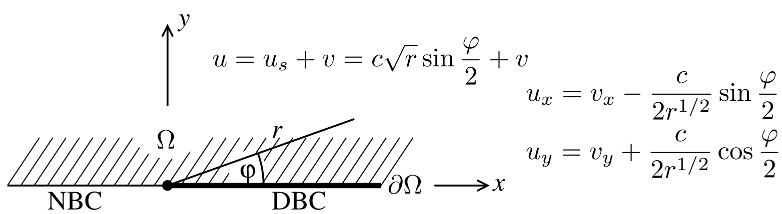


FIG. 6.2. *Coordinates near an interface point.*

To evaluate the singular boundary integral in terms of the singular coefficient, define coordinates as in Figure 6.2, with the boundary point a located at $(0,0)$. Let us assume that the C^1 vector field X is given by $f(x,y)\partial_x$ with the coefficient at the interface $f(0,0) = 1$. It can easily be seen that the regular function v does not contribute to the integral, nor do the mixed terms. We have

$$-2\int_{-t}^{t} L_X u\,\partial_\nu u\,dx = 2\int_{-t}^{t} \frac{\partial u_s}{\partial x}\,\frac{\partial u_s}{\partial y}\,dx = -\frac{c^2}{4}\int_{-t}^{t} \frac{y}{x^2+y^2}\,dx = -\frac{c^2}{2}\arctan\frac{t}{y},$$

and this converges to $-\frac{c^2\pi}{4}$ as $y \to 0+$.

Finally, we estimate the singular coefficient. Choose r so small that $B_r(0)$ intersects $\partial\Omega$ in a straight line as in Figure 6.2, with one radius (N_r) being Neumann boundary and one radius (D_r) Dirichlet boundary; let $S_r := (\partial B_r(0)) \cap \Omega$, and count φ from the Dirichlet to the Neumann boundary. Let

$$-\Delta h = 0 \text{ in } B_r(0)\cap\Omega, \quad \partial_\nu h = 0 \text{ on } N_r,\ h = 0 \text{ on } D_r,\ h = u \text{ on } S_r,$$
$$-\Delta v = \lambda u \text{ in } B_r(0)\cap\Omega, \quad \partial_\nu v = 0 \text{ on } N_r,\ v = 0 \text{ on } D_r,\ v = 0 \text{ on } S_r.$$

Then $u = v + h$ with $v \ge 0$. Evaluation on the boundary implies that the singular coefficient of u is at least as large as the singular coefficient of h. Explicit calculation of the singular coefficient of h by means of Fourier analysis gives exactly

$$c \ge \frac{2}{\pi r^{1/2}}\int_0^\pi u(re^{i\varphi})\,\sin\frac{\varphi}{2}\,d\varphi > 0. \qquad \square$$

The above estimate of the singular coefficient is closely related to formula (2.3) in Dauge, Lubuma, and Nicaise [7], which actually gives the exact coefficient (in terms of u). However, their formula is not designed to show nonvanishing (which relies on using the maximum principle), but is instead built on Fredholm properties. (The distinction that their formula is for a Dirichlet–Dirichlet corner, not a Dirichlet–Neumann corner, is a minor issue.)

Our argument shows that shortening a window infinitesimally at the interface decreases the eigenvalue by an amount proportional to the square of the singular coefficient at the end of the window. Moving a window amounts to shortening it at one end and lengthening it at the other end. To decrease the eigenvalue, the window should be moved in the direction of the smaller singular coefficient (i.e., towards the corner of the square if it is already close to a corner). If the window consists of several intervals, nonlocal changes that lengthen one component at the expense of the other can also be studied in terms of the singular coefficients. Conversely, singular coefficients can be determined graphically from the slopes in Figure 2.1, for the geometric configurations depicted there.

As an immediate consequence of the role of singular coefficients, a window consisting of any number of equidistant and congruent arcs on the boundary of a circle is a critical point for the first eigenvalue. Since these arcs can now be moved independently, these are critical points of arbitrarily large index. The optimal window in a circle is known to be a single arc [9].

Limitations of our result should also be observed. The variations induced by the flow of vector fields correspond to the "weak," C^1-small variations (as opposed to "strong," C^0-small variations) that are exploited in the Euler–Lagrange equations of the classical calculus of variations. It is doubtful how significant a role such variations can play if it comes to show, say, that a certain open-dense set of small measure is *not* an optimal window.

We have not established an analogue of the fundamental lemma of the calculus of variations that would permit elimination of the vector field X. In the absence of a priori regularity for optimal windows, such an attempt seems extremely difficult. There is, however, some hope to get nontrivial boundary regularity for the optimal eigenfunction by selecting vector fields constructed from the eigenfunction in some appropriate way. We plan a further investigation of this issue.

In spite of these limitations, Theorem 6.2 does give some insight into the question of optimal windows, and in particular into the variation of windows with a given a priori regularity.

Acknowledgments. The second author would like to thank Steve Cox for a quick introduction to the MATLAB pdetool that boosted the experimental work, V.G. Maz'ya for an inspiring chat in Darmstadt, and U. Mertins for sharing some of his numerics on the problem, each several years ago.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1978.
[2] F. Brock and A. Yu. Solynin, *An approach to symmetrization via polarization*, Trans. Amer. Math. Soc., 352 (2000), pp. 1759–1796.
[3] J. Brothers and W. P. Ziemer, *Minimal rearrangements of Sobolev functions*, J. Reine Angew. Math., 384 (1988), pp. 153–179.
[4] A. Burchard, *Steiner symmetrization is continuous in $W^{1,p}$*, Geom. Funct. Anal., 7 (1997), pp. 823–860.

[5] R. Courant, *Dirichlet's Principle, Conformal Mapping, and Minimal Surfaces*, Interscience, New York, 1950.

[6] S. J. Cox and P. X. Uhlig, *Where best to hold a drum fast*, SIAM J. Optim., 9 (1999), pp. 948–964.

[7] M. Dauge, M. S. Lubuma, and S. Nicaise, *Coefficients des singularités pour le problème de Dirichlet sur un polygone*, C. R. Acad. Sci. Paris Sér. I Math., 304 (1987), pp. 483–486.

[8] J. Denzler, *Bounds for the heat diffusion through windows of given area*, J. Math. Anal. Appl., 217 (1998), pp. 405–422.

[9] J. Denzler, *Windows of given area with minimal heat diffusion*, Trans. Amer. Math. Soc., 351 (1999), pp. 569–580.

[10] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[11] D. Gilbarg and L. Hörmander, *Intermediate Schauder estimates*, Arch. Rational Mech. Anal., 74 (1980), pp. 297–318.

[12] P. Grisvard, *Singularities in Boundary Value Problems*, Masson, Paris, Springer-Verlag, 1992.

[13] E. M. Harrell, II, P. Kröger, and K. Kurata, *On the placement of an obstacle or a well so as to optimize the fundamental eigenvalue*, SIAM J. Math. Anal., 33 (2001), pp. 240–259.

[14] T. Kato, *Perturbation Theory for Linear Operators*, Die Grundlehren der mathematischen Wissenschaften, Band 132, Springer-Verlag, New York, 1966.

[15] B. Kawohl, *Rearrangements and Convexity of Level Sets in PDE*, Lecture Notes in Math. 1150, Springer-Verlag, Berlin, New York, 1985.

[16] V. A. Kondrat'ev and O. A. Oleinik, *Boundary value problems for partial differential equations in non-smooth domains*, Russian Math. Surveys 38 (1983), pp. 1–86 (in English); Uspekhi Mat. Nauk 38 (1983), pp. 3–76 (in Russian).

[17] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[18] E. H. Lieb and M. Loss, *Analysis*, Grad. Stud. Math. 14, AMS, Providence, RI, 1997.

[19] J. Malý and W. P. Ziemer, *Fine Regularity of Solutions of Elliptic Partial Differential Equations*, AMS, Providence, RI, 1997.

[20] V. G. Maz'ya, S. A. Nazarov, and B. A. Plamenevskii, *Asymptotic Theory of Elliptic Boundary Value Problems in Singularly Perturbed Domains*, Vol. 1, Birkhäuser-Verlag, Basel, Switzerland, 2000.

[21] V. G. Maz'ya and B. A. Plamenevskii, *On the coefficients in the asymptotics of solutions of elliptic boundary value problems in domains with conical points*, Amer. Math. Soc. Transl. 123 (1984), pp. 57–88 (in English); Math. Nachr. 76 (1977), pp. 29–60 (in Russian).

[22] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* IV*: Analysis of Operators*, Academic Press, New York, 1978.

[23] J. Rossmann, *Contributions of V. Maz'ya to the theory of boundary value problems in non-smooth domains*, in The Maz'ya Anniversary Collection, Vol. 1, J. Rossmann, P. Takáč, and G. Wildenhain, eds., Birkhäuser Boston, Cambridge, MA, 1999.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1828–1853

ON THE DIFFERENTIAL EQUATION $u_{xxxx} + u_{yyyy} = f$ FOR AN ANISOTROPIC STIFF MATERIAL*

B. KAWOHL† AND G. SWEERS‡

Abstract. We study the differential operator $L = \frac{\partial^4}{\partial x^4} + \frac{\partial^4}{\partial y^4}$ and investigate positivity preserving properties in the sense that $f \geq 0$ implies that solutions u of $Lu - \lambda u = f$ are nonnegative. Since the operator is of fourth order we have no maximum principle at our disposal. The operator models the deformation of an anisotropic stiff material like a wire fabric, and it has to be complemented by appropriate boundary conditions. Our operator was introduced by Jacob II Bernoulli as the operator that supposedly models the vibrations of an elastic plate. This model was later revised by Kirchhoff, because the operator and its solutions were anisotropic. Modern materials, however, are often anisotropic, and therefore the old model of Bernoulli deserves an updated investigation. It turns out that even our apparently simple model problem contains some hard analytical challenges.

Key words. orthotropic plate, anisotropic operator, vibrations, spectrum, fourth order elliptic, clamped and hinged boundary condition, positivity of the operator, Green's function, Kirchhoff plate

AMS subject classifications. 35J35, 35J40, 35P10, 74E10, 74G55, 74G40, 74H45, 74K10, 74K20

DOI. 10.1137/050624704

1. Introduction.

Small vertical deformations u of an elastic membrane are usually described by a second order differential equation $-\Delta u = f$ with f denoting the load, whereas the deformation of a plate is commonly modeled by a fourth order equation $\Delta^2 u = f$. Suppose that the membrane is replaced by a piece of material or cloth that is woven out of elastic strings. Then the material properties change drastically, and in [4] such a problem was studied for second order differential operators. If a plate is replaced by a stiff woven material [17] (running in Cartesian directions), its deformation energy can be described by

$$\int_\Omega \left(u_{xx}^2 + u_{yy}^2\right)\ dx\,dy \tag{1}$$

rather than by the functional for the elastic plate [40]

$$\int_\Omega \left((\Delta u)^2 - (1-\sigma)\left(u_{xx}u_{yy} - u_{xy}^2\right)\right) dx\,dy. \tag{2}$$

For the energy that corresponds to the reinforcement or wire fabric that is embedded in, for example, concrete, a linear combination of (1) and (2) is appropriate.

In contrast to the plate equation, that is, the Euler equation for (2) which contains the operator $\Delta^2 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy}$, the linearized differential equation for a stiff fabric consisting of perpendicular fibers does not contain mixed terms when these fibers run parallel to the x- and y-axes. Indeed, if the torsional stiffness can

*Received by the editors February 18, 2005; accepted for publication (in revised form) September 7, 2005; published electronically February 21, 2006. This research was financially supported through the RiP program of the Mathematisches Forschungsinstitut Oberwolfach.
http://www.siam.org/journals/sima/37-6/62470.html
†Mathematisches Institut, Universität zu Köln, D-50923 Köln, Germany (kawohl@math.uni-koeln.de).
‡Delft Institute of Applied Mathematics, EWI, Technische Universiteit Delft, P.O. Box 5031, 2600 GA Delft, The Netherlands (g.h.sweers@ewi.tudelft.nl).

be neglected, the energy stored in the grid under a vertical load f is supposed to be given by

$$E(u) = \int_\Omega \left\{ \frac{1}{2} \left(u_{xx}^2 + u_{yy}^2 \right) - fu \right\} dx\, dy.$$

The corresponding Euler equation is $u_{xxxx} + u_{yyyy} = f$. This equation has to be complemented by suitable boundary conditions, and in the present paper we shall study the problem on a planar domain Ω:

- as a general grid that is hinged at the boundary

$$\begin{cases} u_{xxxx} + u_{yyyy} = f & \text{in } \Omega, \\ u = n_1^2\, u_{xx} + n_2^2\, u_{yy} = 0 & \text{on } \partial\Omega, \end{cases} \tag{3}$$

  where $n = n(x,y)$ is the exterior normal at $(x,y) \in \partial\Omega$;
- or as a general grid that is *clamped* at the boundary

$$\begin{cases} u_{xxxx} + u_{yyyy} = f & \text{in } \Omega, \\ u = \frac{\partial}{\partial n} u = 0 & \text{on } \partial\Omega. \end{cases} \tag{4}$$

When we checked the literature for this type of equation, we found a remarkable hint in the chapter on the history of plate theory in Szabó's book [37, p. 409]. Jacob II Bernoulli, inspired by Chladni's experiments on vibrating plates, had attempted to model their behavior by our differential equation in [5], but this was later dismissed for isotropic plates and replaced by Kirchhoff's theory [24]. However, there is more to it. According to [30], Bernoulli had also studied and absorbed Euler's idea that an elastic membrane should be modeled as a fabric of one-dimensional orthogonal elastic strings, and he tried to carry this idea over to modeling a plate as a fabric of one-dimensional beams. Thus he arrived at

$$\frac{\partial^4 z}{\partial x^4} + \frac{\partial^4 z}{\partial y^4} = \frac{z}{c^4}$$

as *"the fundamental equation of the entire theory"* of plate vibrations. In those days church bells were intended as applications for the theory. Both operators, the isotropic plate operator Δ^2 and the anisotropic

$$L = \frac{\partial^4}{\partial x^4} + \frac{\partial^4}{\partial y^4}, \tag{5}$$

retain a certain degree of isotropy. They are special cases of

$$\tilde{L} = \frac{\partial^4}{\partial x^4} + P \frac{\partial^4}{\partial x^2 \partial y^2} + Q \frac{\partial^4}{\partial y^4} \tag{6}$$

with $P \geq 0$ and $Q > 0$ denoting material constants. Notice that $\tilde{L}$ is always invariant under reflections across Cartesian axes, but not always under rotations. Plates whose deformation is described by such operators are called *orthotropic*; see, e.g., [29], [31]. By scaling y and not scaling x one can always force Q to be 1. Realistic values for P and Q in the case of plywood material (birch with bakelite glue) can be found in [25, p. 92], [26, p. 269], or [31]. It is not unrealistic to expect Q to be of order 1–10 and $P \in [0, 1)$. Modern (composite) materials like GLARE (see http://www.lr.tudelft.nl/highlights/glare.asp), a composite of layers of fiberglass and aluminum that is also called "plymetal," can be expected to satisfy similar orthotropic equations. Orthotropic plate equations like $\tilde{L}u = f$ have been rigorously derived by homogenization methods as the right macroscopic model for grid structures as the thickness of the structure and the size of its cells go to zero. To be precise, in [3, p. 130] and using our notation, the limit equation has coefficients $Q = 1$ and $P = 4/(1+\nu)$, where ν denotes the Poisson ratio of the original (solid and unhomogenized) plate material. For $\nu = -1/3$ one gets $P = 6$, as in (19) below. We take the differential equation for granted here and do not address issues of homogenization as in [3].

Section 2 is devoted to proving existence and uniqueness questions, and section 3 to regularity of solutions to these boundary value problems. Regularity near corners of Ω is delicate, and its discussion will be limited to some special cases. Moreover, we address the subject of representations of solutions by series or by means of a Green function at the end of section 3.

In section 4 we study the spectrum of the operator L on rectangular domains and for hinged and clamped boundary conditions. Since the operator is separable, on special domains like rectangles all of its eigenfunctions can be represented in terms of products of one-dimensional eigenfunctions. We learned this from Courant and Hilbert (see [10, Chapter II, Par. 1.6]), who did it for operators of second order. Therefore the one-dimensional cases will always be treated before the rectangular domains. We present all eigenvalues and eigenfunctions for a number of examples and compare spectra for different (parallel or diagonal) alignments of our anisotropic grid.

Section 5 is dedicated to positivity questions. Suppose that the load f on a beam (or grid) is pointing downwards. Does this imply that the deformation u has the same sign everywhere in Ω? The answer is in general negative, unless the geometry of the domain is special or unless the beam (or grid) is embedded in an elastic ambient medium that exerts a restoring force proportional to the deformation. So the modified question is for which (presumably negative) values of λ one can show that $f \geq 0$ implies positivity of the solution to

$$u_{xxxx} + u_{yyyy} = \lambda u + f \qquad \text{in } \Omega,$$

which satisfies the boundary conditions under consideration. This question turns out to be technically most challenging, and its answer is given using different tricks for different alignments or boundary conditions.

For the reader's convenience we finish with a summary in section 6 and an appendix.

2. Existence and uniqueness for hinged and clamped grids. Let $\Omega \subset \mathbb{R}^n$ be a bounded simply connected set. Then the variational problem

$$\text{(7)} \qquad \text{Minimize:} \quad E(v) = \int_\Omega \left(\frac{1}{2} \sum_{i=1}^{n} v_{x_i x_i}^2 - f\, v \right) dx \quad \text{on } W_0^{1,2}(\Omega) \cap W^{2,2}(\Omega)$$

has a unique solution. To see this directly we follow the ideas of [16] and first show that $E(v)$ is coercive on $W^{2,2}(\Omega)$. Obviously $2u_{xx}u_{yy} \leq u_{xx}^2 + u_{yy}^2$, so that

$$E(v) \geq c(n) \int_\Omega (\Delta v)^2 \, dx - \int_\Omega f\, v \, dx.$$

If we denote Δv by g, then a well-known a priori estimate for solutions of Dirichlet problems for second order elliptic differential equations on bounded domains (see, e.g., [14, p. 317]) implies that

$$||D^2 v||_{L^2(\Omega)} \leq C||\Delta v||_{L^2(\Omega)},$$

so that all second derivatives of v are in $L^2(\Omega)$. This and a Poincaré-type inequality show the coerciveness of E on $W_0^{1,2}(\Omega) \cap W^{2,2}(\Omega)$. Now the existence and uniqueness of a solution follow from the direct method in the calculus of variations and from the strict convexity of the functional E. The solution satisfies the Euler equation $\sum_{i=1}^n u_{x_i x_i x_i x_i} = f$ in Ω. To derive the boundary conditions, we note that a weak solution satisfies

$$\int_\Omega \left(\sum_{i=1}^n u_{x_i x_i} \varphi_{x_i x_i} - f\,\varphi \right) dx = 0, \tag{8}$$

and after two integrations by parts we obtain

$$0 = \int_\Omega \left(-\sum_{i=1}^n u_{x_i x_i x_i} \varphi_{x_i} - f\,\varphi \right) dx + \int_{\partial\Omega} \sum_{i=1}^n u_{x_i x_i} \varphi_{x_i} \nu_i \; d\sigma \tag{9}$$

$$= 0 + \int_{\partial\Omega} \sum_{i=1}^n u_{x_i x_i} \varphi_{x_i} \nu_i \; d\sigma - \int_{\partial\Omega} \sum_{i=1}^n u_{x_i x_i x_i} \; \varphi \; \nu_i \; d\sigma \tag{10}$$

$$= \int_{\partial\Omega} \left(\sum_{i=1}^n u_{x_i x_i} \nu_i^2 \right) \frac{\partial \varphi}{\partial \nu} \; d\sigma. \tag{11}$$

Notice that the last integral in (10) vanishes because φ vanishes on the boundary. Therefore the first boundary integral in (10) must vanish too. The vanishing of φ on $\partial\Omega$ implies in particular that the bracket in (11) must vanish on $\partial\Omega$. Thus we have formally derived (3) in the plane case. See also [32].

If the grid or stiff fabric is clamped, we consider the variational problem

$$\text{Minimize:} \quad E(v) = \int_\Omega \left(\frac{1}{2} \sum_{i=1}^n v_{x_i x_i}^2 - f\,v \right) dx \quad \text{on } W_0^{2,2}(\Omega) \tag{12}$$

and observe that the same existence proof works for this problem, too. The solution satisfies

$$\begin{cases} \sum_{i=1}^n u_{x_i x_i x_i x_i} = f & \text{in } \Omega, \\ \qquad u = \frac{\partial u}{\partial \nu} = 0 & \text{on } \partial\Omega. \end{cases} \tag{13}$$

We have the following existence and uniqueness results.

THEOREM 2.1. *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with piecewise smooth boundary, and suppose that $f \in L^2(\Omega)$. Then problems* (7) *and* (12) *have a unique minimizer. Moreover, the corresponding boundary value problems, which in the case $n = 2$ are given by* (3) *and* (4), *have a unique weak solution.*

Remark 2.1. As usual, a weak solution for (3) is a function u in $W_0^{2,2}(\Omega)$ that satisfies (8) for all $\varphi \in W_0^{2,2}(\Omega)$. A weak solution of (4) is a function u in $W^{2,2}(\Omega) \cap W_0^{1,2}(\Omega)$, satisfying (8) for all $\varphi \in W^{2,2}(\Omega) \cap W_0^{1,2}(\Omega)$.

The existence was shown above by variational methods, and the uniqueness of the weak solution follows from the strict convexity of the underlying functional E. Notice that the second order boundary condition holds only in the sense of distributions. To see that it holds pointwise in every smooth point of the boundary, we need to know more about its regularity.

3. Regularity. One may use the standard regularity theory for elliptic operators whenever the elliptic system is of an appropriate type and if the boundary is sufficiently smooth. First we will show that our systems are regular elliptic.

3.1. Regular elliptic. The symbol, that is $L = \mathcal{L}(\frac{\partial}{\partial x}, \frac{\partial}{\partial y})$, of our fourth order operator can be decomposed as follows:

$$\mathcal{L}(\xi_1, \xi_2) := \xi_1^4 + \xi_2^4 = \left(\xi_1^2 + \sqrt{2}\xi_1\xi_2 + \xi_2^2\right)\left(\xi_1^2 - \sqrt{2}\xi_1\xi_2 + \xi_2^2\right). \tag{14}$$

Hence L can be written as the composition of two second order elliptic operators. Notice, however, that the boundary value problem (4) cannot be split into a system of two second order equations with separated boundary conditions. In fact, even for the boundary value problem (3) this seems to be out of reach. The boundary operators have the following symbols:

- for (3): $\mathcal{B}_1(\xi) = 1$ and $\mathcal{B}_2(\xi) = n_1(x)^2\xi_1^2 + n_2(x)^2\xi_2^2$;
- for (4): $\mathcal{B}_1(\xi) = 1$ and $\mathcal{B}_2(\xi) = n_1(x)\xi_1 + n_2(x)\xi_2$.

A necessary condition for obtaining the full classical regularity results is that the corresponding boundary value problem should be regular elliptic in the sense of [27], and this is indeed the case.

LEMMA 3.1. *Problems* (3) *and* (4) *are regular elliptic.*

Proof. The differential operator is regular elliptic of order $2k$ if there is $c > 0$ such that $\mathcal{L}(\xi) \geq c\,|\xi|^{2k}$ for $\xi \in \mathbb{R}^2$. To show that the operator and boundary conditions constitute a regular elliptic problem in the sense of [27], one has to consider the factorization of $\tau \mapsto \mathcal{L}(\xi + \tau\eta)$. One finds that the roots of this polynomial are

$$\tau_k = -\frac{\xi_1 + (-1)^{\frac{2k-1}{4}}\,\xi_2}{\eta_1 + (-1)^{\frac{2k-1}{4}}\,\eta_2} \qquad \text{with } k \in \{1, 2, 3, 4\}.$$

We use $(-1)^\alpha = \cos\pi\alpha + \boldsymbol{i}\sin\pi\alpha$. Depending on ξ and η, which should be taken independently, there are two roots, τ_I and τ_{II}, which have positive imaginary part. We find $\mathcal{L}(\xi + \tau\eta) = a^+(\xi, \eta; \tau)a^-(\xi, \eta; \tau)$ with

$$a^+(\xi, \eta; \tau) = \sqrt{\eta_1^4 + \eta_2^4}\,(\tau - \tau_I)\,(\tau - \tau_{II}),$$
$$a^-(\xi, \eta; \tau) = \sqrt{\eta_1^4 + \eta_2^4}\,(\tau - \bar{\tau}_I)\,(\tau - \bar{\tau}_{II}).$$

Since the imaginary parts of τ_I and τ_{II} have the same sign the first order term in $a^+(\xi, \eta; \tau)$ has a coefficient with a strictly negative imaginary part; indeed,

$$(\tau - \tau_I)\,(\tau - \tau_{II}) = \tau^2 - (\tau_I + \tau_{II})\,\tau - \tau_I\tau_{II}.$$

The condition for regularity that has to be verified is that, for ξ a tangential direction and η a normal direction, the polynomials $\tau \mapsto \mathcal{B}_1(\xi + \tau\eta)$ and $\tau \mapsto \mathcal{B}_2(\xi + \tau\eta)$ are independent modulo $\tau \mapsto a^+(\xi, \eta; \tau)$. Therefore we set $\eta = (n_1, n_2)$ and $\xi = (-n_2, n_1)$.

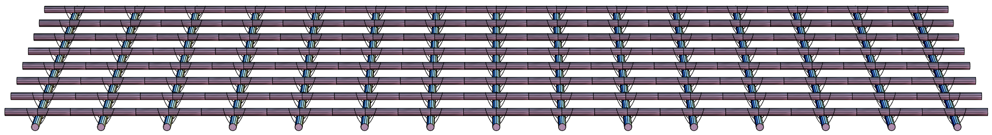


FIG. 1. *A rectangular wire fabric with fibers in Cartesian directions.*

For (3), $\mathcal{B}_1(\xi + \tau\eta) = 1$ and

$$\begin{aligned}\mathcal{B}_2(\xi + \tau\eta) &= n_1^2(-n_2 + n_1\tau)^2 + n_2^2(n_1 + n_2\tau)^2 \\ &= 2n_1^2n_2^2 + (n_2^2 - n_1^2)n_1n_2\tau + \left(n_1^4 + n_2^4\right)\tau^2.\end{aligned}$$

This is a polynomial with only real coefficients. Since $a^+(\xi, \eta; \tau)$ contains a real second order term and an imaginary first order term, both polynomials are linearly independent.

For (4), $\mathcal{B}_1(\xi + \tau\eta) = 1$ and $\mathcal{B}_2(\xi + \tau\eta) = 1 + \tau$. These are clearly independent modulo any second order polynomial. $\square$

3.2. Regularity for smooth domains. Near the smooth boundary part the standard regularity results, e.g., from [27, Ch. 2], may be used, since both the clamped and the hinged problems are regular elliptic. Only the corners need more attention. But to fix the facts let us summarize the regularity results in a theorem.

THEOREM 3.2. *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with piecewise smooth boundary, and let $\Omega' \subset \Omega$ be a subset such that $\overline{\Omega'}$ contains only the smooth boundary points of $\partial\Omega$. If $f \in W^{k,2}(\Omega)$ and $k \in \{0, 1, 2, \dots\}$, then the weak solutions of* (3) *and* (4) *are of class $W^{k+4}(\Omega')$. In particular, for $f \in L^2(\Omega)$ the derivatives $u_{x_ix_i}$ are in $W^{3/2,2}(\partial\Omega \cap \partial\Omega')$, so that the boundary condition in* (3) *holds pointwise a.e. on $\partial\Omega$.*

3.3. Regularity near corners. It remains to discuss the regularity near singular points of the boundary, and this will be done for some special but typical situations. First we will give an explanation for a simple case.

3.3.1. The hinged rectangular grid with aligned fibers. Let $R = (0, a) \times (0, b)$ be the rectangle. It will be relatively easy to study the regularity of the *hinged grid* near a corner, say $(0, 0)$, when the grid is *aligned with the rectangle* as in Figure 1.

Reflection. The first approach is through a reflection argument. As an example we will consider the hinged rectangular grid with horizontally and vertically aligned fibers.

Note that the differential operator and boundary conditions all satisfy

$$\mathcal{L}\left(\pm\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right) = \mathcal{L}\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right) \quad \text{and} \quad \mathcal{B}\left(\pm\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right) = \mathcal{B}\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right).$$

Instead of considering

$$\left\{\begin{aligned} u_{xxxx} + u_{yyyy} &= f && \text{in } R, \\ u = u_{xx} &= 0 && \text{on } \{0, a\} \times [0, b], \\ u = u_{yy} &= 0 && \text{on } [0, a] \times \{0, b\}, \end{aligned}\right. \tag{15}$$

we extend f to $\tilde{f}$ on $(-a, a) \times (0, b)$ by

$$\tilde{f}(x, y) = \operatorname{sign}(x)\ f(|x|, y)$$

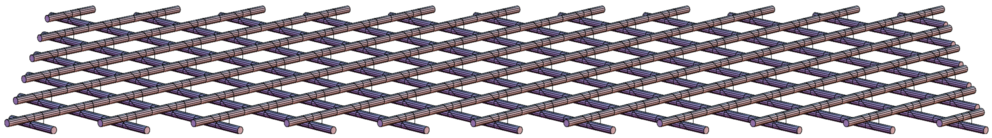


Fig. 2. *Rectangular grid with diagonal fabric.*

and consider

$$(16)\qquad \begin{cases} \tilde{u}_{xxxx} + \tilde{u}_{yyyy} = \tilde{f} & \text{in } \tilde{R} = (-a,a)\times(0,b)\,, \\ \tilde{u} = \tilde{u}_{xx} = 0 & \text{on } \{-a,a\}\times[0,b]\,, \\ \tilde{u} = \tilde{u}_{yy} = 0 & \text{on } [-a,a]\times\{0,b\}\,. \end{cases}$$

If $f \in L^p(R)$, then $\tilde{f} \in L^p(\tilde{R})$, and by the result above there is unique solution $\tilde{u} \in W^{2,p}(\tilde{R})$ and $\tilde{u} \in W^{4,p}(\tilde{R}\backslash N)$ with N some neighborhood of the four corners of $\tilde{R}$; $(0,0)$ has become a regular boundary point. Since the solution $\tilde{u}$ is unique one finds that $\tilde{u}(x,y) = -\tilde{u}(-x,y)$ and hence $\tilde{u}(0,y) = \tilde{u}_{xx}(0,y) = 0$. In other words, $u := \tilde{u}_{|R}$ is the solution of (15) which is in $W^{4,p}(R \cap B_\varepsilon(0))$. Since we may do so for every corner of R we find that $u \in W^{4,p}(R)$.

Separation of eigenfunctions. A second approach can be used if there is a complete orthonormal system of eigenfunctions of the form $\{\varphi_i(x)\psi_j(y); i,j \in \mathbb{N}\}$. For example, for the problem (15) the set $\{\Phi_{ij}; i,j \in \mathbb{N}^+\}$ with

$$\Phi_{i,j}(x,y) = \frac{2}{\sqrt{ab}} \sin\left(i\frac{\pi}{a}x\right) \sin\left(j\frac{\pi}{b}y\right)$$

is a complete orthonormal set of eigenfunctions. Writing $f_{ij} = \langle \Phi_{i,j}, f\rangle$, the solution u is given by

$$u(x,y) = \sum_{i,j=1}^{\infty} \frac{f_{ij}}{\left(i\frac{\pi}{a}\right)^4 + \left(j\frac{\pi}{b}\right)^4} \Phi_{i,j}.$$

Using Parseval's identity, a straightforward computation shows that

$$\left\| \left(\frac{\partial}{\partial x}\right)^k \left(\frac{\partial}{\partial y}\right)^\ell u \right\|^2_{L^2(R)} = \sum_{i,j=1}^{\infty} \frac{\left(i\frac{\pi}{a}\right)^{2k} \left(j\frac{\pi}{b}\right)^{2\ell}}{\left(\left(i\frac{\pi}{a}\right)^4 + \left(j\frac{\pi}{b}\right)^4\right)^2} (f_{ij})^2\,,$$

which is bounded if $f \in L^2(R)$ and $k,l \in \mathbb{N}$ with $k+l \le 4$. Thus $u \in W^{4,2}(R)$.

3.3.2. The hinged rectangular grid with diagonal fibers. Now suppose that the grid runs *diagonally* into the horizontal and vertical axes, as in Figure 2, and that $\hat{x} := \frac{1}{2}\sqrt{2}\,(x+y)$ and $\hat{y} = \frac{1}{2}\sqrt{2}\,(y-x)$. Then a straightforward calculation shows that

$$(17)\qquad u_{xxxx} + u_{yyyy} = \frac{1}{2} u_{\hat{x}\hat{x}\hat{x}\hat{x}} + \frac{6}{2} u_{\hat{x}\hat{x}\hat{y}\hat{y}} + t\frac{1}{2} u_{\hat{y}\hat{y}\hat{y}\hat{y}} = f$$

while the boundary condition from (10) becomes

$$(18)\qquad u_{xx} + u_{yy} = \Delta u = 0 = u_{\hat{x}\hat{x}} + u_{\hat{y}\hat{y}}$$

because $(\nu_1)^2 = (\nu_2)^2 = 1/2$ on the sides of the rectangle and because the Laplacian is invariant under rotations. Since also $u = 0$ on the boundary, this implies $u_{\hat{x}\hat{x}} = 0 = u_{\hat{y}\hat{y}}$. Therefore after an obvious change of notation the deformation u of the *hinged diagonal* grid satisfies again a regular elliptic boundary value problem, namely,

$$\left\{ \begin{aligned} u_{xxxx} + 6u_{xxyy} + u_{yyyy} &= 2f && \text{in } R = (0,a) \times (0,b)\,, \\ u = u_{xx} &= 0 && \text{on } \{0,a\} \times [0,b]\,, \\ u = u_{yy} &= 0 && \text{on } [0,a] \times \{0,b\}\,. \end{aligned} \right. \tag{19}$$

Also for this boundary value problem we find that $\mathcal{L}(\pm\frac{\partial}{\partial x}, \frac{\partial}{\partial y}) = \mathcal{L}(\frac{\partial}{\partial x}, \frac{\partial}{\partial y})$ and $\mathcal{B}(\pm\frac{\partial}{\partial x}, \frac{\partial}{\partial y}) = \mathcal{B}(\frac{\partial}{\partial x}, \frac{\partial}{\partial y})$, and hence we may use the odd reflection argument of (16) to find $u \in W^{4,2}(R)$ that satisfies the boundary conditions for $x = 0$.

Incidentally, the transformed elliptic operator has a symbol that can again be factorized as

$$2\hat{\mathcal{L}}(\xi_1, \xi_2) := \xi_1^4 + 6\xi_1^2\xi_2^2 + \xi_2^4 = \left(\xi_1^2 + \left(3 - 2\sqrt{2}\right)\xi_2^2\right)\left(\xi_1^2 + \left(3 + 2\sqrt{2}\right)\xi_2^2\right). \tag{20}$$

The fact that (19) constitutes a regular elliptic boundary value problem does not need to be checked again, since this property is invariant under changes of the coordinate system. Moreover, the boundary conditions fit nicely with this factorization, and we find a system of two well-posed second order problems:

$$\left\{ \begin{aligned} u_{xx} + \left(3 - 2\sqrt{2}\right) u_{yy} &= v && \text{in } R, \\ u &= 0 && \text{on } \partial R, \\ v_{xx} + \left(3 + 2\sqrt{2}\right) v_{yy} &= 2f && \text{in } R, \\ v &= 0 && \text{on } \partial R. \end{aligned} \right. \tag{21}$$

Using the result of Kadlec [22] for second order operators on convex domains, one finds that $f \in L^2(R)$ implies $v \in W^{2,2}(R) \cap W_0^{1,2}(R)$. Since v satisfies the boundary conditions of u (!), we find not only that $u \in W^{2,2}(R) \cap W_0^{1,2}(R)$ but even that $u \in W^{4,2}(R) \cap W_0^{1,2}(R)$.

We remark that some boundary value problems with different boundary conditions along each side can be treated by a reflection argument. In fact, the same reflection argument works for the aligned rectangular grid if it is clamped on the horizontal parts of the boundary and hinged on the vertical part. To be specific, for $f \in L^2(\Omega)$ the unique solution u of

$$\left\{ \begin{aligned} u_{xxxx} + u_{yyyy} &= f && \text{in } R, \\ u = u_{xx} &= 0 && \text{on } \{0,a\} \times [0,b]\,, \\ u = u_y &= 0 && \text{on } [0,a] \times \{0,b\} \end{aligned} \right. \tag{22}$$

is in $W^{4,2}(R)$.

3.3.3. The clamped rectangular grid with aligned fibers. The regularity of the *clamped* grid near a corner does not follow from such a simple reflection argument, because u_{xx} does in general not vanish on $(0, y)$ with $y \in (0, b)$. However, provided the grid is *aligned* with the rectangle as in Figure 1, we may proceed by "separation of eigenfunctions." To complete this argument we need to borrow some results of subsection 4.2.3 and more specifically Lemmas 4.3 and 4.4.

The set $\{\Phi_{ij}\}$ of eigenfunctions in (34) is a complete orthonormal system in $L^2(\Omega)$. Then, as above for the hinged rectangular grid, the solution of (4) can be represented by

$$u(x,y) = \sum_{i,j=1}^{\infty} \frac{\alpha_{ij}}{\Gamma_{ij}} \Phi_{ij}(x,y),$$

where α_{ij} are the Fourier coefficients from the representation of f with $\sum_{i,j} \alpha_{ij}^2$ being finite by Parseval's identity. The eigenvalues Γ_{ij} are defined by $\Gamma_{ij} = a^{-4}\lambda_i + b^{-4}\lambda_j$, and we find that

$$\frac{\partial^{n+m}}{\partial x^n \partial y^m} u(x,y) = \sum_{i,j=1}^{\infty} \frac{a^{-n}\lambda_i^{n/4} b^{-m}\lambda_j^{m/4}}{a^{-4}\lambda_i + b^{-4}\lambda_j} \alpha_{ij}\Phi_{ij}(x,y),$$

which is bounded when $n+m \le 4$. This shows that $u \in W^{4,2}(R)$ even in this case of a *clamped rectangular grid* aligned with R. From Theorem 2.1 we were allowed to conclude only that $u \in W^{2,2}(R)$.

3.3.4. The clamped rectangular grid with diagonal fibers. How to obtain the regularity of u for a *clamped–hinged* or *clamped–clamped* diagonal grid near a corner is a nontrivial problem and will not be discussed here.

One conceivable way to represent a solution would be using a Green function $g^{0,\xi}(\cdot) = F(\cdot - \xi) + h(\xi, \cdot)$. This can in principle be obtained by adding a solution h of $Lh(\xi,\cdot) = 0$ in R, $h(\xi,\cdot) + F(\cdot - \xi) = 0 = B_2 h(\xi,\cdot) + B_2 F(\cdot - \xi)$ on ∂R to a fundamental solution $F(\cdot - \xi)$, i.e., to a distributional solution of $LF(\cdot - \xi) = \delta_0(\cdot)$. Clearly F is not unique, but just for the record let us quote a fundamental solution F (for L as in (22)) from [39] or [29]:

(23)
$$F(x,y) = -\frac{1}{16\pi\sqrt{2}} \left[(x^2+y^2)\log(x^4+y^4) + 2\sqrt{2}xy \log\left(\frac{x^2+y^2+\sqrt{2}xy}{x^2+y^2-\sqrt{2}xy}\right) + 2\sqrt{2}\left(x^2 \arctan\frac{x^2}{y^2} + y^2 \arctan\frac{y^2}{x^2}\right)\right].$$

An explicit calculation of the Green function even on a quarter plane seems to be beyond reach.

4. Eigenfunctions and eigenvalues.

4.1. Eigenfunctions for a hinged rectangular grid. The eigenfunctions for the *hinged beam*

$$\begin{cases} \varphi_{xxxx} = \lambda\varphi & \text{in } (0,1), \\ \varphi = \varphi_{xx} = 0 & \text{in } \{0,1\} \end{cases} \tag{24}$$

are obviously given by $\phi_i(x) = \sqrt{2}\sin(i\pi x)$, and the eigenvalues are $\lambda_i = i^4\pi^4$.

If a *hinged grid* is rectangular and *aligned with the Cartesian coordinates*, then a calculation shows that the eigenfunctions and eigenvalues of

$$\begin{cases} \Phi_{xxxx} + \Phi_{yyyy} = \Lambda\Phi & \text{in } R, \\ \Phi = \Phi_{xx} = 0 & \text{on } \{0,a\} \times [0,b], \\ \Phi = \Phi_{yy} = 0 & \text{on } [0,a] \times \{0,b\} \end{cases} \tag{25}$$

are given by

$$\Phi_{ij}(x,y) = \frac{2}{\sqrt{ab}} \sin\left(\frac{i\pi x}{a}\right) \sin\left(\frac{j\pi y}{b}\right) \quad \text{and} \quad \Lambda_{ij} = \frac{i^4\pi^4}{a^4} + \frac{j^4\pi^4}{b^4}\ . \tag{26}$$

For $i = j = 1$ one finds the following result.

LEMMA 4.1. *The first eigenfunction for* (25)*, the hinged rectangular grid with aligned fibers, is of fixed sign.*

Even if the *hinged grid* is *diagonally aligned*, we can determine the eigenfunctions and eigenvalues of

$$\begin{cases} \frac{1}{2}\Phi_{xxxx} + 3\Phi_{xxyy} + \frac{1}{2}\Phi_{yyyy} = \tilde{\Lambda}\Phi & \text{in } R, \\ \Phi = \Phi_{xx} = 0 & \text{on } \{0,a\} \times [0,b], \\ \Phi = \Phi_{yy} = 0 & \text{on } [0,a] \times \{0,b\} \end{cases} \tag{27}$$

by a separation of variables. In fact, the eigenfunctions are still given by

$$\Phi_{ij}(x,y) = \frac{2}{\sqrt{ab}} \sin\left(\frac{i\pi x}{a}\right) \sin\left(\frac{j\pi y}{b}\right),$$

but now the eigenvalues are given by

$$2\tilde{\Lambda}_{ij} = \pi^4 \left(\frac{i^4}{a^4} + \frac{6i^2j^2}{a^2b^2} + \frac{j^4}{b^4}\right). \tag{28}$$

We may conclude as before, with the following claim.

LEMMA 4.2. *The first eigenfunction for* (27)*, the hinged rectangular grid with diagonal fibers, is of fixed sign.*

Notice that

$$\tfrac{1}{2}\Lambda_{ij} \le \tilde{\Lambda}_{ij} \le 2\Lambda_{ij}, \tag{29}$$

$$\tilde{\Lambda}_{ij} = \frac{\pi^4}{2}\left(\frac{i^4}{a^4} + \frac{6i^2j^2}{a^2b^2} + \frac{j^4}{b^4}\right) \quad \text{and} \quad \Lambda_{ij} = \pi^4\left(\frac{i^4}{a^4} + \frac{j^4}{b^4}\right).$$

Notice also that the first eigenfunction is of fixed sign.

4.2. Eigenfunctions for clamped problems.

4.2.1. Eigenfunctions for the clamped beam. The set of all normalized eigenfunctions for

$$\begin{cases} \varphi'''' = \lambda\varphi & \text{in } (0,1), \\ \varphi(0) = \varphi'(0) = 0 = \varphi(1) = \varphi'(1) \end{cases} \tag{30}$$

forms a complete orthonormal system in $L^2(0,1)$.

LEMMA 4.3. *These eigenfunctions and eigenvalues are*

$$\varphi_i(x) = \beta_i \cosh\nu_i \left(\frac{\cosh(\nu_i x) - \cos(\nu_i x)}{\cosh\nu_i - \cos\nu_i} - \frac{\sinh(\nu_i x) - \sin(\nu_i x)}{\sinh\nu_i - \sin\nu_i}\right) \quad \textit{and} \quad \lambda_i = \nu_i^4,$$

with $i = 1, 2, \ldots,$ *where* ν_i *is the ith positive zero of* $\cos\nu - \frac{1}{\cosh\nu} = 0$ *and* β_i *is the normalization factor such that* $\int_0^1 \varphi_i(x)^2\, dx = 1$.

Note that the first eigenfunction is of fixed sign.

The statement of this lemma is shown by a lengthy but straightforward calculation.

TABLE 1
Comparison of the of the eigenvalues λ_i and the approximation in Lemma 4.4.

λ_i :	500.56390	3803.5370	14617.630	39943.799	89135.406	173881.31	308208.45
$\left(i-\frac{1}{2}\right)^4\pi^4$:	493.13352	3805.0426	14617.451	39943.815	89135.406	173881.31	308208.45

LEMMA 4.4. *The sequences ν_i and β_i as above have the following asymptotics:*
- $\lim_{i\to\infty} i\pi - \nu_i = \frac{1}{2}\pi$ *and hence* $\lambda_i \approx (i-1/2)^4\,\pi^4$;
- $\lim_{i\to\infty} \beta_i = 1$.

Proof. For obvious reasons two subsequent zeroes ν_i and ν_{i+1} of $\cos\nu - \frac{1}{\cosh\nu} = 0$ are in the interval $((i-\frac{1}{2})\pi, (i+\frac{1}{2})\pi)$ and close to its boundaries. Since $\frac{1}{\cosh\nu_i} \le 2e^{-\nu_i}$ and $|\sin x| > \frac{1}{2}$ in a sufficiently small neighborhood of $(i-\frac{1}{2})\pi$ we have

$$|\nu_i - (i-\tfrac{1}{2})\pi| < 4\ e^{\pi/2}\ e^{-i\pi}.$$

This proves the first statement, as Table 1 illustrates.

Let us now turn to the second statement. With the help of Mathematica one sees that

$$\beta_i = \frac{-Z_i(\cosh\nu_i)^2}{4\,\nu_i\ (\cos\nu_i - \cosh\nu_i)^2\ (\sin\nu_i - \sinh\nu_i)^2}$$

with

$$\begin{aligned} Z_i = \quad & 2\,\nu_i\ \cos 2\,\nu_i + 4\ \cosh\nu_i\ \sin\nu_i - \sin 2\,\nu_i - \cosh 2\,\nu_i\ (2\,\nu_i + \sin 2\,\nu_i) \\ & -\ 4\ \cos\nu_i\ \sinh\nu_i + 8\,\nu_i\ \sin\nu_i\ \sinh\nu_i + \sinh 2\,\nu_i + \cos 2\,\nu_i\ \sinh 2\,\nu_i. \end{aligned}$$

Now the second statement follows by a straightforward computation. □

4.2.2. Comparing eigenvalues of clamped plates and grids. In [33] Philippin, following ideas of Hersch [20], obtained estimates for the eigenvalues of a clamped plate through these for clamped rectangular and diagonal grids. Let us state a special result in this direction that compares the first eigenvalues of a clamped grid,

$$\left\{ \begin{array}{rl} \Phi_{xxxx} + \Phi_{yyyy} = \Gamma\Phi & \text{in } \Omega, \\ \Phi = |\nabla\Phi| = 0 & \text{on } \partial\Omega, \end{array} \right. \tag{31}$$

with those of the clamped plate,

$$\left\{ \begin{array}{rl} \Delta^2\Phi = \Upsilon\Phi & \text{in } \Omega, \\ \Phi = |\nabla\Phi| = 0 & \text{on } \partial\Omega. \end{array} \right. \tag{32}$$

LEMMA 4.5. *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with a $C^{0,1}$-boundary. Let Γ_1 and Υ_1 be the first eigenvalues of* (31), *respectively* (32). *Then it holds that* $\frac{1}{2}\Upsilon_1 \le \Gamma_1 \le \Upsilon_1$.

For approximations of the lowest eigenvalues for the clamped aligned square grid and the clamped square plate, see Tables 2 and 3.

Proof. The result follows from the definition of the eigenvalue by Rayleigh's quotient and some energy estimates. For the first inequality one uses

$$\frac{1}{4}\int_\Omega (\Delta u)^2 dx\,dy = \frac{1}{4}\int_\Omega \left(u_{xx}^2 + 2u_{xx}u_{yy} + u_{yy}^2\right) dx\,dy \le \frac{1}{2}\int_\Omega \left(u_{xx}^2 + u_{yy}^2\right) dx\,dy.$$

TABLE 2

Numerical eigenvalues Λ_{ij}, with $i, j \le 5$, of a clamped square grid of length 1 that is aligned with Cartesian coordinates (without repeating the multiple ones like $\Lambda_{1,2} = \Lambda_{2,1}$).

Γ_{ij} :	1001.13	4304.10	15118.2	40444.4	89636.0
		7607.07	18421.2	43747.3	92938.9
			29235.3	54561.	103753.
				79887.	129079.
					178271.

TABLE 3

Numerical eigenvalues for a clamped square plate of length 1*. We used the values found in* [15, p. 79] *and scaled these.*

Υ_{ij} :	1294.93	5386.63
		11710.3

For the second one proceeds via an integration by parts that shows, due to the clamped boundary conditions,

$$\begin{aligned} \int_\Omega u_{xx} u_{yy}\, dx\, dy &= \int_{\partial\Omega} (u_x u_{yy} n_1 - u_x u_{xy} n_2)\, d\sigma + \int_\Omega u_{xy}^2\, dx\, dy \\ &= \int_\Omega u_{xy}^2\, dx\, dy \ge 0, \end{aligned}$$

and hence

$$\frac{1}{2} \int_\Omega \left(u_{xx}^2 + u_{yy}^2\right) dx\, dy \le \frac{1}{2} \int_\Omega \left(u_{xx}^2 + 2u_{xy}^2 + u_{yy}^2\right) dx\, dy = \frac{1}{2} \int_\Omega (\Delta u)^2\, dx\, dy.$$

This completes the proof. $\square$

4.2.3. Eigenfunctions for the clamped rectangular grid. A complete orthonormal system of eigenfunctions and eigenvalues for the grid *aligned with the Cartesian coordinates*

$$\left\{ \begin{array}{ll} \Phi_{xxxx} + \Phi_{yyyy} = \Gamma\Phi & \text{in } R, \\ \Phi = |\nabla\Phi| = 0 & \text{on } \partial R \end{array} \right. \tag{33}$$

with $R = (0, a) \times (0, b)$ is given in terms of the one-dimensional eigenfunctions and eigenvalues φ_j and λ_j from Lemma 4.3 by

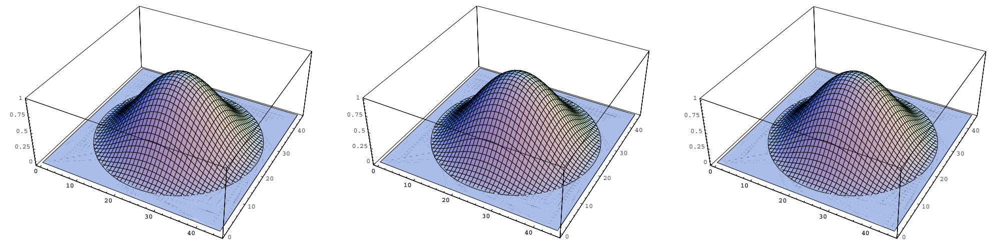
$$\Phi_{ij}(x, y) = \frac{1}{\sqrt{ab}} \varphi_i \left(\frac{x}{a}\right) \varphi_j \left(\frac{y}{b}\right) \quad \text{and} \quad \Gamma_{ij} = a^{-4}\lambda_i + b^{-4}\lambda_j. \tag{34}$$

LEMMA 4.6. *The first eigenfunction for* (33)*, the clamped rectangular grid with aligned fibers, is of fixed sign.*

This is in marked contrast to the biharmonic operator, whose first eigenfunction under Dirichlet conditions on a rectangle is known to change sign infinitely often (see [8]), and positivity of the ground state for our anisotropic operator cannot be expected for a general domain.

An explicit determination of all eigenfunctions and eigenvalues for the *diagonally aligned clamped grid*, however,

$$\left\{ \begin{array}{ll} \frac{1}{2}\Phi_{xxxx} + 3\Phi_{xxyy} + \frac{1}{2}\Phi_{yyyy} = \tilde{\Gamma}\Phi & \text{in } R, \\ \Phi = \Phi_x = 0 & \text{on } \{0, a\} \times [0, b], \\ \Phi = \Phi_y = 0 & \text{on } [0, a] \times \{0, b\}, \end{array} \right. \tag{35}$$



$L_1 = \frac{\partial^4}{\partial x^4} + \frac{\partial^4}{\partial y^4}$ $L_2 = \frac{1}{2}\frac{\partial^4}{\partial x^4} + 3\frac{\partial^4}{\partial x^2 \partial y^2} + \frac{1}{2}\frac{\partial^4}{\partial y^4}$ $L_3 = \Delta^2$

FIG. 3. *Numerical approximations of the first "clamped" eigenfunctions on a disk for L_i, $i = 1, 2, 3$. One sees hardly any difference. We remark that the eigenfunctions for the first and the second operator differ "analytically" just by a 45° rotation. The finite difference scheme, however, is different since in each case the discrete version of the corresponding operator L_i has been used.*

seems to be a nontrivial problem. From Lemma 4.5 we can find an estimate, namely $\Gamma_1 \le 2\tilde{\Gamma}_1$, by using $\Gamma_1 \le \Upsilon_1$ and $\frac{1}{2}\Upsilon_1 \le \tilde{\Gamma}_1$, and similarly $\tilde{\Gamma}_1 \le 2\Gamma_1$. This is consistent with inequality (29) for hinged grids. Note that the estimate $\frac{1}{2}\tilde{\Gamma}_1 \le \Gamma_1 \le 2\tilde{\Gamma}_1$ even holds on general domains.

Remark 4.1. We do not know if the first eigenfunction for (35), the clamped rectangular grid with diagonal fibers, is of fixed sign. Some evidence against a fixed sign follows from Coffman's result in [8].

4.2.4. Eigenfunctions for the clamped circular grid. For a clamped circular plate there are radially symmetric eigenfunctions, and these can be expressed in terms of the (modified) Bessel functions J_0 and I_0. Since Boggio [2] gave an explicit formula for the Dirichlet biharmonic on a circular disk Jentzsch's theorem implies that the first eigenfunction is positive (of fixed sign) and unique and hence radially symmetric. Although a numerical approximation shows that the first eigenfunction of the clamped grid looks similar to the one for the clamped plate (see Figure 3), this eigenfunction is not radially symmetric.

LEMMA 4.7. *Let D denote the unit disk. There is no radial eigenfunction for*

$$\begin{cases} \Phi_{xxxx} + \Phi_{yyyy} = \Gamma\Phi & \textit{in } D, \\ \Phi = |\nabla\Phi| = 0 & \textit{on } \partial D. \end{cases} \tag{36}$$

Remark 4.2. Of course, since the differential equation $u_{xxxx} + u_{yyyy} = \lambda u$ is not rotation invariant, this result should not come as a surprise. A nasty consequence, however, is that the first eigenfunction does not seem to have an "easy" explicit representation. We do not even have analytical proof that this eigenfunction has a fixed sign or that it is unique.

Remark 4.3. The first eigenvalue $\lambda_{cp,1} \approx 104.363$ for the circular clamped plate may be found in [1]. The first one for the clamped grid is approximately 75% of this value.

Proof of Lemma 4.7. Suppose that Φ is a radial eigenfunction. Then we can rotate this eigenfunction by $\pi/4$ and it is still an eigenfunction. However, the rotated Φ now satisfies (see (19))

$$\begin{cases} \frac{1}{2}\Phi_{xxxx} + 3\Phi_{xxyy} + \frac{1}{2}\Phi_{yyyy} = \Gamma\Phi & \text{in } D, \\ \Phi = |\nabla\Phi| = 0 & \text{on } \partial D. \end{cases} \tag{37}$$

Consequently we can add (36) to (37) to arrive at $\frac{3}{2}\Phi_{xxxx}+3\Phi_{xxyy}+\frac{3}{2}\Phi_{yyyy}=2\Gamma\Phi$ or

$$\left\{\begin{array}{ll} \Delta^2\ \Phi=\frac{4}{3}\Gamma\Phi & \text{in } D,\\ \Phi=|\nabla\Phi|=0 & \text{on } \partial D. \end{array}\right. \tag{38}$$

But then Φ must be a radial eigenfunction of the plate equation, an unlikely coincidence. To show that this cannot be the case suppose that $\Phi(r)$ is such a radial function. Then

$$\Phi_{xx}=\Phi''\frac{x^2}{r^2}+\Phi'\frac{y^2}{r^3}\quad\text{and}\quad\Phi_{xx}=\Phi''\frac{y^2}{r^2}+\Phi'\frac{x^2}{r^3}$$

and

$$\begin{aligned}&\Phi_{xxxx}+\Phi_{yyyy}\\ &=\Phi''''\frac{r^4-2x^2y^2}{r^4}+\Phi'''\frac{12x^2y^2}{r^5}+\Phi''\frac{3r^4-30x^2y^2}{r^6}+\Phi'\frac{-3r^4+30x^2y^2}{r^7}=\Gamma\Phi\ ,\end{aligned}$$

or equivalently

$$2x^2y^2\left(-\Phi''''+6r^{-1}\Phi'''-15r^{-2}\Phi''+15r^{-3}\Phi'\right)=r^4\Gamma\Phi-r^4\Phi''''-3r^2\Phi''+3r\Phi'.$$

But this implies that either x^2y^2 is a function of r or both sides are identical zero. Thus we have to show that this second case cannot occur. Suppose that both sides are identical zero. The solutions of

$$-\Phi''''+6r^{-1}\Phi'''-15r^{-2}\Phi''+15r^{-3}\Phi'=0$$

that satisfy the boundary conditions are the functions $(ar^2+b)(1-r^2)^2$. No nonzero a and b will make the right-hand side identically zero, a contradiction. $\square$

5. Positivity questions. From the Krein–Rutman theorem one knows that for a regular elliptic problem strong positivity of the solution operator implies that the first eigenfunction has multiplicity one and, moreover, is of fixed sign. If the solution operator has an integral kernel, one may even use a much earlier result of Jentzsch [21]. Let us be more precise and consider

$$\left\{\begin{array}{ll} Lu=\lambda u+f & \text{in } \Omega,\\ Bu=0 & \text{on } \partial\Omega. \end{array}\right. \tag{39}$$

If the solution operator $(L-\lambda)_B^{-1}:X\mapsto X$ for (39) in the Banach lattice X is compact, positive, and irreducible for some λ_0, then there exists an eigenvalue $\lambda_1\in(\lambda_0,\infty)$ with a positive eigenfunction. For a precise statement, see [6]. Moreover, for all $\lambda\in[\lambda_0,\lambda_1)$ and $f\in X$ one finds that there is a solution $u_\lambda\in X$ and

$$f>0\quad\text{implies}\quad u_\lambda>0.$$

5.1. Known results for plates. Let us recall some of the known positivity preserving results for plates.

For the *hinged plate*

$$\left\{\begin{array}{ll} \Delta^2 u=\lambda u+f & \text{in } \Omega,\\ u=\Delta u=0 & \text{on } \partial\Omega; \end{array}\right. \tag{40}$$

this question was studied in [23] on a general bounded domain Ω. The problem is positivity preserving if $\lambda \in [-\lambda_c(\Omega), \lambda_1(\Omega)^2)$. Here λ_c is a critical number which is bounded above by $\lambda_1(\Omega)\lambda_2(\Omega)$, and $\lambda_i(\Omega)$ are the eigenvalues of the Laplacian operator under Dirichlet conditions. If Ω is a rectangle R with sides a and $b < a$, one can easily calculate $\lambda_1(R) = \pi^2(a^2 + b^2)$ and $\lambda_2(R) = \pi^2(a^2 + 4b^2)$, so that (40) with $\Omega = R$ is positivity preserving for

$$-\pi^4(a^2 + b^2)(a^2 + 4b^2) \leq -\lambda_c(R) \leq \lambda < \pi^4(a^2 + b^2)^2. \tag{41}$$

The *clamped plate*

$$\begin{cases} \Delta^2 u = \lambda u + f & \text{in } \Omega, \\ u = |\nabla u| = 0 & \text{on } \partial\Omega \end{cases} \tag{42}$$

is a more delicate problem. In general (42) is not positivity preserving for $\lambda = 0$; see [13] or [36]. The boundary value problem in (42) is positivity preserving only in the case of the following special domains Ω:

- If Ω is a ball or a disk, Boggio's explicit formula for the solution operator with $\lambda = 0$ implies positivity.
- For small perturbations of the disk, positivity has been shown in [19].
- For Ω some limaçons positivity can be found in [11].
- For a combination of the above results with Möbius transformations, see [12].

5.2. Positivity under hinged boundary conditions.

5.2.1. Hinged beam. For the *hinged beam*

$$\begin{cases} u_{xxxx} = \lambda u + f & \text{in } (0,1), \\ u = u_{xx} = 0 & \text{in } \{0,1\}, \end{cases} \tag{43}$$

the boundary value problem (43) is positivity preserving, provided that (see [23])

$$-950.884 \approx \lambda_c \leq \lambda < \pi^4 \approx 97.409. \tag{44}$$

Here the lower bound λ_c equals $4(\kappa_0)^4$, where κ_0 is the first positive zero of $\tan(x) + \tanh(x)$.

The Green function of the hinged beam. For the sake of completeness we list some facts about the Green function of the hinged beam problem (43). We set $\nu = \sqrt[4]{\lambda}$ and $\mu = \sqrt[4]{-\frac{1}{4}\lambda}$,

$$\phi(\lambda; x) = \begin{cases} \frac{\sinh \nu x - \sin \nu x}{\nu^3} & \text{if } \lambda > 0, \\ \frac{1}{3}x^3 & \text{if } \lambda = 0, \\ \frac{\cosh(\mu x)\sin \mu x - \cos(\mu x)\sinh \mu x}{2\mu^3} & \text{if } \lambda < 0, \end{cases}$$

$$\psi(\lambda; x) = \begin{cases} \frac{\sinh \nu x + \sin \nu x}{2\nu} & \text{if } \lambda > 0, \\ x & \text{if } \lambda = 0, \\ \frac{\cosh(\mu x)\sin \mu x + \cos(\mu x)\sinh \mu x}{2\mu} & \text{if } \lambda < 0. \end{cases}$$

TABLE 4
These values for the coefficients in (45) were obtained using Mathematica.

	$0<\lambda\neq\lambda_i$ set $\nu=\sqrt[4]{\lambda}$	$\lambda=0$	$\lambda<0$ set $\mu=\sqrt[4]{-\frac{1}{4}\lambda}$
α_λ	$\frac{\nu^3}{8}\left(\frac{1}{\sin\nu}-\frac{1}{\sinh\nu}\right)\ \frac{\nu}{2\sin\nu}$	0	$\mu^3\frac{\cos\mu\sinh\mu-\cosh\mu\sin\mu}{\cosh 2\mu-\cos 2\mu}$
β_λ	$-\frac{\nu}{4}\left(\frac{1}{\sin\nu}+\frac{1}{\sinh\nu}\right)$	$-\frac{1}{2}$	$-\mu\frac{\sinh\mu\cos\mu+\cosh\mu\sin\mu}{\cosh 2\mu-\cos 2\mu}$
γ_λ	$-\frac{1}{2\nu}\ \left(\frac{1}{\sin\nu}-\frac{1}{\sinh\nu}\right)$	$\frac{1}{6}$	$\frac{\cosh\mu\sin\mu-\cos\mu\sinh\mu}{\mu(\cosh 2\mu-\cos 2\mu)}$

Now $g_\lambda(x,y)$ can be represented as follows:

(45)
$$g_\lambda(x,y)=\begin{cases}\begin{array}{l}\alpha_\lambda\ \phi(\lambda;x)\,\phi(\lambda;1-y)+\beta_\lambda\ \psi(\lambda;x)\,\phi(\lambda;1-y)\\ +\ \beta_\lambda\ \phi(\lambda;x)\,\phi'(\lambda;1-y)+\gamma_\lambda\ \psi(\lambda;x)\,\psi(\lambda;1-y)\end{array} & \text{if } 0\le x\le y\le 1,\\ \begin{array}{l}\alpha_\lambda\ \phi(\lambda;y)\,\phi(\lambda;1-x)+\beta_\lambda\ \psi(\lambda;y)\,\phi(\lambda;1-x)\\ +\ \beta_\lambda\ \phi(\lambda;y)\,\psi(\lambda;1-x)+\gamma_\lambda\ \psi(\lambda;y)\,\psi(\lambda;1-x)\end{array} & \text{if } 0\le y<x\le 1,\end{cases}$$

and the constants α_λ, β_λ, and γ_λ are suitably chosen in order to accommodate the boundary values in 1 and the continuity requirements of g_λ. Some tedious calculations lead to the coefficients in Table 4.

For $\lambda\ge 0$ formula (45) can be simplified to

$$(46)\qquad \begin{array}{ll} g_\lambda(x,y)=\frac{\sin(x\,\nu)\,\sin(\nu(1-y))}{2\,\nu^3\sin\nu}-\frac{\sinh(x\,\nu)\,\sinh(\nu(1-y))}{2\,\nu^3\sinh\nu} & \text{if } 0\le x\le y\le 1,\\ g_0(x,y)=\frac{1}{6}x\,(1-y)-\frac{1}{6}x\,(1-y)^3-\frac{1}{6}x^3\,(1-y) & \text{if } 0\le x\le y\le 1.\end{array}$$

5.2.2. Hinged rectangular grid with aligned fibers. In this section it will be convenient to use (x_1,x_2) instead of (x,y). An investigation of positivity preserving properties for the *hinged rectangular grid* that is *aligned with the Cartesian axes* seems to be difficult. The eigenfunctions are

$$\Phi_{ij}\,(x_1,x_2)=\frac{1}{\sqrt{ab}}\varphi_i\left(a^{-1}x_1\right)\varphi_j\left(b^{-1}x_2\right)$$

with $\varphi_i\,(t)=\sqrt{2}\sin(i\pi t)$. Recall that the first eigenfunction is of fixed sign and has multiplicity one. Using these eigenfunctions and the Green function g_λ from (45) or (46), the solution of

$$(47)\qquad \left\{\begin{array}{ll}\left(\frac{\partial^4}{\partial x_1^4}+\frac{\partial^4}{\partial x_2^4}\right)u=\lambda u+f & \text{in } R,\\ u=\Delta u=0 & \text{on } \partial R,\end{array}\right.$$

can be written as

$$\begin{aligned}u(x)&=\sum_{i,j=1}^{\infty}\frac{1}{a^{-4}\lambda_i+b^{-4}\lambda_j-\lambda}\left\langle\Phi_{ij},f\right\rangle\Phi_{ij}(x)\\ &=\frac{1}{\sqrt{ab}}\sum_{j=1}^{\infty}\ \left\langle\varphi_j\left(\frac{\circ}{b}\right),\int_{s=0}^{a}g_{b^{-4}\lambda_j-\lambda}\,(x_1,s)\,f\,(s,\circ)\,ds\right\rangle_{(0,b)}\varphi_j\left(\frac{x_2}{b}\right).\end{aligned}$$

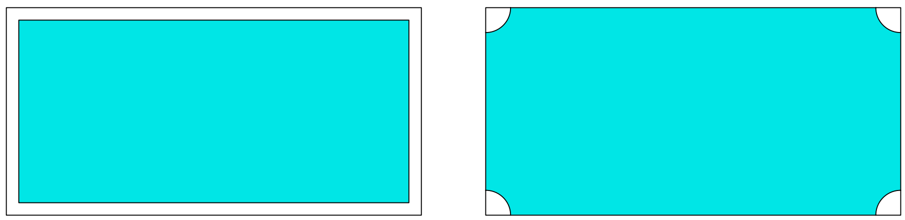


FIG. 4. *The sets* R_ε *and* $R \setminus (C + B^\varepsilon)$ *from Lemma* 5.1.

An inspection of the series representation above suggests that for nonnegative and nontrivial f, for $\lambda < \Lambda_{11}$ and λ close to Λ_{11} the coefficient in front of Φ_{11} becomes very large and positive. This suggests that the first term in the series determines the sign of u. But estimating the remainder of the series in terms of $\Phi_{1,1}$ turns out to be a hard technical problem.

In order to verify that problem (47) is positivity preserving at least for λ in some interval $[\Lambda_{11} - \gamma, \Lambda_{11})$ it suffices to show that the solution of (47) with $f = \delta_y$ is positive for every $y \in R$, where δ_y is the delta function at y.

Since the first eigenfunction is strictly positive in the interior we may prove the following result, in which we use this notation for a domain Ω:

- the ε-interior: $A_\varepsilon = \{x \in A; d(x, \partial\Omega) > \varepsilon\}$,
- the ε-neighborhood: $A + B^\varepsilon = \{x \in \Omega; d(x, \partial A) < \varepsilon\}$.

LEMMA 5.1. *Let* u^λ *be the solution of* (47)*. For every* $\varepsilon > 0$ *there is a positive* $\gamma > 0$ *such that for* $\lambda \in [\Lambda_{11} - \gamma, \Lambda_{11})$ *and* $f \geq 0$ *the following two statements hold (here* C *denotes the set of corner points):*

- *if* support $f \in R_\varepsilon$, *then* $u^\lambda(x) \geq 0$ *for all* $x \in R \setminus (C + B^\varepsilon)$,
- *if* support $f \in R \setminus (C + B^\varepsilon)$, *then* $u^\lambda(x) \geq 0$ *for all* $x \in R_\varepsilon$.

See Figure 4.

Proof. It is sufficient to show such a result for $f = \delta_y$, the delta function, with $y \in R_\varepsilon$. Formally we have

$$\delta_y(\cdot) = \sum_{i,j=1}^{\infty} \Phi_{ij}(y)\, \Phi_{ij}(\cdot),$$

but since $\delta_y \notin L^2(R)$ this series does not converge. The distributional solution $u^{y,\lambda}$ of (47) with $f = \delta_y$, that is,

$$u^{y,\lambda}(\cdot) = \sum_{i,j=1}^{\infty} \frac{\Phi_{ij}(y)}{\Lambda_{ij} - \lambda} \Phi_{ij}(\cdot), \tag{48}$$

lies in $L^2(R)$ since its coefficients are in ℓ_2 :

$$\sum_{i,j=1}^{\infty} \left(\frac{\Phi_{ij}(y)}{\Lambda_{ij} - \lambda} \right)^2 \leq \sum_{i,j=1}^{\infty} \left(\frac{1}{\pi^4 \left(\frac{i^4}{a^4} + \frac{j^4}{b^4} \right) - \lambda} \right)^2 < \infty.$$

We even find for $\alpha + \beta < 3$ that

$$\left\{ i^\alpha j^\beta \frac{\Phi_{ij}(y)}{\Lambda_{ij} - \lambda} \right\} \in \ell_2,$$

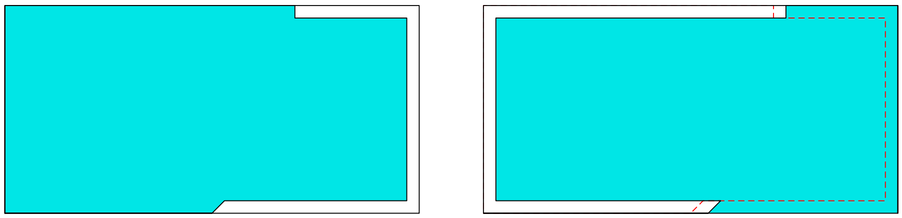


FIG. 5. *The sets* supp$f$ *and* $R_\varepsilon \cup (R \setminus (\text{supp} f + B_\varepsilon))$ *from Lemma* 5.2.

and hence $u^{y,\lambda} \in W^{3-t,2}(R)$ for all $t > 0$.

We split $u^{y,\lambda} = u_1^{y,\lambda} + u_2^{y,\lambda}$, where

$$u_1^{y,\lambda}(\cdot) = \frac{\Phi_{11}(y)}{\Lambda_{11} - \lambda} \Phi_{11}(\cdot).$$

By our assumption we have

$$|\Phi_{ij}(y)| \le \frac{c}{\varepsilon} \Phi_{11}(y) \qquad \text{and} \qquad |\Phi_{ij}(x)| \le c \max(i,j)\, \Phi_{11}(x).$$

This implies that we find

$$\begin{aligned} \left| u_2^{y,\lambda}(x) \right| &\le \sum_{\substack{i,j=1 \\ (i,j) \neq (1,1)}}^{\infty} \left| \frac{\Phi_{ij}(y)}{\Lambda_{ij} - \lambda} \Phi_{ij}(x) \right| \\ &\le \frac{c^3}{\varepsilon^2} \Phi_{11}(x)\, \Phi_{11}(y) \sum_{\substack{i,j=1 \\ (i,j) \neq (1,1)}}^{\infty} \frac{\max(i,j)}{\pi^4 \left(\frac{i^4}{a^4} + \frac{j^4}{b^4} \right) - \lambda}. \end{aligned}$$

Since Λ_{12} and Λ_{21} are greater than Λ_{11}, a straightforward computation shows that the last sum is bounded uniformly with respect to $\lambda < \Lambda_{11}$ by a constant $\gamma = C(a,b)\, \varepsilon^{-2}$. For $\lambda \in [\Lambda_{11} - \gamma, \Lambda_{11})$ we find

$$\left| u_2^{y,\lambda}(x) \right| \le \gamma \Phi_{11}(x)\, \Phi_{11}(y) \le u_1^{y,\lambda}(x)$$

and hence that $u^{y,\lambda}(x) > 0$. □

LEMMA 5.2. *For every $\varepsilon > 0$ there is a $\gamma > 0$ such that if $\lambda \in [\Lambda_{11} - \gamma, \Lambda_{11})$ and $f \ge 0$, then the solution of* (47) *satisfies $u^\lambda(x) \ge 0$ for all $x \in R_\varepsilon \cup (R \setminus (\text{supp} f + B_\varepsilon))$; see Figure* 5.

Proof. If supp$f \in R_\varepsilon$, then the previous lemma yields that $u^\lambda(x) \ge 0$ except near the corners C. By Proposition A.1 in the appendix and using duality, we find that $\left\| u_2^\lambda \right\|_{W^{28,2}(C + B_{\varepsilon/2})} \le c(\varepsilon) \|f\|_{W^{-4,2}(\Omega)}$. Let us denote by $d_h(x)$ and $d_v(x)$ the distance of $x \in R$ to the horizontal and vertical parts of its boundary, and by $\langle v, f \rangle$ the $L^2(R)$ product, when applicable. Then one continues with the imbedding of $W_0^{4,2}(\Omega)$ in $C^2(\bar{\Omega}) \cap C_0^1(\bar{\Omega})$, through

$$\begin{aligned} \|f\|_{W^{-4,2}(\Omega)} &= \sup \left\{ \langle v, f \rangle ; v \in W_0^{4,2}(\Omega) \text{ with } \|v\|_{W^{4,2}(\Omega)} \le 1 \right\} \\ &\le c \sup \left\{ \langle v, f \rangle ; v \in C^2(\bar{\Omega}) \cap C_0^1(\bar{\Omega}) \text{ with } \|v\|_{C^2(\bar{\Omega})} \le 1 \right\} \\ &\le c \sup \left\{ \langle v, f \rangle ; |v(x)| \le d_h(x) d_v(x) | \right\} \\ &\le c' \langle \Phi_{11}, f \rangle. \end{aligned}$$

The last inequality is due to the fact that Φ_{11} can be bounded above and below by multiples of $d_h(x)d_v(x)$.

Similarly, again with an imbedding, we find for the function $u_2^\lambda \in C^2(\bar{\Omega}) \cap C_0^1(\bar{\Omega})$ and for $x \in C + B_{\varepsilon/2}$ that

$$\begin{aligned} u_2^\lambda(x) &\leq c_1 \left\| u_2^\lambda \right\|_{C^2(C+B_{\varepsilon/2})} \Phi_{11}(x) \leq c_2 \left\| u_2^\lambda \right\|_{W^{28,2}\left(C+B_{\varepsilon/2}\right)} \Phi_{11}(x) \\ &\leq c(\varepsilon) \left\| f \right\|_{W^{-3,2}(\Omega)} \Phi_{11}(x) \leq c'(\varepsilon) \left\langle \Phi_{11}, f \right\rangle \Phi_{11}(x). \end{aligned}$$

Since $u_1^\lambda(x) = (\Lambda_{11} - \lambda)^{-1} \langle \Phi_{11}, f \rangle \Phi_{11}(x)$ we find $u^\lambda(x) > 0$ near the corners for $\Lambda_{11} - \lambda$ chosen sufficiently small. A similar proof can be followed for the remaining claim. $\square$

Let us summarize our results in terms of positivity for the Green function $u^{y,\lambda}$ from (48) belonging to the *hinged rectangular grid*.

COROLLARY 5.3. *For every $\varepsilon > 0$ there is a $\gamma(\varepsilon) > 0$ such that $u^{y,\lambda}(x) \geq 0$ for all $x \in R$, all $y \in R_{2\varepsilon}$, and all $\lambda \in [\Lambda_{11} - \gamma(\varepsilon), \Lambda_{11})$.*

Proof. We approximate $\delta_y(\cdot)$ in $\mathcal{D}'(\Omega)$ by a sequence of smooth f_n with support in $B_\varepsilon(y)$, and note that the corresponding solutions $u_n^{y,\lambda}(x)$ of (47) are nonnegative for all $x \in R$ and all $y \in R_{2\varepsilon}$ due to Lemma 5.2. Then we send $n \to \infty$. Since f_n converges in $W^{-1,2}(R)$, the sequence u_n converges pointwise. $\square$

Notice that when sending ε to zero it is conceivable (although it seems unlikely) that $\gamma(\varepsilon) \to 0$. In that case, as $\varepsilon_n \to 0$ there exist sequences $\lambda_n < \Lambda_{11}$ with $\lambda_n \to \Lambda_{11}$, $y_n \in R \setminus R_{2\varepsilon_n}$ with $y_n \to y_0 \in \partial R$, and $x_n \to x_0 \in R$ such that

$$z_n := u^{y_n,\lambda_n}(x_n) < 0 \quad \text{ for all } n \in \mathbb{N}.$$

At present we are unable to derive a contradiction from this.

We will end this section by a another nonuniform positivity result near Λ_{11} by using the fact that the projection on the first eigenfunction will dominate near Λ_{11}. We proceed as for the nonuniform version of the antimaximum principle in [7] to obtain the following nonuniform result.

PROPOSITION 5.4. *For all $f \in L^2(R)$ with $f \geq 0$ there exists $\lambda_f < \Lambda_{11}$ such that for $\lambda \in [\lambda_f, \Lambda_{11})$ the solution u_λ of (47) satisfies $u_\lambda \geq 0$.*

Proof. We will adjust the arguments in [7] for the present situation. Let $\mathcal{L}$: $W^{4,2}(R) \cap W_0^{2,2}(R) \to L^2(R)$ be the operator that corresponds to (47). Fix P_0 to be the projection on the first eigenfunction, that is, $P_0 f = \langle \Phi_{11}, f \rangle_R \ \Phi_{11}$, and set $\tilde{\Lambda} \in (\Lambda_{11}, \min(\Lambda_{12}, \Lambda_{21}))$. Then, using our regularity result for (47) as in [7], we find that there exists a constant C such that for all $\lambda \in \left[0, \tilde{\Lambda}\right]$ the following holds:

$$\left\| (\mathcal{L} - \lambda)^{-1} (I - P_0) f \right\|_{W^{4,2}(R)} \leq C \left\| f \right\|_{L^2(R)}.$$

Since the domain R satisfies a uniform interior cone condition we find by [18, Theorem 7.26] that $W^{4,2}(R)$ is imbedded in $C^{2,\alpha}(\bar{\Omega})$ for any $\alpha \in (0,1)$. Since

$$(\mathcal{L} - \lambda)^{-1} (I - P_0) f \in W_0^{2,2}(\Omega)$$

we find that $u \in C_0(\bar{\Omega})$ and hence that

$$\left\| \frac{(\mathcal{L} - \lambda)^{-1} (I - P_0) f}{\Phi_{11}} \right\|_\infty \leq C' \left\| (\mathcal{L} - \lambda)^{-1} (I - P_0) f \right\|_{W^{4,2}(R)}.$$

FIG. 6. *Top to bottom: a hinged plate, a hinged grid with rectangular fibers, and a hinged grid with diagonal fibers. The arrow denotes the location of the pointed force, and the red (dark) part represents the part of the grid with a negative deviation.*

The solution u_λ of (47) can be written as

$$\begin{aligned} u_\lambda(x) &= \frac{\langle \Phi_{11}, f\rangle_R}{\Lambda_{11} - \lambda}\Phi_{11}(x) + \left((\mathcal{L} - \lambda)^{-1} (I - P_0) f \right)(x) \\ &\geq \left(\frac{\langle \Phi_{11}, f\rangle_R}{\Lambda_{11} - \lambda} - C'' \|f\|_{L^2(R)} \right) \Phi_{11}(x), \end{aligned}$$

which is positive for $0 \leq \Lambda_{11} - \lambda$ sufficiently small. □

5.2.3. Hinged rectangular grid with diagonal fibers. The positivity question is much simpler to decide if the grid runs diagonally. For the *diagonally hinged grid on the rectangle* R as in (17)–(20),

$$\text{(49)} \qquad \begin{cases} \frac{1}{2}u_{xxxx} + 3u_{xxyy} + \frac{1}{2}u_{yyyy} = \lambda u + f & \text{in } R, \\ u = u_{xx} = 0 & \text{on } \{0,a\} \times [0,b], \\ u = u_{yy} = 0 & \text{on } [0,a] \times \{0,b\}, \end{cases}$$

one may decouple the fourth order equation (19) (or (49) with $\lambda = 0$) into a system of two second order equations by using (20).

Since the boundary conditions decouple nicely with the two second order operators, one may use the substitution $v := -u_{xx} - (3 + 2\sqrt{2})u_{yy}$ and two iterations of the standard maximum principle for second order differential operators to find that (49) is positivity preserving for $\lambda = 0$.

Going back to the fourth order problem, one has a strongly positive and compact solution operator that maps $f \in C(\overline{\Omega})$ to $u \in C(\overline{\Omega})$. From the Krein–Rutman theorem one finds that there exists a first eigenvalue, and this eigenvalue corresponds to an eigenfunction of fixed sign. But then one can show the following as in [35].

PROPOSITION 5.5. *For* $\lambda \in \left[0, \pi^4 \left(\frac{1}{2}a^{-4} + 3a^{-2}b^{-2} + \frac{1}{2}b^{-4}\right)\right)$ *the problem* (49) *is positivity preserving.*

The upper bound for λ is the first eigenvalue Γ_{11} given in (28).

5.2.4. Numerical comparison for hinged rectangles. For the hinged rectangular plate and grids we obtained the numerical result shown in Figure 6 by a finite difference method.

5.3. Positivity under clamped boundary conditions.

Table 5
These values for the coefficients in (52) *were obtained using Mathematica.*

	$0<\lambda\neq\lambda_i$ set $\nu=\sqrt[4]{\lambda}$	$\lambda=0$	$\lambda<0$ set $\mu=\sqrt[4]{-\frac{1}{4}\lambda}$
α_λ	$\frac{\nu^3(\sinh\nu+\sin\nu)}{4-4\cosh\nu\cos\nu}$	3	$\frac{2\mu^3(\cos\mu\sinh\mu+\cosh\mu\sin\mu)}{\cosh 2\mu+\cos 2\mu-2}$
β_λ	$\frac{\nu^2(\cos\nu-\cosh\nu)}{4-4\cosh\nu\cos\nu}$	$-\frac{3}{2}$	$\frac{-2\mu^2\sinh\mu\sin\mu}{\cosh 2\mu+\cos 2\mu-2}$
γ_λ	$\frac{\nu(\sinh\nu-\sin\nu)}{4-4\cosh\nu\cos\nu}$	$\frac{1}{2}$	$\frac{\mu(\cosh\mu\sin\mu-\cos\mu\sinh\mu)}{\cosh 2\mu+\cos 2\mu-2}$

5.3.1. Clamped beam. What can be said about positivity preservation for the *clamped beam* (50)?

$$\left\{\begin{array}{ll} u_{xxxx}=\lambda u+f & \text{in } (0,1), \\ u=u_x=0 & \text{in } \{0,1\}. \end{array}\right. \tag{50}$$

This requires more effort. If λ is not an eigenvalue, there exists a Green function g_λ for the clamped beam problem (50) such that the solution can be represented as

$$u(x)=\int_0^1 g_\lambda(x,y)f(y)\,dy.$$

Let us define

$$\phi(\lambda;x)=\left\{\begin{array}{ll} \nu^{-3}\left(\sinh(\nu x)-\sin(\nu x)\right) & \text{if } \lambda>0, \\ \frac{1}{3}x^3 & \text{if } \lambda=0, \\ \frac{1}{2}\mu^{-3}\left(\cosh(\mu x)\sin(\mu x)-\sinh(\mu x)\cos(\mu x)\right) & \text{if } \lambda<0, \end{array}\right. \tag{51}$$

where $\nu=\sqrt[4]{\lambda}$ and $\mu=\sqrt[4]{-\frac{1}{4}\lambda}$. The functions $\phi(\lambda;\cdot)$ and $\frac{\partial}{\partial x}\phi(\lambda;\cdot)$ are two linearly independent solutions of the differential equation and the boundary conditions of (50) in the left end point 0. By the definition of the Green function it follows that

(52)
$$g_\lambda(x,y)=\left\{\begin{array}{ll} \begin{array}{l}\alpha_\lambda\ \phi(\lambda;x)\,\phi(\lambda;1-y)+\beta_\lambda\ \phi'(\lambda;x)\,\phi(\lambda;1-y)\\ +\ \beta_\lambda\ \phi(\lambda;x)\,\phi'(\lambda;1-y)+\gamma_\lambda\ \phi'(\lambda;x)\,\phi'(\lambda;1-y)\end{array} & \text{if } 0\le x\le y\le 1, \\ \begin{array}{l}\alpha_\lambda\ \phi(\lambda;y)\,\phi(\lambda;1-x)+\beta_\lambda\ \phi'(\lambda;y)\,\phi(\lambda;1-x)\\ +\ \beta_\lambda\ \phi(\lambda;y)\,\phi'(\lambda;1-x)+\gamma_\lambda\ \phi'(\lambda;y)\,\phi'(\lambda;1-x)\end{array} & \text{if } 0\le y<x\le 1, \end{array}\right.$$

with appropriate constants to accommodate the boundary values in 1 and the continuity requirements of g_λ. Some tedious calculations lead to the coefficients given in Table 5.

For $\lambda=0$ formula (52) can be simplified to

$$g_0(x,y)=\left\{\begin{array}{ll} \frac{1}{2}x^2(1-y)^2\left(y-x+\frac{2}{3}x(1-y)\right) & \text{if } 0\le x\le y\le 1, \\ \frac{1}{2}y^2(1-x)^2\left(x-y+\frac{2}{3}y(1-x)\right) & \text{if } 0\le y<x\le 1. \end{array}\right.$$

Problem (50) is positivity preserving if and only if the Green function is positive, and for g_0 this is now easily seen to be the case. Instead of directly computing for which λ the Green function g_λ is in fact positive, one may proceed through the results of Schröder in [34]. The Green function changes sign for some λ if and only if this λ is an eigenvalue of either (30) or of

$$\left\{ \begin{array}{c} \varphi'''' = \lambda\varphi \text{ in } (0,1)\,, \\ \varphi(0) = \varphi'(0) = \varphi'''(0) = 0 = \varphi(1). \end{array} \right. \tag{53}$$

The "first" solution of (53) is $g_\lambda(x,1)$ with $\lambda_c = -4\nu_0^4$, where ν_0 is the first positive zero of $\tanh\nu = \tan\nu$.

LEMMA 5.6. *Problem* (50) *is positivity preserving if and only if* $\lambda \in [\lambda_c, \lambda_1)$, *where*

- λ_1 *is the first eigenvalue of* (30)*, that is, the fourth power of the first positive solution of*

$$\cos\lambda = \frac{1}{\cosh\lambda},$$

- λ_c *is the "first" eigenvalue of* (53)*, that is, the first negative solution of*

$$\tan\sqrt[4]{-\frac{1}{4}\lambda} = \tanh\sqrt[4]{-\frac{1}{4}\lambda}. \tag{54}$$

The numerical approximations are $\lambda_1 \approx 4.7300$ and $\lambda_c \approx -950.884$. Notice that this is the same λ_c as in (44) for problem (43).

Proof. The arguments are similar to the ones in [23] and reflect the ideas from [34].

Direct inspection shows that g_0 is strictly positive. To study the case of positive λ, notice that (50) can be rewritten as $(I - \lambda L^{-1})u = f$, where $Lu = u_{xxxx}$, so that by a Neumann series argument $u = \sum_{k=1}^{\infty}(\lambda L^{-1})^k f$ converges and is positive for all $\lambda \in [0, \lambda_1)$. For $\lambda = \lambda_1$ no solution exists when $f = \varphi_1$. For $\lambda > \lambda_1$ and $f = \varphi_1$ the solution is $u = (\lambda_1 - \lambda)^{-1}\varphi_1$, and this is negative.

For $\lambda < 0$ one finds from (52)–(51) and the coefficients in Table 4 that $\lambda \mapsto g_\lambda(x,y)$ is continuous for $\lambda \le 0$ in almost every sense. Let $\lambda_c < 0$ be the first number after which positivity fails. Suppose that for a fixed $y \in (0,1)$ the value of $g_{\lambda_c}(x,y)$ is nonnegative but equals 0 for some $x_y \in (0,1)$. And suppose w.l.o.g. that $x_y \le y$. Then $g_{\lambda_c}(x_y,y) = \frac{\partial}{\partial x} g_{\lambda_c}(x_y,y) = \frac{\partial}{\partial x} g_{\lambda_c}(0,y) = \frac{\partial}{\partial x} g_{\lambda_c}(0,y) = 0$ and we have found an eigenfunction scaled to $[0, x_y]$, a contradiction. Hence $x_y = 0$. Using the symmetry $g_\lambda(x,y) = g_\lambda(y,x)$, we may assume that y is at the boundary, say $y = 1$. We may repeat the argument above for $\tilde{g}_{\lambda_c}$ defined by $\tilde{g}_\lambda(x) = \lim_{y\uparrow 1}(1-y)^{-2} g_\lambda(x,y)$, which is a nontrivial function. Again if $\tilde{g}(x_1) = 0$ for some $x_1 \in (0,1)$, we find an eigenfunction by scaling on $[0, x_1]$. Since $\tilde{g}'(1) < 0 = \tilde{g}(1)$ it remains that $x_1 = 0$. One finds that $\tilde{g}$ is an eigenfunction of (53). The first eigenfunction of that eigenvalue problem is

$$\psi_1(x) = \cosh(\mu x)\sin(\mu x) - \sinh(\mu x)\cos(\mu x)$$

with μ the first positive root of $\cosh\mu\sin\mu = \sinh\mu\cos\mu$ and $\lambda_c = 4\mu^2$. This can be rephrased to (54). For $\lambda < \lambda_c$ one finds that $\tilde{g}_\lambda$ is sign changing, implying that for y near 1 the function $x \mapsto g_\lambda(x,y)$ is sign changing. $\square$

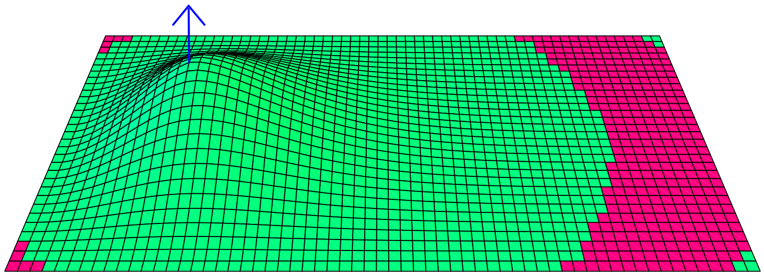


FIG. 7. *Numerical simulation of a clamped rectangularly aligned grid;* (55) *with* $\mu = 0$ *and a point source* f.

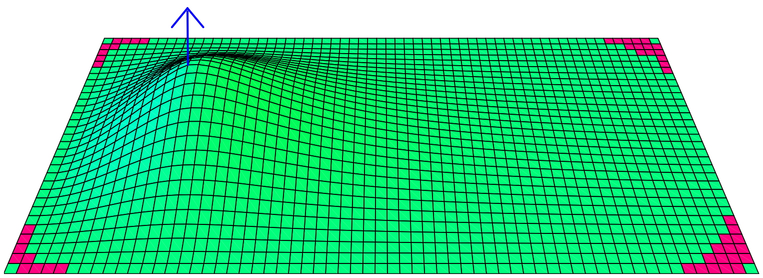


FIG. 8. *Numerical simulation of a clamped diagonally aligned grid.*

5.3.2. Clamped rectangular grid with aligned fibers. In this section we investigate the problem

$$\left\{ \begin{array}{ll} u_{xxxx} + u_{yyyy} = \lambda u + f & \text{in } R, \\ u = |\nabla u| = 0 & \text{on } \partial R. \end{array} \right. \tag{55}$$

Numerical calculations suggest that for $\lambda = 0$ a point load $f = \delta_y(\cdot)$ can lead to a sign changing solution; see Figure 7, in which the sign of u is color coded. This behavior is also known and recorded in [9] for isotropic rectangular plates, whose deformation solves $\Delta^2 u = f$ instead.

However, since the first eigenfunction is positive, by using the eigenfunction expansion one finds the following solution formula for (55):

$$u(x, y) = \sum_{i,j=1}^{\infty} \frac{1}{\Lambda_{ij} - \lambda} \langle \Phi_{ij}, f \rangle_R \; \Phi_{ij}(x, y).$$

As for the hinged plate one might hope that for λ near Λ_{11} the projection on the first eigenfunction will dominate the sign. But to find such a result we would need a C^4 estimate near corner points, which, unfortunately, we do not have at our disposal.

5.3.3. Clamped diagonal grid. Since we do not know if the first eigenfunction is of fixed sign for this grid we can only give some numerical evidence. With the same point source and domain as in rectangularly aligned grid from Figure 7, the area where the solutions change sign seems to be much smaller for the diagonally aligned grid; see Figure 8.

5.3.4. Numerical comparison for clamped rectangles. Duffin's famous counterexample in [13] for the conjecture of Boggio and Hadamard (the clamped plate problem on convex domains is positivity preserving) uses a long thin rectan-

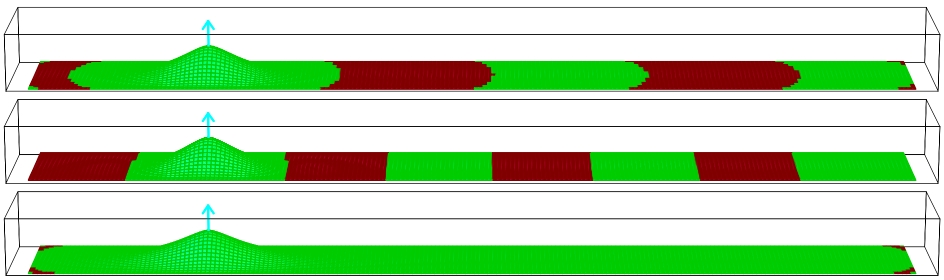


FIG. 9. *Top to bottom: a clamped plate, a clamped grid with rectangular fibers, and a clamped grid with diagonal fibers. The arrow denotes the location of the pointed force and the red (dark) part represents the part of the grid with a negative deviation.*

TABLE 6
Overview for rectangular plates and grids.

		Positive eigenfunction	Positivity preserving
Hinged	Plate	$\Phi_1 > 0$	for $\lambda \in [0, \Lambda_1)$
	Grid aligned with sides	$\Phi_1 > 0$	conditionally near Λ_1
	Grid with diagonal fibers	$\Phi_1 > 0$	for $\lambda \in [0, \Lambda_1)$
Clamped	Plate	Φ_1 changes sign	no
	Grid aligned with sides	$\Phi_1 > 0$	conditionally near Λ_1
	Grid with diagonal fibers	?	?

gle. In Figure 9 we present numerical results for long clamped rectangular plates and grids. Rather surprisingly, the numerical result for a long thin rectangle with a diagonal fabric hardly shows any sign change.

The numerical illustrations have been obtained using a finite difference method.

6. Summary for rectangular grids. We set out to study positivity for rectangular grids with aligned and with diagonal fibers. An overview of the results we obtained for those problems can be found in Table 6. For the sake of comparison we include the known results for the rectangular plate.

Numerics for the clamped plate with diagonal fabric suggest that the first question mark in the table above should be answered affirmatively; the second question mark might have a positive answer for λ near Λ_1. Of course "near Λ_1" always means in a left neighborhood of Λ_1.

Appendix. Nonlocal smoothness. The standard regularity statement for $2m$-th order elliptic problems is usually a statement of the form that $f \in W^{k,p}(\Omega)$ implies $u \in W^{k+2m,p}(\Omega)$ or $f \in C^{k,\gamma}(\bar{\Omega})$ implies $u \in C^{k+2m,\gamma}(\bar{\Omega})$. Such a maximal regularity result is optimal. However, for a function $f \in L^p(\Omega)$ which has its support in $\Omega' \subset \Omega$ one may show that the corresponding solution is smooth outside of Ω'. Although this result is well known, we are not aware of any reference. So allow us to formulate a corresponding statement.

Consider a regular elliptic problem with L of order $2m$ and Ω a domain in $\mathbb{R}^n$:

$$\begin{cases} Lu = f & \text{in } \Omega, \\ B_i u = 0 & \text{on } \partial\Omega \quad \text{for } i = 0, \ldots, m. \end{cases} \tag{56}$$

PROPOSITION A.1. *Let Ω_1, Ω_2 be two disjoint subdomains of Ω having a positive distance r, that is, $r := \inf\{|x-y|\,; x \in \Omega_1, y \in \Omega_2\} > 0$. Suppose that there exists $c > 0$ such that for all $k \in \{0,\dots,\kappa\}$ and all $f \in W^{k,2}(\Omega)$ there is a solution $u \in W^{2m+k,2}(\Omega)$ of (56) with*

$$\|u\|_{W^{2m+k,2}(\Omega)} \le c\,\|f\|_{W^{k,2}(\Omega)}\,; \tag{57}$$

then there exists $C(c,\kappa,r)$ such that for all $f \in L_2(\Omega)$ with support$f$ $\subset \Omega_1$ *the following holds true:*

$$\|u\|_{W^{2m+\kappa,2}(\Omega_2)} \le C(c,\kappa,r)\,\|f\|_{L_2(\Omega_1)}\,. \tag{58}$$

Proof. We will prove this by induction. For $k = 0$ the estimate (58) follows from (57) and the fact that supp $f \subset \Omega_1$. Next we do the induction from k to $k+1$ and suppose that $\|u\|_{W^{2m+k,2}(\Omega_2)} \le C(c,k,r)\,\|f\|_{L_2(\Omega_1)}$ for some $k \ge 0$. One may construct a cut-off function χ such that for some $c_1 \in \mathbb{R}^+$

1. $\chi \in C^\infty(\bar\Omega)$ with $\chi_{|\Omega_1} = 0$ and $\chi_{|\Omega_2} = 1$;
2. $\bar\Omega_2 \Subset \text{support}\chi \Subset \bar\Omega\backslash\Omega_1$;
3. $\|\chi\|_{C^i(\bar\Omega)} \le c_1 r^{-i}$ for $i \in \{0,\dots,k\}$.

Note that $L(\chi u) = \chi L u + \text{l.o.t.} = 0 + \text{l.o.t.}$ and that χu satisfies the boundary conditions from (56). Since the right-hand-side l.o.t. lies in $W^{k+1,2}(\Omega)$ we find $\chi u \in W^{2m+k+1,2}(\Omega)$. Moreover,

$$\begin{aligned} &\|u\|_{W^{2m+k+1,2}(\Omega_2)} \le \|\chi u\|_{W^{2m+k+1,2}(\Omega)} \le c_1\,\|l.o.t.\|_{W^{k+1,2}(\Omega)} \\ &= c_1\,\|l.o.t.\|_{W^{k+1,2}(\text{support}\chi)} \le c(r)\ \|u\|_{W^{2m+k,2}(\tilde\Omega_2)} \le C'(c,k,r/2)\,\|f\|_{L_2(\Omega_1)}\,. \end{aligned}$$

Here $\tilde\Omega_2$ can be chosen so that $d(\Omega_1,\tilde{\Omega_2}) < r/(2k)$. ∎

Acknowledgments. Special thanks go to W. Jäger for bringing [17] to our attention and also to P. Seidel for her help in locating some of the older literature.

REFERENCES

[1] M. S. Ashbaugh and R. S. Laugesen, *Fundamental tones and buckling loads of clamped plates*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 23 (1996), pp. 383–402.

[2] T. Boggio, *Sulle funzioni delle piastre elastiche incastrata*, Rend. Accad. Lincei, 10 (1901), pp. 197–205.

[3] H. T. Banks, D. Cioranescu, and R. A. Rebnord, *Homogenization models for two-dimensional grid structures*, Asymptot. Anal., 11 (1995), pp. 107–130.

[4] M. Belloni and B. Kawohl, *The pseudo-p-Laplace eigenvalue problem and viscosity solutions as $p \to \infty$*, ESAIM Control Optim. Calc. Var., 10 (2004), pp. 28–52.

[5] J. Bernoulli, II, *Essai théoretique sur les vibrations des plaques élastiques, rectangulaires et libres*, Nova Acta Acad. Sci. Imperialis Petropolitanae, V (1789), pp. 197–219.

[6] I. Birindelli, E. Mitidieri, and G. Sweers, *Existence of the principal eigenfunction for cooperative elliptic systems in a general domain*, Differential Equations, 35 (1999), pp. 326–334.

[7] Ph. Clément and G. Sweers, *Uniform anti-maximum principles*, J. Differential Equations, 164 (2000), pp. 118–154.

[8] C. V. Coffman, *On the structure of solutions to $\Delta^2 u = \lambda u$ which satisfy the clamped plate conditions on a right angle*, SIAM J. Math. Anal., 13 (1982), pp. 746–757.

[9] C. V. Coffman and R. J. Duffin, *On the structure of biharmonic functions satisfying the clamped plate conditions on a right angle*, Adv. Appl. Math., 1 (1980), pp. 373–389.

[10] R. Courant and D. Hilbert, *Methoden der Mathematischen Physik*, Springer Verlag, Berlin, 1993.

[11] A. Dall'Acqua and G. Sweers, *The clamped-plate equation on the limaçon*, Ann. Mat. Pura Appl., 184 (2005), 361–374.
[12] A. Dall'Acqua and G. Sweers, *On domains for which the clamped plate system is positivity preserving*, in Partial Differential Equations and Inverse Problems, Contemp. Math. 362, AMS, Providence, RI, 2004, pp. 133–144.
[13] R. J. Duffin, *On a question of Hadamard concerning super-biharmonic functions*, J. Math. Phys., 27 (1949), pp. 253–258.
[14] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, AMS, Providence, RI, 1998.
[15] G. Fichera, *Numerical and Quantitative Analysis*, Pitman, London, 1978.
[16] K. Friedrichs, *Die Randwert- und Eigenwertprobleme aus der Theorie der elastischen Platte*, Math. Ann., 98 (1928), pp. 205–247.
[17] H. Finckh, *Numerical simulation of mechanical behaviour of dynamically stressed fencing mask*, in High Performance Computing in Science and Engineering '98, E. Krause and W. Jäger, eds., Springer-Verlag, Berlin, 1999, pp. 387–395.
[18] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Grundlehren Math. Wiss., 224, Springer-Verlag, Berlin, 1983.
[19] H.-Ch. Grunau and G. Sweers, *Positivity for perturbations of polyharmonic operators with Dirichlet boundary conditions in two dimensions*, Math. Nachr., 179 (1996), pp. 89–102.
[20] J. Hersch, *Lower bounds for membrane eigenvalues by anisotropic auxiliary problems*, Appl. Anal., 3 (1973), pp. 241–245.
[21] P. Jentzsch, *Über Integralgleichungen mit positivem Kern*, J. Reine Angew. Math., 141 (1912), pp. 235–244.
[22] J. Kadlec, *The regularity of the solution of the Poisson problem in a domain whose boundary is similar to that of a convex domain*, Czechoslovak Math. J., 14 (1964), pp. 386–393 (in Russian).
[23] B. Kawohl and G. Sweers, *On 'anti'–eigenvalues for elliptic systems and a question of McKenna and Walter*, Indiana Univ. Math. J., 51 (2002), pp. 1023–1040.
[24] G. Kirchhoff, *Über das Gleichgewicht und die Bewegung einer elastischen Scheibe*, J. Reine Angew. Math., 40 (1850), pp. 51–88.
[25] J. R. Kuttler and V. G. Sigillito, *Estimating eigenvalues with a posteriori/a priori inequalities*, Pitman Res. Notes Math. 135, Pitman, Boston, 1985.
[26] S. G. Lekhnitskii, *Theory of Elasticity of an Anisotropic Elastic Body*, transl. P. Fern, J. J. Brandstatter, ed., Holden-Day, San Francisco, 1963.
[27] J.-L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, Vol. I, Springer Grundlehren 181, Springer-Verlag, Heidelberg, 1972.
[28] A. E. H. Love, *A Treatise on the Mathematical Theory of Elasticity*, Cambridge University Press, Cambridge, UK, 1934.
[29] J. Mossakowski, *Singular solutions in the theory of orthotropic plates*, Arch. Mech., 6 (1954), pp. 413–432.
[30] D. Ó Mathúna, *Jacob II Bernoulli and the problem of the vibrating plate*, in Entre mécanique et Architecture/Between Mechanics and Architecture, P. Radelet-de-Grave and E. Benvenuto, eds., Birkhäuser-Verlag, Basel, 1995, pp. 165–177.
[31] J. L. Nowinski, *Bilateral bounds for the solution of a generalized biharmonic boundary value problem*, SIAM J. Appl. Math., 39 (1980), pp. 193–200.
[32] L. E. Persson, and T. Strömberg, *Green's method applied to the plate equation in mechanics*, Comment. Math. Prace Mat., 33 (1993), pp. 119–133.
[33] G. Philippin, *Évaluations par défaut de l'énergie potentielle et des valeurs propres pour certains problèmes relatifs aux plaques élastiques*, J. Anal. Math., 28 (1975), pp. 138–170.
[34] J. Schröder, *Operator Inequalities*, Math. Sci. Engrg. 147, Academic Press, New York, 1980.
[35] G. Sweers, *Strong positivity in $C(\bar{\Omega})$ for elliptic systems*, Math. Z., 209 (1992), pp. 251–271.
[36] G. Sweers, *When is the first eigenfunction for the clamped plate equation of fixed sign?*, in Proceedings of the USA-Chile Workshop on Nonlinear Analysis (Viña del Mar-Valparaiso, 2000), Electron. J. Differential Equations Conf. 6, Southwest Texas State University, San Marcos, TX, 2001, pp. 285–296 (electronic).
[37] I. Szabó, *Geschichte der mechanischen Prinzipien und ihrer wichtigsten Anwendungen*, Wissenschaft und Kultur 32, Birkhäuser-Verlag, Basel, 1977.
[38] S. T. Timoshenko and J. M. Gere, *Theory of Elastic Stability*, Mc Graw-Hill, New York, 1961.
[39] H. J. Töpfer, *Über Eindeutigkeit und praktische Bedeutung der Lösungen von Randwertaufgaben in der Theorie der orthogonal-anisotropen Platten*, Z. Angew. Math. Mech., 42 (1962), pp. 142–162.
[40] P. Villaggio, *Mathematical Models for Elastic Structures*, Cambridge University Press, Cambridge, UK, 1997.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1854–1889

QUASI-NEUTRAL LIMIT OF THE DRIFT DIFFUSION MODELS FOR SEMICONDUCTORS: THE CASE OF GENERAL SIGN-CHANGING DOPING PROFILE*

SHU WANG†, ZHOUPING XIN‡, AND PETER A. MARKOWICH§

Abstract. In this paper the vanishing Debye length limit (space charge neutral limit) of bipolar time-dependent drift-diffusion models for semiconductors with p-n junctions (i.e., with a fixed bipolar background charge) is studied in one space dimension. For general sign-changing doping profiles, the quasi-neutral limit (zero-Debye-length limit) is justified rigorously in the spatial mean square norm uniformly in time. One main ingredient of our analysis is the construction of a more accurate approximate solution, which takes into account the effects of initial and boundary layers, by using multiple scaling matched asymptotic analysis. Another key point of the proof is the establishment of the structural stability of this approximate solution by an elaborate energy method which yields the uniform estimates with respect to the scaled Debye length.

Key words. quasi-neutral limit, drift-diffusion equations, multiple scaling asymptotic expansions, singular perturbation, classical energy methods, λ-weighted Liapunov-type functional

AMS subject classifications. 35B25, 35B40, 35K57

DOI. 10.1137/S0036141004440010

1. Introduction.

The scaled one-dimensional isothermal drift-diffusion model for semiconductors reads

$$n_t^\lambda = (n_x^\lambda - n^\lambda \Phi_x^\lambda)_x, \quad 0 < x < 1, \quad t > 0, \tag{1}$$

$$p_t^\lambda = (p_x^\lambda + p^\lambda \Phi_x^\lambda)_x, \quad 0 < x < 1, \quad t > 0, \tag{2}$$

$$\lambda^2 \Phi_{xx}^\lambda = n^\lambda - p^\lambda - D, \quad 0 < x < 1, \quad t > 0, \tag{3}$$

$$n_x^\lambda - n^\lambda \Phi_x^\lambda = p_x^\lambda + p^\lambda \Phi_x^\lambda = \Phi_x^\lambda = 0, \quad x = 0, 1, \quad t > 0, \tag{4}$$

$$n^\lambda(x, 0) = n_0^\lambda(x), \quad p^\lambda(x, 0) = p_0^\lambda(x), \quad 0 \le x \le 1. \tag{5}$$

The variables $n^\lambda, p^\lambda, \Phi^\lambda$ are the electron density, the hole density, and the electric potential, respectively. The constant λ is the scaled Debye length of the semiconductor device under consideration. $D = D(x)$ is the given function of space and models the doping profile (i.e., the preconcentration of electrons and holes). Because of the occurrence of p-n junctions in realistic semiconductor devices, the doping profile $D(x)$ typically changes its sign.

*Received by the editors January 18, 2004; accepted for publication (in revised form) February 21, 2005; published electronically March 3, 2006.
http://www.siam.org/journals/sima/37-6/44001.html

†College of Science, Beijing University of Technology, Ping Le Yuan 100, Chaoyang District, Beijing 100022, P. R. China (wangshu@bjut.edu.cn). The research of this author was supported by NSFC grant 10471009 and BSFC grant 1052001 of China, the Wittgenstein-Award 2000 of Peter A. Markowich, funded by the Austrian Science Fund FWF, the Zheng Ge Ru Foundation, and grants from RGC of HKSAR CUHK4040/02P.

‡The IMS and Department of Mathematics, The Chinese University of Hong Kong, Shatin, N. T., Hong Kong (zpxin@ims.cuhk.edu.hk). The research of this author was supported in part by grants from RGC of HKSAR CUHK4028/04P, CUHK4279/00P, CUHK4040/02P, and the Zheng Ge Ru Foundation.

§Institut für Mathematik, Universität Wien, Boltzmanngasse 9, A-1090 Wien, Austria (peter.markowich@univie.ac.at). The research of this author was supported by the EU-funded network HYKE and the Wittgenstein-Award 2000 of Peter A. Markowich, funded by the Austrian Science Fund FWF.

In this paper, we assume that $D(x)$ is a smooth function.

Note that for the sake of simplicity we take insulating boundary conditions modeled by outward electric field and current density components, namely (4).

A necessary solvability condition for the Poisson equation (3) subject to the Neumann boundary condition for the field in $(4)_3$ is global space charge neutrality

$$\int_0^1 (n^\lambda - p^\lambda - D)dx = 0.$$

Since the total numbers of electrons and holes are conserved, it is sufficient to require the following corresponding condition for the initial data:

$$\int_0^1 (n_0^\lambda - p_0^\lambda - D)dx = 0. \tag{6}$$

Usually semiconductor physics are concerned with large-scale structures with respect to the Debye length λ (λ takes small values, typically $\lambda^2 \approx 10^{-7}$). For such scales, the semiconductor is almost electrically neutral, i.e., there is no space charge separation or electric field. This is the so-called quasi-neutrality assumption of semiconductors or plasma physics, which was applied by Shockley [31] in the first theoretical studies of semiconductor devices in 1949, but was also applied in other contexts such as the modeling of plasmas [32] and ionic membranes [29]. Under the assumption of space charge neutrality, i.e., $\lambda = 0$, we formally arrive at the following quasi-neutral drift-diffusion model:

$$n_t = (n_x + n\mathcal{E})_x, \tag{7}$$

$$p_t = (p_x - p\mathcal{E})_x, \tag{8}$$

$$\begin{aligned} 0 &= n - p - D, \\ \mathcal{E} &= -\Phi_x. \end{aligned} \tag{9}$$

This formal limit was obtained by Roosbroeck [28] in 1950. For further formal asymptotic analysis, see [25, 27, 20, 33].

Generally speaking, it should be expected at least formally that $(n^\lambda, p^\lambda, -\Phi_x^\lambda) \to (n, p, \mathcal{E})$ as $\lambda \to 0$ in the interior of the interval $[0,1]$, while it cannot be expected a priori that all boundary and initial value conditions are maintained for the limit problem because of the singular perturbation character of the problem (the Poisson equation becomes an algebraic equation in the limit). However, by the conservation form of the continuity equations the property of zero fluxes through the boundary will prevail in the limit

$$n_x + n\mathcal{E} = 0, \qquad p_x - p\mathcal{E} = 0, \quad x = 0, 1, \quad t > 0, \tag{10}$$

while the boundary condition for the electric field E^λ does not.

Similarly, we can expect a priori that quasi-neutral drift-diffusion models (7)–(9) are supplemented by the initial data

$$n(x,0) = n_0(x), \quad p(x,0) = p_0(x) \tag{11}$$

satisfying locally initial time space charge neutrality

$$n_0 - p_0 - D = 0. \tag{12}$$

The aim of this paper is to justify rigorously the above formal limit for $O(1)$-time and sufficiently smooth solutions.

It is important to mention that the quasi-neutral limit is a well-known challenging and physically very complex modeling problem for (bipolar) fluid dynamic models and for kinetic models of semiconductors and plasmas. In both cases there exist only partial results. In particular, for time-dependent transport models, the limit $\lambda \to 0$ has been performed for the Vlasov–Poisson system by Brenier [2], Grenier [12, 13], and Masmoudi [21]; for the Schrödinger–Poisson system by Puel [26] and Jüngel and Wang [16]; for the drift-diffusion-Poisson system by Gasser et al. [10, 11], Jüngel and Peng [15], and Schmeiser and Wang [30] under much more restrictive assumptions on the doping profile that is used in this paper (no sign changes of D are allowed); and for the Euler–Poisson system by Cordier and Grenier [5, 6], Cordier et al. [4], and Wang [35]. However, as already mentioned, all these results are restricted to *the special cases of doping profiles*, i.e., *either assuming that $D(x)$ is constant zero or assuming that $D(x)$ does not change its sign*. But p-n junctions are of great importance both in modern electronic applications and in understanding semiconductor devices since the p-n junction theory serves as the foundation of the physics of semiconductor devices (see Sze [34]). For physically interesting doping profiles with p-n junctions, i.e., for the case where the doping profile can change its sign, there is no rigorous result available for time-dependent semiconductor models for either fluid dynamic models or for kinetic models up to now. Therefore, it is natural to study the quasi-neutral limit on the level of the drift-diffusion-Poisson models first. For stationary drift-diffusion-Poisson models, rigorous convergence results for p-n junction devices with contacts can be found in Markowich [19], and recent extensions were done by Caffarelli et al. [3] and Dolbeault, Markowich, and Unterreiter [7].

In this paper we consider the quasi-neutral limit of the time-dependent drift-diffusion model (1)–(6) for semiconductors with p-n junctions in the general case of physically relevant sign-changing doping profiles.

Our main result can be summarized as follows: The convergence of the drift diffusion models (1)–(5) to (7)–(11) is rigorously proved for general sign-changing and smooth doping profiles in one-space dimension case on time intervals, on which a smooth nonvacuum solution of the reduced problem (7)–(11) exists (the precise statement will be given in section 2).

We mention that one of the main difficulties in dealing with quasi-neutral limits is the oscillatory behavior of the electric field. Usually it is difficult to obtain uniform estimates on the electric field with respect to the Debye length λ due to a possible vacuum set of the density, in particular, the occurrence of the depletion region.

In this paper, we overcome this difficulty by the following strategy. First, we introduce a new transformation (see (13) in section 2) that changes the original problem, (1)–(5), into an equivalent one, (14)–(17), where the dissipation for the electric field becomes apparent. Second, we construct carefully a better uniformly valid approximation solution to (14)–(17) by using a method of matched asymptotic analysis. Finally, we show the asymptotic structural stability of the resulting approximate solution by energy methods based on studies on a weighted entropy and entropy dissipation. This approach is strongly motivated by the analysis of boundary layers in the fluid-dynamic limit of a nonlinear Boltzmann equation by Liu and Xin in [18] and viscous boundary layers by Xin in [36]. This ansatz deviates from the solutions to (7)–(11) slightly in a region away from the parabolic boundaries, since it changes more rapidly in the electronic field in the parabolic boundaries (boundary layers and initial layers). It is this kind of structure of the ansatz that governs the main oscillatory feature of the electric field and plays a key role in our analysis. Due to the special structure of the quasi-neutral drift-diffusion model (7)–(11), and the more or less explicit forms of the boundary layer functions and initial layer functions, the asymptotic ansatz can be estimated rather easily. Thus, the quasi-neutral limit problem is reduced to show the structural stability of such an ansatz. To achieve this goal, our basic idea in the whole analysis in this paper is that the more accurate approximate solutions make it easier to obtain their structural stability for the following reasons.

First, though the linearized system around the approximate solutions is complictaed and involves many scales due to the presence of boundary and initial layers, the detailed forms of the layer structure in the ansatz and the explicit form of the dissipation for the electric field make it possible to derive the uniform energy estimates for the linearized operator. Furthermore, the ansatz is constructed in such a way so that the deviation from the ansatz, the error, satisfies the so-called "error" system (see (104)–(105)) which is only weakly nonlinear with source terms suitably small in appropriate norms. Since the detailed estimates, given in section 4, are very lengthy and involved, we would like to outline some main ingredients here. Note that although the "error" system is very complicated due to the layer structure of the ansatz and the complexity of the problem involved, it is in fact a coupled nonlinear parabolic system with nonhomogeneous source terms, and the major part of its linearized operator has an extra dissipative term for the electric field due to the uniform positivity of the total interior density $\mathcal{Z}^0$, which yields the uniform energy estimates of the electric field with respect to the Debye length λ if one just deals with only the linearized system of the error system. Moreover, according to the structure of the ansatz, we can decompose the inhomogeneous source term of the error system into the inner part, the boundary layer part, the initial layer part, and the mixed boundary and initial layer part, which has a uniform decay rate with respect to λ in a weighted Sobolev's norm. Hence an extra power of λ can be created in the energy bound, as is the main purpose of the approximate solution constructed in this paper. Because there is no maximal principle for the error system (104)–(105), the classical energy method is employed to establish the uniform energy estimates with respect to the Debye length λ in Sobolev's space. However, due to the complexity of the "error" equations, it is not easy to obtain the uniformly a priori energy bound with respect to the Debye length λ or the decay rate (a power of λ) if one just uses only the standard energy estimate techniques. So one needs some new techniques for establishing the uniformly a priori estimate in order to derive the energy bound. One basic point of our approach here is to derive high order spatial energy estimates by combining the basic L^2-estimates of $(z_R^\lambda, E_R^\lambda)$ (see below) and the L^2-estimates of the time derivatives of $(z_R^\lambda, E_R^\lambda)$ to avoid meeting the strong singularity caused by the boundary layer function appearing in the ansatz. Hence the inhomogeneous source terms can be controlled by further using the decay rate of the boundary layer and initial layer functions. Another technique involved here is to control strong nonlinear oscillation, caused by the nonlinearity of the error system, by a combination of the entropy and the entropy dissipation, which is the reason why one can establish the uniform energy bound. Namely, one controls the deviation of the solution to (14)–(17) from the ansatz by introducing the following two λ-weighted Liapunov-type functionals:

$$\Gamma^\lambda(t) = \int_0^1 \big(|z_R^\lambda|^2 + |z_{R,x}^\lambda|^2 + |z_{R,t}^\lambda|^2 + \lambda^2(|E_R^\lambda|^2 + |E_{R,x}^\lambda|^2 + |E_{R,t}^\lambda|^2) + |E_R^\lambda|^2\big)dx$$

and

$$G^\lambda(t) = \int_0^1 \big(|z_{R,x}^\lambda|^2 + |z_{R,xt}^\lambda|^2 + |E_R^\lambda|^2 + |E_{R,t}^\lambda|^2 + \lambda^2(|E_{R,x}^\lambda|^2 + |E_{R,xt}^\lambda|^2)\big)dx,$$

where $z_R^\lambda = n_R^\lambda + p_R^\lambda$, $E_R^\lambda = -\Phi_{R,x}^\lambda$, and $(n_R^\lambda, p_R^\lambda, \Phi_R^\lambda)^T$ denotes the difference between the solution to (14)–(17) from the ansatz (see section 2 for details). Notice that, physically, $\Gamma^\lambda(t)$ is called the entropy while $G^\lambda(t)$ is called the entropy dissipation because it is an entropy derived from the dissipation term of the error system (14)–(17). By a careful energy method, we are able to prove the entropy production integration inequality

$$\begin{aligned}
&\Gamma^\lambda(t) + \int_0^t G^\lambda(s)ds \\
&\leq M\Gamma^\lambda(t=0) + M\int_0^t (\Gamma^\lambda(s) + (\Gamma^\lambda(s))^\iota)ds \\
&\quad + M\int_0^t \Gamma^\lambda(s)G^\lambda(s)ds + M\lambda^q, \quad t \geq 0,
\end{aligned}$$

for some $\iota > 1$, $q > 0$, and $M > 0$, independent of λ, which implies the desired convergence result of this paper.

Finally, we also mention that for drift-diffusion models there are many results on existence, uniqueness, large time asymptotic behavior, stability of stationary states, regularity of weak solutions, etc. For example, see [1, 8, 9, 14, 20, 22, 23, 24].

The plan of this paper is as follows. In section 2 we reformulate our problem and state the main results of this paper; furthermore, the existence and regularity of solutions to the quasi-neutral drift-diffusion models are discussed. In section 3 we give the approximate solutions by a method of matched asymptotic analysis and study the properties of initial and boundary layer function. Finally, section 4 is devoted to the energy estimates for the main theorems of this paper.

2. Reformulation of the equations and main results. Introduce the new variables (z^λ, E^λ) by the following transformation:

$$E^\lambda = -\Phi_x^\lambda, \quad n^\lambda = \frac{z^\lambda + D - \lambda^2 E_x^\lambda}{2}, \quad p^\lambda = \frac{z^\lambda - D + \lambda^2 E_x^\lambda}{2} \quad (z^\lambda = n^\lambda + p^\lambda). \tag{13}$$

Adding (1) and (2) and using (3); taking ∂_t of (3) and replacing n_t, p_t of the resulting equation by (1), (2) and then integrating over $[0, x]$ with respect to x; and using the boundary condition (4) and the transformation (13), we can reduce the initial boundary value problem (1)–(6) to the following equivalent system for (z^λ, E^λ):

$$z_t^\lambda = (z_x^\lambda + DE^\lambda)_x - \lambda^2(E^\lambda E_x^\lambda)_x, \quad 0 \leq x \leq 1, \quad t > 0, \tag{14}$$

$$\lambda^2(E_t^\lambda - E_{xx}^\lambda) = -(D_x + z^\lambda E^\lambda), \quad 0 \leq x \leq 1, \quad t > 0, \tag{15}$$

$$z_x^\lambda = E^\lambda = 0, \quad x = 0, 1, \quad t > 0, \tag{16}$$

$$z^\lambda(x, 0) = z_0^\lambda(x), \quad E^\lambda(x, 0) = E_0^\lambda(x), \quad 0 \leq x \leq 1. \tag{17}$$

Note that the equivalence between system (1)–(6) and system (14)–(17) is easily verified for classical solutions by using the transformation (13). Thus, we have the following proposition.

PROPOSITION 1 (existence and uniqueness). *Assume that* $(z_0^\lambda, E_0^\lambda) \in (C^2)^2$ *satisfies the compatibility conditions*

$$z_{0,x}^\lambda = E_0^\lambda = 0, \quad -\lambda^2 E_{0,xx}^\lambda = -D_x \text{ at } x = 0, 1. \tag{18}$$

Then system (14)–(17) *has a unique, global, and classical solution* $(z^\lambda, E^\lambda) \in C^{2,1}([0,1] \times [0, \infty))$.

Remark 1. The existence in Proposition 1 is obtained by the known existence results for (1)–(6) (see, for example, [24, 9]) and by the transformation (13), while uniqueness in Proposition 1 can be proved easily for H^1-solutions of (14)–(17).

Let us assume that the initial datum $(z_0^\lambda, E_0^\lambda)$ is taken to guarantee that boundary-initial consistency for the initial boundary value problem (14)–(17) for $\lambda > 0$ holds. In particular, the compatibility condition (18) is assumed and the initial datum $(z_0^\lambda, E_0^\lambda)$ is assumed to have an expansion of the form

$$
\begin{aligned}
(z_0^\lambda, E_0^\lambda)^T = \Bigg(& z_0^0(x) + \lambda\left(f(x)z_+^1\left(\frac{x}{\lambda}\right) + g(x)z_-^1\left(\frac{1-x}{\lambda}\right)\right) + \lambda z_{0R}^\lambda(x), \\
& E_0^0(x) + f(x)E_+^0\left(\frac{x}{\lambda}\right) + g(x)E_-^0\left(\frac{1-x}{\lambda}\right) + \lambda E_{0R}^\lambda(x)\Bigg)^T .
\end{aligned}
\tag{19}
$$

To justify the rigorous quasi-neutral assumptions, we make the following ansatz for the approximate solution:

$$
\begin{aligned}
(z^\lambda, E^\lambda)_{app}^T = \Bigg(& \mathcal{Z}^0(x,t) + \sum_{i=0}^{2} \lambda^i \big(f(x)z_+^i(\xi,t) + g(x)z_-^i(\eta,t) + z_I^i(x,s)\big), \\
& \mathcal{E}^0(x,t) + f(x)E_+^0(\xi,t) + g(x)E_-^0(\eta,t) + E_I^0(x,s)\Bigg)^T ,
\end{aligned}
\tag{20}
$$

where the inner function $(\mathcal{Z}^0, \mathcal{E}^0)^T$ is independent of λ; $z_+^i, E_+^0, z_-^i, E_-^0$, $i = 0, 1, 2$, are the left boundary layer functions near $x = 0$ and the right boundary layer functions near $x = 1$, respectively; and z_I^i, $i = 0, 1, 2, E_I^0$, are the initial time layer functions near $t = 0$. The cut-off functions $f(x)$ and $g(x)$ are smooth C^2 functions satisfying $f(0) = g(1) = 1$ and $f(1) = f'(1) = f''(1) = f'(0) = f''(0) = g(0) = g'(0) = g''(0) = g'(1) = g''(1) = 0$. Here we set

$$
\xi = \frac{x}{\lambda}, \quad \eta = \frac{1-x}{\lambda}, \quad s = \frac{t}{\lambda^2},
$$

which corresponds physically to the dielectric relaxation time scale, and $(\cdot,\cdot)^T$ represents transposition. We will discuss in detail the construction of the inner, boundary layer, and initial layer functions in the next section; however, we summarize the results here.

First, the inner function $(\mathcal{Z}^0, \mathcal{E}^0)^T$ is determined as a solution of the following initial boundary value problems for the transformed quasi-neutral drift/diffusion equations:

$$\mathcal{Z}_t^0 = (\mathcal{Z}_x^0 + D\mathcal{E}^0)_x, \quad 0 < x < 1, \quad t > 0, \tag{21}$$

$$0 = -(D_x + \mathcal{Z}^0\mathcal{E}^0), \quad 0 < x < 1, \quad t > 0, \tag{22}$$

$$(\mathcal{Z}_x^0 + D\mathcal{E}^0)(0, 1; t) = 0, \quad t > 0, \tag{23}$$

$$\mathcal{Z}^0(x, 0) = z_0^0(x), \quad 0 \le x \le 1. \tag{24}$$

The existence of the above inner problem is guaranteed by the following proposition.

PROPOSITION 2. *Assume that $D \in C^{2(l+1)+1}$ and that $z_0^0 \in C^{2(l+1)}$ for some integer $l \ge 0$. Also assume that $z_0^0 \ge \delta_0 > 0$ satisfy the compatibility condition of order l for (21)–(24). Then there exist a $T_0 \in (0, +\infty]$ and a unique classical* *solution* $(\mathcal{Z}^0, \mathcal{E}^0)$*, well defined on* $[0,1] \times [0,T_0]$*, of* (21)–(24) *satisfying* $\mathcal{Z}^0, \mathcal{E}^0 \in C^{2(l+1),l+1}([0,1] \times [0,T_0])$ *and* $\mathcal{Z}^0(x,t) \geq \delta_1 > 0$ *on* $[0,1] \times [0,T_0]$ *for some positive constant* δ_1*. In particular, if* $D \in C^\infty([0,1])$ *and* $z_0^0 \in C^\infty([0,1])$ *satisfying the compatibility condition of any order, then* $\mathcal{Z}^0, \mathcal{E}^0 \in C^\infty([0,1] \times [0,T_0])$*.*

Moreover, if δ_0 *is suitably large, then* $T_0 = \infty$*.*

Proof of Proposition 2. The proof is elementary. For completeness, we outline it here. First, it follows from (80) that $\mathcal{E}^0(x,t) = -\frac{D_x}{\mathcal{Z}^0(x,t)}$. Then the problem (79)–(82) is reduced to the following system:

$$\mathcal{Z}_t^0 = \left(\mathcal{Z}_x^0 - \frac{DD_x}{\mathcal{Z}^0}\right)_x, \quad 0 < x < 1, \quad t > 0, \tag{25}$$

$$\left(\mathcal{Z}_x^0 - \frac{DD_x}{\mathcal{Z}^0}\right)(0,1;t) = 0, \quad t > 0, \tag{26}$$

$$\mathcal{Z}^0(x,0) = z_0^0(x), \quad 0 \leq x \leq 1. \tag{27}$$

For $z_0^0 \geq \delta_0 > 0$, the standard parabolic theory yields the desired local existence of classical positive solution $\mathcal{Z}^0$. This concludes the first part of Proposition 2.

To prove the global existence of large classical solutions for large initial data, one introduces the transformation

$$(\mathcal{Z}^0)^2 - D^2 = w. \tag{28}$$

Then it follows from the system (25)–(27) that w satisfies

$$w_t = w_{xx} - \frac{w_x^2 + 2DD_x w_x}{2(w + D^2)}, \quad 0 < x < 1, \quad t > 0, \tag{29}$$

$$w_x(0,1,t) = 0, \quad t > 0, \tag{30}$$

$$w(x,0) = w_0(x) = (z_0^0)^2 - D^2. \tag{31}$$

If $\delta_0 \geq \sqrt{D^2 + \delta_2}$ for some $\delta_2 > 0$, then $w_0 \geq \delta_2 > 0$.

By the standard parabolic theory [17], we know that there exists a unique, classical, and global solution w for (29)–(31) satisfying $0 < \delta_2 \leq w \in C^{2(l+1),l+1}([0,1] \times [0,T])$ for any $T > 0$. By transformation (28), we conclude the second part of Proposition 2. The proof of Proposition 2 is complete. ∎

Remark 2. By the transformation

$$n(x,t) = \frac{\mathcal{Z}^0(x,t) + D(x)}{2}, \quad p(x,t) = \frac{\mathcal{Z}^0(x,t) - D(x)}{2}, \quad \mathcal{E}(x,t) = \mathcal{E}^0(x,t)$$

it is easy to verify that the system (7)–(12) and the system (21)–(24) are equivalent. Thus, by Proposition 2, one obtains the existence of the classical nonvacuum solution of the quasi-neutral drift-diffusion system (7)–(12). The uniform positivity of $z_0^0(x)$, together with (12), excludes singularities of the solution of the quasi-neutral drift-diffusion system (7)–(12). Indeed, if $z_0^0(x) = D(x)$, then (21)–(24) has a stationary solution

$$\mathcal{Z}^0(x,t) = D(x), \quad \mathcal{E}^0 = -(lnD(x))_x.$$

In this case, the electric field $\mathcal{E}^0$ has a singularity in the vacuum set of the density $\mathcal{Z}^0$. In the present paper, the case of singular solutions of the quasi-neutral drift-diffusion models is thus not allowed due to our assumption that $z_0^0 \geq \delta_0 > 0$. But the singular solution case is interesting and will be investigated in the future.

Remark 3. Proposition 2 cannot hold true in the unipolar case. In fact, in the unipolar case we must have $z_0^0(x) = D(x)$ or $z_0^0(x) = -D(x)$ due to the local quasi-neutrality assumption (12) of the initial data, and hence we have no uniform positivity of $z_0^0(x)$ if the doping profile $D(x)$ changes sign. This comes back to the above vacuum singular solution case.

Next, the boundary layer functions z_B^i, E_B^0, $B = +/-$, $i = 0, 1, 2$, are governed by the following boundary value problems for the elliptic equations:

$$-E^0_{+,\xi\xi} = J^0_+, \quad -E^0_{-,\eta\eta} = J^0_-, \quad 0 < \xi, \quad \eta < \infty, \quad t > 0, \tag{32}$$

$$E^0_+(\xi = 0, t) = -\mathcal{E}^0(x = 0, t), \quad E^0_-(\eta = 0, t) = -\mathcal{E}^0(x = 1, t), \quad t > 0, \tag{33}$$

$$E^0_+(\xi \to \infty, t) = E^0_-(\eta \to \infty, t) = 0, \quad t > 0, \tag{34}$$

and

$$z^0_+ = z^0_- = z^2_+ = z^2_+ = 0, \quad 0 < \xi, \quad \eta < \infty, \quad t > 0, \tag{35}$$

$$z^1_{+,\xi} + D(0)E^0_+ = 0, \quad 0 < \xi, \quad \eta < \infty, \quad t > 0, \tag{36}$$

$$-z^1_{-,\eta} + D(1)E^0_- = 0, \quad 0 < \xi, \quad \eta < \infty, \quad t > 0, \tag{37}$$

$$z^0_+(\xi \to \infty, t) = z^0_-(\eta \to \infty, t) = 0, \quad t > 0, \tag{38}$$

where

$$J^0_+ = -\mathcal{Z}^0(0, t)E^0_+, \quad J^0_- = -\mathcal{Z}^0(1, t)E^0_-. \tag{39}$$

Finally, the initial layer functions z_I^i, $i = 0, 1, 2, E_I^0$, are given by the following equations (initial value problems):

$$E^0_{I,s} = J^0_I, \quad s > 0, \quad 0 < x < 1, \tag{40}$$

$$E^0_I(x, 0) = E^0_0(x) - \mathcal{E}^0(x, 0), \quad 0 < x < 1, \tag{41}$$

and

$$z^0_I = z^1_I = 0, \quad 0 < x < 1, \quad s > 0, \tag{42}$$

$$z^2_{I,s} = (DE^0_I)_x, \quad 0 < x < 1, \quad s > 0, \tag{43}$$

$$z^2_I(x, 0) = 0, \quad 0 < x < 1, \tag{44}$$

where

$$J^0_I = -\mathcal{Z}^0(x, 0)E^0_I. \tag{45}$$

It follows from the special structures of the boundary layer problem (32)–(38) and the initial layer problem (40)–(44) that the existence of solutions of these equations is immediate. We will solve these equations explicitly in section 3.

Define the error term $(z_R^\lambda, E_R^\lambda)^T$ of the approximation solution (20) to (14)–(17) with the initial datum

$$\begin{aligned}
&(z_0^\lambda, E_0^\lambda)^T \\
&\quad = \left(z_0^0(x) + \lambda\left(f(x)z^1_+\left(\frac{x}{\lambda}, 0\right) + g(x)z^1_-\left(\frac{1-x}{\lambda}, 0\right)\right) + \lambda z^\lambda_{0R}(x),\right. \\
&\qquad \left. E_0^0(x) + f(x)E^0_+ = \left(\frac{x}{\lambda}, 0\right) + g(x)E^0_-\left(\frac{1-x}{\lambda}, 0\right) + \lambda E^\lambda_{0R}(x)\right)^T
\end{aligned} \tag{46}$$

by

$$(z_R^\lambda(x,t), E_R^\lambda(x,t))^T = (z^\lambda, E^\lambda)^T - (z^\lambda, E^\lambda)^T_{app}. \tag{47}$$

THEOREM 3. *Let $l \geq 1$ and all assumptions of Proposition 2 hold. Assume also that the initial datum $(z_0^\lambda, E_0^\lambda)$ satisfies (46) with $E_0^0 \in C^{2(l+1)}([0,1])$,*

$$E_0^0(x)|_{x=0,1} = -\frac{D_x(x)}{z_0^0(x)}|_{x=0,1} (= \mathcal{E}^0(x = 0, 1, \quad t = 0)), \tag{48}$$

and

$$\|z_{0R}^\lambda(x)\|_{H^1} \leq M\sqrt{\lambda}, \quad \|\partial_x^2 z_{0R}^\lambda(x)\|_{L_x^2} \leq M\lambda^{-\frac{1}{2}}, \tag{49}$$

$$\|\partial_x^j E_{0R}^\lambda(x)\|_{L_x^2} \leq M\lambda^{\frac{1}{2}-j}, \quad j = 0, 1, 2. \tag{50}$$

Then, for any $T \in (0, T_0)$, where T_0 is given by Proposition 2, there exist positive constants M and $\lambda_0, \lambda_0 \ll 1$ such that, for any $\lambda \in (0, \lambda_0]$,

$$\sup_{0\leq t\leq T} \left(\|(z_R^\lambda, E_R^\lambda, z_{R,x}^\lambda, z_{R,t}^\lambda)\|_{L_x^2} + \lambda\|(E_R^\lambda, E_{R,x}^\lambda, E_{R,t}^\lambda)\|_{L_x^2} \right) \leq M\sqrt{\lambda^{1-\delta}} \tag{51}$$

for any δ with $0 < \delta < 1$.

In particular, if $(z_0^\lambda, E_0^\lambda)$ satisfies (46) with $(z_{0R}^\lambda, E_{0R}^\lambda) = (0, 0)$, then

$$\sup_{0\leq t\leq T} \|(z^\lambda - \mathcal{Z}^0)(\cdot, t)\|_{L_x^\infty} \leq M\sqrt{\lambda^{1-\delta}}.$$

Remark 4. The compatibility assumption (48) in Theorem 3 is important in our analysis. It guarantees that one can take the "well-prepared" initial datum (46) instead of the general initial datum (19), and hence the ansatz (20) is appropriate in this case while, generally speaking, its breakdown will introduce an extra layer $W_{IB}(x, \xi, \eta, s)$ of mixing of fast time and fast space scales. The main strategy involved here can be applied to this case too. This will be done in the future.

It should also be noted that assumptions (49) and (50) are just technical ones. In general, $(z_{0R}^\lambda, E_{0R}^\lambda)^T$ in (19) can be written as

$$(z_{0R}^\lambda, E_{0R}^\lambda)^T = (z_0^1(x), E_0^1(x))^T + (\tilde{z}_{0R}^\lambda, \tilde{E}_{0R}^\lambda)^T, \tag{52}$$

where

$$(\tilde{z}_{0R}^\lambda, \tilde{E}_{0R}^\lambda)^T = \lambda O(1). \tag{53}$$

Here $O(1)$ is a smooth bounded function in $x, \frac{x}{\lambda}, \frac{1-x}{\lambda}$, so that the general assumptions on the initial data become

$$\|(z_{0R}^\lambda - z_0^1)(x)\|_{H^1} \leq M\sqrt{\lambda}, \quad \|\partial_x^2(z_{0R}^\lambda - z_0^1)(x)\|_{L_x^2} \leq M\lambda^{-\frac{1}{2}}, \tag{54}$$

$$\|\partial_x^j(E_{0R}^\lambda - E_0^1)(x)\|_{L_x^2} \leq M\lambda^{\frac{1}{2}-j}, \quad j = 0, 1, 2. \tag{55}$$

In this case, it turns out that an additional correction term $\lambda(z_0^1, E_0^1)$, and hence an extra initial layer term $(\lambda^3 z_I^3, \lambda E_I^1)$, caused by z_0^1, will appear in the solution. Thus, we have more general results as follows.

THEOREM 4. *Under the assumptions of Theorem 3, with assumptions (49) and (50) replaced by (54) and (55) with $(z_0^1, E_0^1) \in C^3$, we have that, for any*

$T \in (0, T_0)$, where T_0 is given by Proposition 2, there exist positive constants M and $\lambda_0, \lambda_0 \ll 1$ such that, for any $\lambda \in (0, \lambda_0]$,

$$\sup_{0\le t\le T}\left(\|(\tilde z_R^\lambda, \tilde E_R^\lambda, \tilde z_{R,x}^\lambda, \tilde z_{R,t}^\lambda)\|_{L^2_x} + \lambda\|(\tilde E_R^\lambda, \tilde E_{R,x}^\lambda, \tilde E_{R,t}^\lambda)\|_{L^2_x}\right) \le M\sqrt{\lambda^{1-\delta}} \tag{56}$$

for any $\delta \in (0,1)$, where

$$\begin{aligned}(\tilde z_R^\lambda(x,t), \tilde E_R^\lambda(x,t))^T &= (z_R^\lambda(x,t), E_R^\lambda(x,t))^T - (\lambda z_0^1 + \lambda^3 z_I^3, \lambda(E_0^1 + E_I^1))^T\\ &= (z^\lambda, E^\lambda)^T - (z^\lambda, E^\lambda)^T_{app} - (\lambda z_0^1 + \lambda^3 z_I^3, \lambda(E_0^1 + E_I^1))^T.\end{aligned}$$

The initial layer functions z_I^3 and E_I^1 solve the following problems (initial value problems):

$$E^1_{I,s} = -\mathcal{Z}^0(x,0)E_I^1 - z_0^1(x)E_I^0, \quad s>0, \quad 0<x<1, \tag{57}$$

$$E_I^1(x,0) = 0, \quad 0<x<1, \tag{58}$$

and

$$z^3_{I,s} = (DE_I^1)_x, \quad 0<x<1, \quad s>0, \tag{59}$$

$$z_I^3(x,0) = 0, \quad 0<x<1. \tag{60}$$

Remark 5. It follows from Theorems 3 and 4 above that the approximation of vanishing space charge holds in the interior part of the parabolic domain, but it cannot be valid uniformly up to the boundary in the case where the doping profile changes its sign.

Remark 6. Note that our smoothness assumption on the doping profile D excludes so-called abrupt p-n junctions, where the doping profile has a jump discontinuity. An additional layer thus has to be introduced locally at abrupt junctions. This will be studied further in the future.

Remark 7. It should be noted that in Theorems 3 and 4 the quasi-neutral limits justified rigorously only in spatial L^2-norm. In order to justify this limit in supernorm, a more accurate ansatz than (20) has to be constructed by using higher order corrections. This is left for the future.

3. Approximate solutions and matched asymptotic analysis. In this section we derive the limit equation and the forms of the boundary layers and of the initial time layers by the multiple scaling asymptotic expansion of a singular perturbation with respect to the scaled Debye length.

Let us look for $W^\lambda = (z^\lambda, E^\lambda)^T$ of the form

$$W^\lambda = \sum_{i=0}^N \lambda^i W^i\left(x, \frac{x}{\lambda}, \frac{1-x}{\lambda}, t, \frac{t}{\lambda^2}\right) + W_R^\lambda(x,t),$$

where λ and λ^2 are the lengths of the boundary layer and of the initial time layer, respectively, and

$$W^i = W^i_{Inn}(x,t) + W^i_B(x,\xi,\eta,t) + W^i_I(x,s)$$

is the sum of an interior term W^i_{Inn}, the boundary layer term W^i_B near $x=0$ and $x=1$, and the initial time layer term W^i_I near $t=0$. Here we set $\xi = \frac{x}{\lambda}$, $\eta = \frac{1-x}{\lambda}$, $s = \frac{t}{\lambda^2}$, and $W = (z,E)^T$.

For simplicity of presentation, we will carry out the constructions of boundary layers $W_B^i = W_+^i(\xi, t)$ only near the left boundary, $x = 0$; the parts at $x = 1$ can be done similarly. Thus, we enforce

$$\lim_{\xi\to\infty} W_B(\xi, t) = 0. \tag{61}$$

In this section, without explicitly writing out the scaled variables, the functions marked by Inn, B, I, BI, and R are ones with respect to (x,t), (ξ,t), (x,s), (x,ξ,t,s), and (x,t). In the following we denote (z_{Inn}, E_{Inn}) by $(\mathcal{Z}, \mathcal{E})$.

Our primary interests lie in the rigorous justification of the quasi-neutral assumptions. Thus, we will ignore the higher corrections to the drift-diffusion equations. Hence, we impose the following decomposition for the solution (z^λ, E^λ) of (14)–(17):

$$\begin{aligned}(z^\lambda, E^\lambda)^T = \big(&\mathcal{Z}^0 + z_B^0 + z_I^0 + \lambda(z_B^1 + z_I^1) + \lambda^2(z_B^2 + z_I^2) + z_R^\lambda(x,t),\\ &\mathcal{E}^0 + E_B^0 + E_I^0 + E_R^\lambda(x,t)\big)^T.\end{aligned} \tag{62}$$

Thus, we obtain an approximation of the solution (z^λ, E^λ) of (14)–(17). The expansion (62) will satisfy the differential equations (14)–(15), the boundary condition (16), and the initial condition (17) for arbitrary "well-prepared" initial data $(z_0^\lambda, E_0^\lambda)$ satisfying (46).

Inserting (62) into (14) and (15), by direct computations one gets

$$\begin{aligned}\mathcal{Z}_t^0 &+ \sum_{i=0}^{2}\lambda^i z_{B,t}^i + \frac{1}{\lambda^2} z_{I,s}^0 + \frac{1}{\lambda} z_{I,s}^1 + z_{I,s}^2 + z_{R,t}^\lambda\\ &= \Big[(z_{R,x}^\lambda + DE_R^\lambda)_x + (\mathcal{Z}_x^0 + D\mathcal{E}^0)_x\\ &\quad + \frac{1}{\lambda}\left(\frac{1}{\lambda} z_{B,\xi}^0 + (z_{B,\xi}^1 + D(0)E_B^0) + \lambda z_{B,\xi}^2 + (D(\lambda\xi) - D(0))E_B^0\right)_\xi\\ &\quad + \sum_{i=0}^{2}\lambda^i z_{I,xx}^i + (DE_I^0)_x\Big] - \lambda^2\Big[K_{Inn}^0 + \tilde{K}_B + K_I^\lambda + \tilde{K}_{IB}^\lambda + \tilde{F}_R^\lambda\Big]_x\end{aligned} \tag{63}$$

and

$$\begin{aligned}&\lambda^2(\mathcal{E}_t^0 - \mathcal{E}_{xx}^0) + \big(E_{I,s}^0 - \lambda^2 E_{I,xx}^0\big) + \big(\lambda^2 E_{B,t}^0 - E_{B\xi\xi}^0\big) + \lambda^2(E_{R,t}^\lambda - E_{R,xx}^\lambda)\\ &\quad = J_{Inn}^0 + (\tilde{J}_B^0 + \tilde{J}_{BR}^0) + (J_I^0 + J_{IR}^0) + \tilde{J}_{BI}^0 + \sum_{i=1}^{2}\lambda^i\big(\tilde{J}_B^i + J_I^i + \tilde{J}_{BI}^i\big) + \tilde{G}_R^\lambda,\end{aligned} \tag{64}$$

where K_{Inn}^0, $\tilde{K}_B^i$, K_I^λ, $\tilde{K}_{IB}^\lambda$, and $\tilde{F}_R^\lambda$ are defined by

$$\begin{aligned}K_{Inn}^0 &= \mathcal{E}^0\mathcal{E}_x^0,\\ \tilde{K}_B^\lambda &= (\mathcal{E}^0(0,t) + E_B^0)E_{B,\xi}^0 + E_B^0\mathcal{E}_x^0(0,t)\\ &\quad + (\mathcal{E}^0(\lambda\xi,t) - \mathcal{E}^0(0,t))E_{B,\xi}^0 + E_B^0(\mathcal{E}_x^0(\lambda\xi,t) - \mathcal{E}_x^0(0,t)),\\ K_I^\lambda &= \big(\mathcal{E}^0(x,\lambda^2 s)E_{I,x}^0 + E_I^0(E_{Inn,x}^0(x,\lambda^2 s) + E_{I,x}^0)\big),\\ \tilde{K}_{IB}^\lambda &= E_B^0 E_{I,x}^0 + E_I^0\frac{1}{\lambda}E_{B,\xi}^0,\\ \tilde{F}_R^\lambda &= (\mathcal{E}^0 + E_B^0 + E_I^0)E_{R,x}^\lambda + E_R^\lambda\left(\mathcal{E}_x^0 + E_{B,\xi}^0\frac{1}{\lambda} + E_{I,x}^0\right) + E_{R,x}^\lambda,\end{aligned}$$

and $J^0_{Inn}, \tilde{J}^0_B, \tilde{J}^0_{BR}, J^0_I, J^0_{IR}, \tilde{J}^0_{BI}, \tilde{J}^i_B, J^i_I, \tilde{J}^i_{BI}$, $i = 1, 2$, and $\tilde{G}^\lambda_R$ are defined by the following:

$$
\begin{aligned}
J^0_{Inn} &= -(D_x + \mathcal{Z}^0\mathcal{E}^0),\\
\tilde{J}^0_B &= -(\mathcal{Z}^0(0,t)E^0_B + z^0_B(\mathcal{E}^0(0,t) + E^0_B)),\\
\tilde{J}^0_{BR} &= -\big((\mathcal{Z}^0 - \mathcal{Z}^0(0,t))E^0_B + z^0_B(\mathcal{E} - \mathcal{E}^0(0,t))\big),\\
J^0_I &= -\big(\mathcal{Z}^0(x,0)E^0_I + z^0_I(\mathcal{E}^0(x,0) + E^0_I)\big),\\
J^0_{IR} &= -\big((\mathcal{Z}^0 - \mathcal{Z}^0(x,0))E^0_I + z^0_I(\mathcal{E}^0 - \mathcal{E}^0(x,0))\big),\\
\tilde{J}^0_{BI} &= -(z^0_B E^0_I + z^0_I E^0_B),\\
\tilde{J}^i_B &= -z^i_B(\mathcal{E}^0 + E^0_B), \quad i = 1, 2,\\
J^i_I &= -z^i_I(\mathcal{E}^0 + E^0_I), \quad i = 1, 2,\\
\tilde{J}^i_{BI} &= -z^i_B E^0_I + z^i_I E^0_B, \quad i = 1, 2,
\end{aligned}
$$

and

$$
G^\lambda_R = -\left((\mathcal{E}^0 + E^0_B + E^0_I)z^\lambda_R + \left(\mathcal{Z}^0 + z^0_B + z^0_I + \sum_{i=1}^{2}\lambda^i(z^i_B + z^i_I)\right)E^\lambda_R\right) - z^\lambda_R E^\lambda_R.
$$

Similarly, inserting (62) into the boundary condition (16) yields an expansion at the boundary $x = 0$. Since the boundary expansion is expected to correct the boundary conditions of inner solutions to quasi-neutral drift-diffusion equations well, according to the expansion at the boundary $x = 0$ we may impose the following boundary conditions:

$$z^0_{B,\xi}(\xi = 0; t) = 0, \tag{65}$$

$$z^1_{B,\xi}(\xi = 0; t) = -\mathcal{Z}^0_x(x = 0; t), \tag{66}$$

$$z^2_{B,\xi}(\xi = 0; t) = 0, \tag{67}$$

$$E^0_B(\xi = 0; t) = -\mathcal{E}^0(x = 0; t). \tag{68}$$

Now we start to derive the equations of the inner solution $(\mathcal{Z}^0, \mathcal{E}^0)$ of the various orders of boundary layer and initial time layer functions in the above expansion (62) by comparing coefficients of $O(\lambda^k)$ of (63) and (64). At the leading order λ^{-2} of (63), one gets

$$z^0_{I,s}(x, s) = 0. \tag{69}$$

For z^0_I we take the initial data

$$z^0_I(x, 0) = 0. \tag{70}$$

The only solution of (69) and (70) is given as

$$z^0_I(x, s) = 0, \quad x \in [0, 1], \quad s \geq 0. \tag{71}$$

Similarly

$$z^0_{B,\xi\xi} = 0. \tag{72}$$

We also expect the decay condition at the infinity for z_B^0 such that

$$z_B^0(\xi, t) \to 0 \text{ as } \xi \to \infty. \tag{73}$$

The only solution of (65), (72), and (73) is given as

$$z_B^0(\xi, t) = 0, \quad \xi \geq 0, \quad t \geq 0, \tag{74}$$

which partially explains that the Neumann boundary condition of the density does not produce the boundary layer at the leading order.

At the order λ^{-1} of (63), one gets

$$z_{I,s}^1(x, s) = 0, \text{ hence } z_I^1 = 0, \quad x \in [0, 1], \quad s \geq 0, \tag{75}$$

since $z_I^1(x, 0) = 0$.

One also has from the order λ^{-1} of (63) that

$$z_{B,\xi\xi}^1 + D(0)E_{B,\xi}^0 = 0. \tag{76}$$

As before, we impose the decay condition at the infinity such that

$$z_B^1(\xi, t) \to 0, \quad E_B^0(\xi, t) \to 0 \quad \text{as } \xi \to \infty. \tag{77}$$

It follows from (76) and (77) that

$$z_{B,\xi}^1 + D(0)E_B^0 = 0. \tag{78}$$

Next, we determine the limit equations, which form a system satisfied by the first-order term $(\mathcal{Z}^0, \mathcal{E}^0)$ in the above expansion. At the order of λ^0 of (63) and (64), one gets

$$\mathcal{Z}_t^0 = (\mathcal{Z}_x^0 + D\mathcal{E}^0)_x, \quad 0 < x < 1, \quad t > 0, \tag{79}$$

$$0 = -(D_x + \mathcal{Z}^0\mathcal{E}^0), \quad 0 < x < 1, \quad t > 0. \tag{80}$$

This is nothing but the well-known quasi-neutral drift-diffusion model, which can be formally obtained by setting λ equal to zero in (14)–(15), too. In the context of semiconductor device physics, problems (79)–(80) are referred to as "space charge approximation."

Notice that (80) is an algebraic equation. If $\mathcal{Z}^0(x, t) \geq C_0 > 0$, then

$$\mathcal{E}^0(x, t) = -\frac{D_x(x)}{\mathcal{Z}^0(x, t)}.$$

Generally speaking, $\mathcal{E}^0(x, t)|_{x=0,1} \neq 0$, but $E^\lambda(x, t)|_{x=0,1} = 0$. Therefore, $\mathcal{E}^0(x, t)$ has to be supplemented by a boundary layer term there. Similarly, owing to the arbitrariness of the initial data $E_0^\lambda(x)$, there is an initial layer. Furthermore, it should be clear that the boundary and initial layers are caused by the electric field.

Now we supplement the limit equations (79)–(80) by the appropriate boundary conditions. According to conditions (66) and (68) and equation (78), one gets

$$\mathcal{Z}_x^0(x = 0, t) = -z_{B,\xi}^1(\xi = 0, t) = D(0)E_B^0(\xi = 0) = -D(0)\mathcal{E}^0(x = 0, t), \qquad t \geq 0,$$

i.e.,

$$\mathcal{Z}_x^0 + D(0)\mathcal{E}^0 = 0, \quad x = 0, \quad t \geq 0. \tag{81}$$

For the initial data of $\mathcal{Z}^0(x, t)$, we can take this as

$$\mathcal{Z}^0(x, 0) = z_0^0(x) \geq \delta > 0, \quad 0 \leq x \leq 1. \tag{82}$$

Here $z_0^0(x)$ is given by (19).

Finally, one also gets from the order λ^0 of (63) and (64) that

$$E^0_{I,s}(x,s) = J^0_I(x,s), \quad 0 \le x \le 1, \quad s \ge 0, \tag{83}$$

$$z^2_{I,s}(x,s) = (D(x)E^0_I)_x, \quad 0 \le x \le 1, \quad s \ge 0, \tag{84}$$

and

$$-E^0_{B,\xi\xi}(\xi,t) = \tilde{J}^0_B(\xi,t), \quad \xi > 0, \quad t > 0, \tag{85}$$

$$z^2_{B,\xi\xi} = 0, \quad \xi > 0, \quad t > 0. \tag{86}$$

The initial data of E^0_I can be taken as

$$E^0_I(x,0) = E^0_0(x) - \mathcal{E}^0(x,0), \quad 0 \le x \le 1. \tag{87}$$

The only solution to (83) and (87) can be given explicitly by

$$E^0_I(x,s) = (E^0_0(x) - \mathcal{E}^0(x,0)) \exp(-z^0_0(x)s). \tag{88}$$

The initial data of z^2_I are

$$z^2_I(x,0) = 0, \quad 0 \le x \le 1. \tag{89}$$

The unique solution of (84) and (89) can be given, using (88), by

$$\begin{aligned} z^2_I(x,s) &= \int_0^s (D(x)E^0_I)_x ds \\ &= b(x) + (b_0(x) + b_1(x)s) \exp\{-z^0_0(x)s\}, \quad 0 \le x \le 1, \quad s \ge 0, \end{aligned} \tag{90}$$

where $b(x)$, $b_0(x)$, and $b_1(x)$ depend only upon $D(x)$ and (z^0_0, E^0_0) and satisfy $b_0(x) = -b(x) \neq 0$.

For E^0_B, we impose the decay condition at infinity as

$$E^0_B(\xi,t) = 0 \text{ as } \xi \to \infty, \tag{91}$$

and we also take the boundary condition at $\xi = 0$ as

$$E^0_B(\xi = 0,t) = -\mathcal{E}^0(x = 0,t), \quad t > 0. \tag{92}$$

The unique solution of (85), (91), and (92) can be given by

$$E^0_B(\xi,t) = -\mathcal{E}^0(0,t) \exp(-\sqrt{\mathcal{Z}^0(0,t)}\xi). \tag{93}$$

For z^2_B, the decay condition at infinity is

$$z^2_B(\xi,t) = 0 \text{ as } \xi \to \infty. \tag{94}$$

Then the only solution of (86), (67), and (94) is given by

$$z^2_B(\xi,t) = 0, \quad \xi > 0, \quad t > 0. \tag{95}$$

Similarly, we can construct the boundary layer functions near $x = 1$ and hence deduce the similar boundary conditions of the inner solutions at $x = 1$.

We end this section by summarizing some properties of the boundary layers and the initial time layer and discussing their decay rate, which will be useful for the energy estimates in the next section.

Using (93), one gets from (78) and (77) that

$$\begin{aligned} z_B^1(\xi,t) &= -D(0)\mathcal{E}^0(0,t)\int_\xi^\infty e^{-\sqrt{\mathcal{Z}^0(0,t)}y}dy \\ &= -\frac{D(0)\mathcal{E}^0(0,t)}{\sqrt{\mathcal{Z}^0(0,t)}}e^{-\sqrt{\mathcal{Z}^0(0,t)}\xi}. \end{aligned} \tag{96}$$

Thus, we obtained exact formulae of all initial layer functions and the left boundary layer functions. In particular, we can determine the values of $z_B^1(\xi)$ and $E_B^0(\xi)$, depending only upon $D(0)$, $z_0^0(0)$, and $\mathcal{E}^0(0,0) = E_0^0(0)$, which are given by

$$\begin{aligned} z_B^1(\xi) &= z_B^1(x,0) \\ &= -\frac{D(0)E_0^0(0)}{\sqrt{z_0^0(0)}}e^{-\sqrt{z_0^0(0)}\xi}, \quad \xi > 0, \end{aligned} \tag{97}$$

$$E_B^0(\xi) = E_B^0(x,0) = -E_0^0(0)e^{-\sqrt{z_0^0(0)}\xi}, \quad \xi > 0. \tag{98}$$

Similarly, the right boundary layer functions at $x = 1$, denoted by $z_-^i(\eta,t)$, $i = 0,1,2, E_-^0(\eta,t)$, satisfy similar equations and have exactly the same properties as the left boundary layer functions $z_B^i(\xi,t)$, $i = 0,1,2, E_B^0(\xi,t)$ at $x = 0$, denoted by $z_+^i(\xi,t)$, $i = 0,1,2, E_+^0(\xi,t)$. We omit this.

Thus, we have the following properties of the boundary layer and initial layer functions.

LEMMA 5. (i) $z_+^0 = z_-^0 = z_+^2 = z_-^2 = z_I^0 = z_I^1 = 0$.

(ii) *Assume that the inner solution $(\mathcal{Z}^0,\mathcal{E}^0)$ is C^∞. Then,*

(a) *for any $T > 0$, there exists a positive constant M independent of λ such that*

$$\|\partial_t^{k_1}(\xi^{k_2}\partial_\xi^{k_3}(z_+^1,E_+^0),\eta^{k_4}\partial_\eta^{k_5}(z_-^1,E_-^0))\|_{L^\infty_{(x,t)}([0,1]\times[0,T])} \le M \tag{99}$$

and

$$\|\partial_t^{k_1}(\xi^{k_2}\partial_\xi^{k_3}(z_+^1,E_+^0),\eta^{k_4}\partial_\eta^{k_5}(z_-^1,E_-^0))\|_{L_t^\infty([0,T];L_x^2([0,1]))} \le M\lambda^{\frac{1}{2}} \tag{100}$$

for any nonnegative integer k_j, $j = 0,\dots,5$;

(b) *for any $T > 0$, there exists a positive constant M independent of λ such that*

$$\|\partial_x^{k_6}(z_I^2,s^{k_7}(\partial_s^{k_8}z_I^2,\partial_s^{k_9}E_I^0))\|_{L^\infty_{(x,t)}([0,1]\times[0,T])} \le M \tag{101}$$

and

$$\|\partial_x^{k_{10}}s^{k_{11}}(\partial_s^{k_{12}}z_I^2,\partial_s^{k_{13}}E_I^0)\|_{L_t^2([0,T];L_x^\infty([0,1]))} \le M\lambda \tag{102}$$

for any nonnegative integer k_j, $j = 6,7,9,10,11,13$, and any positive integers k_8,k_{12}.

4. Energy estimates. In this section we investigate the asymptotic behavior of the solution to the problem (14)–(17) as $\lambda \to 0$ and prove our main theorems, Theorems 3 and 4. From now on, we may assume $0 < \lambda \le 1$.

4.1. The proof of Theorem 3. In this subsection we prove Theorem 3 by a careful energy method based on the approximate solutions constructed in the previous section.

Let $\mathcal{Z}^0, \mathcal{E}^0, E_+^0, E_-^0, E_I^0, z_+^1, z_-^1, z_I^2$ be the functions constructed in the previous sections.

Let us assume

$$
\begin{aligned}
&(z^\lambda(x,0), E^\lambda(x,0))^T \\
&\quad = \left(z_0^0(x) + \lambda \left(f(x) z_+^1 \left(\frac{x}{\lambda}, 0\right) + g(x) z_-^1 \left(\frac{1-x}{\lambda}, 0\right)\right) + \lambda z_{0R}^\lambda(x), \right. \\
&\qquad \left. E_0^0(x) + f(x) E_+^0 \left(\frac{x}{\lambda}, 0\right) + g(x) E_-^0 \left(\frac{1-x}{\lambda}, 0\right) + \lambda E_{0R}^\lambda(x) \right)^T,
\end{aligned}
$$

where $f(x)$ and $g(x)$ are two smooth C^2 cut-off functions satisfying $f(0) = g(1) = 1$ and $f(1) = f'(1) = f''(1) = f'(0) = f''(0) = g(0) = g'(0) = g''(0) = g'(1) = g''(1) = 0$, and $(z_{0R}^\lambda, E_{0R}^\lambda)$ satisfies assumptions (49) and (50). In this case, one gets

$$
(z_R^\lambda, E_R^\lambda)^T(x,0) = \lambda (z_{0R}^\lambda(x), E_{0R}^\lambda(x))^T.
$$

Replacing $(z^\lambda, E^\lambda)^T$ by

$$
\begin{aligned}
(z^\lambda, E^\lambda)^T = &\Big(\mathcal{Z}^0 + \lambda (f(x) z_+^1 + g(x) z_-^1) + \lambda^2 z_I^2 + z_R^\lambda(x,t), \\
&\quad \mathcal{E}^0 + f(x) E_+^0 + g(x) E_-^0 + E_I^0 + E_R^\lambda(x,t) \Big)^T
\end{aligned} \tag{103}
$$

in the system (14)–(15) and using the equations of the inner solutions, the boundary layers, and the initial layers, one gets

$$
z_{R,t}^\lambda = H_x^\lambda + f^\lambda, \quad 0 < x < 1, \quad t > 0, \tag{104}
$$

$$
\lambda^2 (E_{R,t}^\lambda - E_{R,xx}^\lambda) + \mathcal{Z}^0 E_R^\lambda = g^\lambda, \quad 0 < x < 1, \quad t > 0, \tag{105}
$$

where

$$
\begin{aligned}
&H^\lambda = z_{R,x}^\lambda + D E_R^\lambda + H_{Inn} + H_B^\lambda + H_I^\lambda + H_{IB}^\lambda + H_R^\lambda, \\
&f^\lambda = -\lambda^1 (f(x) z_{+,t}^1 + g(x) z_{-,t}^1), \quad g^\lambda = G_{Inn} + G_B^\lambda + G_I^\lambda + G_{IB}^\lambda + G_R^\lambda,
\end{aligned}
$$

and H_{Inn} (G_{Inn}), H_B^λ (G_B^λ), H_I^λ (G_I^λ), H_{IB}^λ (G_{IB}^λ), H_R^λ (G_R^λ) represent the inner part, the boundary layer part, the initial layer part, the mixed boundary and initial layer part, and the error parts involving nonlinearities, respectively, and are defined by the following:

$$
H_{Inn}(x,t) = -\lambda^2 \mathcal{E}^0 \mathcal{E}_x^0,
$$

$$
\begin{aligned}
H_B^\lambda(x,t,\xi,\eta) = &\big((D(x) - D(0)) f(x) E_+^0 + (D(x) - D(1)) g(x) E_-^0 \big) \\
&+ \lambda \Big(f'(x) z_+^1 + g'(x) z_-^1 - \mathcal{E}^0 (f(x) E_{+,\xi}^0 - g(x) E_{-,\eta}^0) \\
&\qquad + (f(x) E_+^0 + g(x) E_-^0)(f(x) E_{+,\xi}^0 - g(x) E_{-,\eta}^0) \Big) \\
&+ \lambda^2 \Big(-\mathcal{E}^0 (f'(x) E_+^0 + g'(x) E_-^0) \\
&\qquad - (f(x) E_+^0 + g(x) E_-^0)(\mathcal{E}_x^0 + f'(x) E_+^0 + g'(x) E_-^0) \Big) \\
= &\Big((D(x) - D(0)) f(x) E_+^i + (D(x) - D(1)) g(x) E_-^i \Big) + \lambda \{\cdots\}_{HB}^R.
\end{aligned}
$$

Note that $\{\cdots\}_{HB}^{R}$ is the sum of the boundary layer functions z_+^1, z_-^1, E_+^0, E_-^0, $E_{+,\xi}^0$, $E_{-,\eta}^0$, $E_+^0E_+^0$, $E_-^0E_-^0$, $E_+^0E_-^0$, $E_+^0E_{+,\xi}^0$, $E_-^0E_{-,\eta}^0$, $E_+^0E_{-,\eta}^0$, and $E_-^0E_{+,\xi}^0$ with the coefficients consisting of $D(x)$, $f(x)$, $g(x)$, $\mathcal{E}^0$, $f'(x)$, $g'(x)$, and $\mathcal{E}_x^0$, and that $\{\cdots\}_{HB}^{R}$ does not depend upon the fast dielectric relaxation time scale. Hence, by (99) and (100), we easily obtain that there exists a constant M, independent of λ, such that

$$\|\{\cdots\}_{HB}^{R}(t)\|_{L_x^2}^2 + \int_0^t \|\partial_t\{\cdots\}_{HB}^{R}(t)\|_{L_x^2}^2 dt \le M\lambda, \tag{106}$$

$$\begin{aligned} H_I^\lambda(x,s) &= \lambda^2 z_{I,x}^2 - \lambda^2\big(\mathcal{E}^0E_I^0 + E_I^0(\mathcal{E}_x^0 + E_{I,x}^0)\big),\\ H_{IB}^\lambda(x,\xi,\eta,t,s) &= -\lambda\Big(E_I^0(f(x)E_{+,\xi}^0 - g(x)E_{-,\eta}^0)\Big)\\ &\quad -\lambda^2\Big(E_I^0(f'(x)E_+^0 + g'(x)E_-^0) + (f(x)E_+^0 + g(x)E_-^0)E_{I,x}^0\Big), \end{aligned}$$

$$\begin{aligned} H_R^\lambda &= -\lambda E_R^\lambda(f(x)E_{+,\xi}^0 - g(x)E_{-,\eta}^0)\\ &\quad -\lambda^2\Big((\mathcal{E}^0 + f(x)E_+^0 + g(x)E_-^0)E_{R,x}^\lambda + (\mathcal{E}_x^0 + f'(x)E_+^0 + g'(x)E_-^0)E_R^\lambda\Big)\\ &\quad -\lambda^2\big(E_I^0E_{R,x}^\lambda + E_{I,x}^0E_R^\lambda\big) - \lambda^2E_R^\lambda E_{R,x}^\lambda, \end{aligned}$$

$$G_{Inn}(x,t) = -\lambda^2\big(\mathcal{E}_t^0 - \mathcal{E}_{xx}^0\big),$$

$$\begin{aligned} G_B^\lambda(x,\xi,\eta,t) &= \Big(-f(x)(\mathcal{Z}^0(x,t) - \mathcal{Z}^0(0,t))E_+^0 - g(x)(\mathcal{Z}^0(x,t) - \mathcal{Z}^0(1,t))E_-^0\Big)\\ &\quad +\lambda\{\cdots\}_{GB}^{R}. \end{aligned}$$

Here $\{\cdots\}_{GB}^{R}$ is the sum of the boundary layer functions $z_{+,t}^1$, $z_{-,t}^1$, E_+^0, E_-^0, $E_{+,\xi}^0$, $E_{-,\eta}^0$, $z_+^1E_+^0$, $z_-^1E_-^0$, $z_+^1E_-^0$, and $z_-^1E_+^0$ with the coefficients consisting of $f(x)$, $g(x)$, $\mathcal{E}^0$, $f'(x)$, $g'(x)$, $f''(x)$, $g''(x)$, and $\mathcal{Z}^0$. Like $\{\cdots\}_{HB}^{R}$, $\{\cdots\}_{GB}^{R}$ does not depend upon the fast dielectric relaxation time scale, and hence it easily follows from (99) and (100) that there exists a constant M, independent of λ, such that

$$\|\{\cdots\}_{GB}^{R}(t)\|_{L_x^2}^2 + \int_0^t \|\partial_t\{\cdots\}_{GB}^{R}(t)\|_{L_x^2}^2 dt \le M\lambda, \tag{107}$$

$$\begin{aligned} G_I^\lambda &= (\mathcal{Z}^0 - \mathcal{Z}^0(x,0))E_I^0 + \lambda^2E_{I,xx}^0 + \lambda^2z_I^2(\mathcal{E}^0 + E_I^0),\\ G_{IB}^\lambda &= -\lambda(f(x)z_+^1 + g(x)z_-^1)E_I^0 - \lambda^2z_I^2(f(x)E_+^0 + g(x)E_-^0),\\ G_R^\lambda &= -(\mathcal{E}^0 + f(x)E_+^0 + g(x)E_-^0 + E_I^0)z_R^\lambda\\ &\quad -\lambda(f(x)z_+^1 + g(x)z_-^1)E_R^\lambda - \lambda^2z_I^2E_R^\lambda - z_R^\lambda E_R^\lambda. \end{aligned}$$

We now derive the boundary conditions for the error functions.

First, the assumption

$$E_0^0(x=0,1) = -\frac{D_x(x=0,1)}{z_0^0(x=0,1)} = \mathcal{E}^0(x=0,1;t=0),$$

together with the initial layer function (88), gives

$$E_I^0(x=0,1;t)=0, \quad t>0. \tag{108}$$

Then it follows from (17), (68), and (108) that

$$E_R^\lambda(x=0,1;t)=0, \quad t>0. \tag{109}$$

Next we claim that

$$H^\lambda(x=0,1;t)=0, \quad t>0. \tag{110}$$

In fact, we can rewrite $H^\lambda(x,t)$ as

$$\begin{aligned} H^\lambda &= z_{R,x}^\lambda + DE_R^\lambda + H_{Inn} + H_B^\lambda + H_I^\lambda + H_{IB}^\lambda + H_R^\lambda \\ &= z_{R,x}^\lambda + \lambda(f'(x)z_+^1 + g'(x)z_-^1) + \lambda^2 z_{I,x}^2 + f(x)(D(x)-D(0))E_+^0 \\ &\quad + g(x)(D(x)-D(1))E_-^0 + DE_R^\lambda - \lambda^2 E^\lambda E_x^\lambda. \end{aligned} \tag{111}$$

Then, by the definitions of cut-off functions $f(x)$ and $g(x)$, the boundary condition $E^\lambda(x=0,1;t)=0$, and (109), one gets from (111) that

$$H^\lambda(x=0,1;t)=(z_{R,x}^\lambda + \lambda^2 z_I^2)|_{x=0,1}. \tag{112}$$

Also, replacing z^λ by (103) in the boundary condition $z_x^\lambda(x=0,1;t)=0$ and using

$$z_{+,\xi}^1(\xi=0;t) = -\mathcal{Z}_x^0(x=0;t), \quad z_{-,\eta}^1(\eta=0;t)=\mathcal{Z}_x^0(x=1;t), \quad t>0,$$

one gets

$$(z_{R,x}^\lambda + \lambda^2 z_I^2)|_{x=0,1}=0,$$

which, together with (112), gives (110).

Now we start the energy estimates. In the following, we use c_i, δ_i, ϵ, and $M(\epsilon)$ or M to denote the constants which are independent of λ and may differ from one line to another.

First we derive the basic energy estimates on $(z_R^\lambda, E_R^\lambda)$.

LEMMA 6. *Under the assumptions of Theorem 3, we have*

$$\begin{aligned} &\|z_R^\lambda(t)\|_{L_x^2}^2 + \lambda^2\|E_R^\lambda(t)\|_{L_x^2}^2 + \int_0^t \|(z_{R,x}^\lambda, E_R^\lambda)\|_{L_x^2}^2 dt + \lambda^2 \int_0^t \|E_{R,x}^\lambda\|_{L_x^2}^2 dt \\ &\quad \le \|z_R^\lambda(x,0)\|_{L_x^2}^2 + \lambda^2 \|E_R^\lambda(x,0)\|_{L_x^2}^2 \\ &\qquad + M\int_0^t \|z_R^\lambda\|_{L_x^2}^2 dt + M\lambda^4 \int_0^t \|(E_R^\lambda, E_{R,x}^\lambda)\|_{L_x^2}^2 \|E_{R,x}^\lambda\|_{L_x^2}^2 dt \\ &\qquad + M\int_0^t \|(z_R^\lambda, z_{R,x}^\lambda)\|_{L_x^2}^2 \|E_R^\lambda\|_{L_x^2}^2 dt + M\lambda. \end{aligned} \tag{113}$$

Proof of Lemma 6. Multiplying (104) by z_R^λ and integrating the resulting equation over $[0,1]$ with respect to x, by (110) and integrations by parts one gets

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|z_R^\lambda\|_{L_x^2}^2 &= -\int_0^1 H^\lambda z_{R,x}^\lambda dx + \int_0^1 f^\lambda z_R^\lambda dx \\ &= -\int_0^1 (z_{R,x}^\lambda + DE_R^\lambda) z_{R,x}^\lambda dx + \int_0^1 f^\lambda z_R^\lambda dx \\ &\quad - \int_0^1 (H_{Inn} + H_B^\lambda + H_I^\lambda + H_{IB}^\lambda + H_R^\lambda) z_{R,x}^\lambda dx. \end{aligned} \tag{114}$$

Now we estimate each term in the right-hand side of (114).

First, by the Cauchy–Schwarz inequality and the properties of the boundary layers, one gets

$$-\int_0^1 (z^\lambda_{R,x} + DE^\lambda_R) z^\lambda_{R,x} dx \le -\frac{1}{2}\|z^\lambda_{R,x}\|^2_{L^2_x} + M(\epsilon)\|E^\lambda_R\|^2_{L^2_x} \tag{115}$$

and

$$\int_0^1 f^\lambda z^\lambda_R dx \le M\|z^\lambda_R\|^2_{L^2_x} + M\|f^\lambda\|^2_{L^2_x} dx \le M\|z^\lambda_R\|^2_{L^2_x} + M\lambda^3. \tag{116}$$

Here we used $\|(z^1_{+,t}, z^1_{-,t})\|^2_{L^2_x} \le M\lambda$ due to (100).

Then, using the regularity of inner solutions, the properties (99) and (100) of boundary layer functions, the properties (101) and (102) of initial layer functions, and the definitions of H_{Inn}, H^λ_B, H^λ_I, and H^λ_{IB}, one easily gets

$$\|H_{Inn}\|^2_{L^2_x} + \|H^\lambda_B\|^2_{L^2_x} + \|H^\lambda_I\|^2_{L^2_x} + \|H^\lambda_{IB}\|^2_{L^2_x} \le M\lambda.$$

This, combined with the Cauchy–Schwarz inequality, yields

$$\begin{aligned} &-\int_0^1 (H_{Inn} + H^\lambda_B + H^\lambda_I + H^\lambda_{IB}) z^\lambda_{R,x} dx \\ &\le \epsilon\|z^\lambda_{R,x}\|^2_{L^2_x} + M(\epsilon)(\|H_{Inn}\|^2_{L^2_x} + \|H^\lambda_B\|^2_{L^2_x} + \|H^\lambda_I\|^2_{L^2_x} + \|H^\lambda_{IB}\|^2_{L^2_x}) \\ &\le \epsilon\|z^\lambda_{R,x}\|^2_{L^2_x} + M\lambda. \end{aligned} \tag{117}$$

Finally, for the nonlinear term, using $\mathcal{E}^0, \mathcal{E}^0_x \in C^0([0,1]\times[0,T])$, (99), and (101), one gets, with the aid of the Cauchy–Schwarz inequality and Sobolev's lemma, that

$$\begin{aligned} &\int_0^1 H^\lambda_R z^\lambda_{R,x} dx \\ &\quad \le \epsilon\|z^\lambda_{R,x}\|^2_{L^2_x} + M(\epsilon)\|H^\lambda_R\|^2_{L^2_x} \\ &\quad \le \epsilon\|z^\lambda_{R,x}\|^2_{L^2_x} + M\lambda^2\|E^\lambda_R\|^2_{L^2_x} + M\lambda^4\|E^\lambda_{R,x}\|^2_{L^2_x} + M\lambda^4\int_0^1 |E^\lambda_R E^\lambda_{R,x}|^2 dx \\ &\quad \le \epsilon\|z^\lambda_{R,x}\|^2_{L^2_x} + M\lambda^2\|E^\lambda_R\|^2_{L^2_x} + M\lambda^4\|E^\lambda_{R,x}\|^2_{L^2_x} \\ &\qquad + M\lambda^4\|(E^\lambda_R, E^\lambda_{R,x})\|^2_{L^2_x}\|E^\lambda_{R,x}\|^2_{L^2_x}. \end{aligned} \tag{118}$$

Thus, combining (114) with (115)–(118) and taking ϵ small enough, one gets

$$\begin{aligned} \frac{d}{dt}\|z^\lambda_R\|^2_{L^2_x} + c_1\|z^\lambda_{R,x}\|^2_{L^2_x} &\le M\|(z^\lambda_R, E^\lambda_R)\|^2_{L^2_x} + M\lambda^4\|E^\lambda_{R,x}\|^2_{L^2_x} \\ &\quad + M\lambda^4(\|(E^\lambda_R, E^\lambda_{R,x})\|^2_{L^2_x}\|E^\lambda_{R,x}\|^2_{L^2_x} + M\lambda. \end{aligned} \tag{119}$$

Integrating (119) with respect to t over $[0,t]$, one gets

$$\begin{aligned} &\|z^\lambda_R(t)\|^2_{L^2_x} + c_1\int_0^t \|z^\lambda_{R,x}\|^2_{L^2_x} dt \\ &\quad \le \|z^\lambda_R(x,0)\|^2_{L^2_x} + M\int_0^t \|(z^\lambda_R, E^\lambda_R)\|^2_{L^2_x} dt + M\lambda^4\int_0^t \|E^\lambda_{R,x}\|^2_{L^2_x} dt \\ &\qquad + M\lambda^4\int_0^t \|(E^\lambda_R, E^\lambda_{R,x})\|^2_{L^2_x}\|E^\lambda_{R,x}\|^2_{L^2_x} dt + M\lambda. \end{aligned} \tag{120}$$

Multiplying (105) by E_R^λ and integrating the resulting equation over $[0,1]$ with respect to x, by (109) and integrations by parts one gets

$$\frac{\lambda^2}{2}\frac{d}{dt}\|E_R^\lambda\|_{L_x^2}^2 + \lambda^2\|E_{R,x}^\lambda\|_{L_x^2}^2 + \int_0^1 \mathcal{Z}^0|E_R^\lambda|^2dx = \int_0^1 g^\lambda E_R^\lambda dx. \tag{121}$$

By the Cauchy–Schwarz inequality, we have

$$\int_0^1 g^\lambda E_R^\lambda dx \le \epsilon\|E_R^\lambda\|_{L_x^2}^2 + M(\epsilon)\|g^\lambda\|_{L_x^2}^2.$$

On one hand, noting that

$$\begin{aligned}
&\int_0^1 \left|(\mathcal{Z}^0(x,t) - \mathcal{Z}^0(x,0))E_I^0\left(x, \frac{t}{\lambda^2}\right)\right|^2 dx\\
&= \int_0^1 \left|\int_0^1 \partial_t\mathcal{Z}^0(x,t\theta)d\theta t \cdot E_I^0\left(x, \frac{t}{\lambda^2}\right)\right|^2 dx\\
&\le M\sup_{s\ge 0}\left(\max_{0\le x\le 1}|sE_I^0(x,s)|^2\right)\lambda^4\\
&\le M\lambda^4,
\end{aligned}$$

and using $\mathcal{Z}^0, \mathcal{E}^0 \in C^{2,1}([0,1]\times[0,T])$, (99), (100), (101), and (102), and the definitions of G_{Inn}, G_B^λ, G_I^λ, and G_{IB}^λ, we have

$$\int_0^1 (|G_{Inn}|^2 + |G_B^\lambda|^2 + |G_I^\lambda|^2 + |G_{IB}^\lambda|^2)dx \le M\lambda.$$

On the other hand, as in (118), with the aid of Sobolev's lemma we have that

$$\|G_R^\lambda\|_{L_x^2}^2 \le M\|z_R^\lambda\|_{L_x^2}^2 + M\lambda^2\|E_R^\lambda\|_{L_x^2}^2 + M\|(z_R^\lambda, z_{R,x}^\lambda)\|_{L_x^2}^2\|E_R^\lambda\|_{L_x^2}^2.$$

Thus

$$\int_0^1 g^\lambda E_R^\lambda dx \le \epsilon\|E_R^\lambda\|_{L_x^2}^2 + M\|z_R^\lambda\|_{L_x^2}^2 + M\lambda^2\|E_R^\lambda\|_{L_x^2}^2 + M\|(z_R^\lambda, z_{R,x}^\lambda)\|_{L_x^2}^2\|E_R^\lambda\|_{L_x^2}^2 + M\lambda. \tag{122}$$

Then, combining (121) with (122), using the positivity of $\mathcal{Z}^0$, taking ϵ small enough, and then restricting λ to be small enough, one gets

$$\begin{aligned}
&\lambda^2\frac{d}{dt}\|E_R^\lambda\|_{L_x^2}^2 + \lambda^2\|E_{R,x}^\lambda\|_{L_x^2}^2 + c_2\|E_R^\lambda\|_{L_x^2}^2\\
&\quad\le M\|z_R^\lambda\|_{L_x^2}^2 + M\|(z_R^\lambda, z_{R,x}^\lambda)\|_{L_x^2}^2\|E_R^\lambda\|_{L_x^2}^2 + M\lambda.
\end{aligned} \tag{123}$$

Integrating (123) with respect to t, one gets

$$\begin{aligned}
&\lambda^2\|E_R^\lambda(t)\|_{L_x^2}^2 + \lambda^2\int_0^t \|E_{R,x}^\lambda\|_{L_x^2}^2 dt + c_2\int_0^t \|E_R^\lambda\|_{L_x^2}^2 dt\\
&\quad\le \lambda^2\|E_R^\lambda(x,0)\|_{L_x^2}^2 + M\int_0^t \|z_R^\lambda\|_{L_x^2}^2 dt\\
&\qquad + M\int_0^t \|(z_R^\lambda, z_{R,x}^\lambda)\|_{L_x^2}^2\|E_R^\lambda\|_{L_x^2}^2 dt + M\lambda.
\end{aligned} \tag{124}$$

The desired estimate (113) follows from (120) and (124). This completes the proof of Lemma 6. ∎

Next we show the estimates of the time derivatives $\partial_t(z_R^\lambda, E_R^\lambda)$ of $(z_R^\lambda, E_R^\lambda)$.

LEMMA 7. *Under the assumptions of Theorem 3, we have*

(125)

$$\begin{aligned}
&\|z_{R,t}^\lambda(t)\|_{L_x^2}^2 + \lambda^2\|E_{R,t}^\lambda(t)\|_{L_x^2}^2 + \int_0^t \|(z_{R,xt}^\lambda, E_{R,t}^\lambda)\|_{L_x^2}^2 dt + \lambda^2\int_0^t \|E_{R,xt}^\lambda\|_{L_x^2}^2 dt\\
&\le M(\|z_{R,t}^\lambda(x,0)\|_{L_x^2}^2 + \lambda^2\|E_{R,t}^\lambda(x,0)\|_{L_x^2}^2)\\
&+ M\int_0^t \|(E_R^\lambda, z_R^\lambda, z_{R,t}^\lambda)\|_{L_x^2}^2 dt + M\lambda^2\int_0^t \|E_{R,t}^\lambda\|_{L_x^2}^2 dt + M\lambda^4\int_0^t \|(E_{R,x}^\lambda, E_{R,xt}^\lambda)\|_{L_x^2}^2 dt\\
&+ M\lambda^4\int_0^t \big(\|(E_{R,t}^\lambda, E_{R,xt}^\lambda)\|_{L_x^2}^2\|E_{R,x}\|_{L_x^2}^2 + \|E_R^\lambda\|_{L_x^2}^2\|E_{R,xt}^\lambda\|_{L_x^2}^2\big)dt\\
&+ M\int_0^t \big(\|(z_{R,t}^\lambda, z_{R,xt}^\lambda)\|_{L_x^2}^2\|E_R^\lambda\|_{L_x^2}^2 + \|(z_R^\lambda, z_{R,x}^\lambda)\|_{L_x^2}^2\|E_{R,t}^\lambda\|_{L_x^2}^2\big)dt + M\lambda.
\end{aligned}$$

Proof of Lemma 7. Differentiating (104) with respect to t, multiplying the resulting equations by $z_{R,t}^\lambda$, then integrating it over $[0,1]\times[0,t]$ and noting that H_t^λ also satisfies the same boundary condition as in (110), one gets by integration by parts that

$$\begin{aligned}
\|z_{R,t}^\lambda(t)\|_{L_x^2}^2 = \|z_{R,t}^\lambda(x,0)\|_{L_x^2}^2 &+ \int_0^t\int_0^1 f_t^\lambda z_{R,t}^\lambda dxdt - \int_0^t\int_0^1 (z_{R,xt}^\lambda + DE_{R,t}^\lambda)z_{R,xt}^\lambda dxdt\\
&- \int_0^t\int_0^1 H_{Inn,t} z_{R,xt}^\lambda dxdt - \int_0^t\int_0^1 H_{B,t}^\lambda z_{R,xt}^\lambda dxdt - \int_0^t\int_0^1 H_{I,t}^\lambda z_{R,xt}^\lambda dxdt\\
&- \int_0^t\int_0^1 H_{IB,t}^\lambda z_{R,xt}^\lambda dxdt - \int_0^t\int_0^1 H_{R,t}^\lambda z_{R,xt}^\lambda dxdt.
\end{aligned}\tag{126}$$

One needs to estimate the terms on the right in the above carefully.

First, it follows from the Cauchy–Schwarz inequality that

$$\int_0^t\int_0^1 (z_{R,xt}^\lambda + DE_{R,t}^\lambda)z_{R,xt}^\lambda dxdt \le -\frac{1}{2}\int_0^t \|z_{R,xt}^\lambda\|_{L_x^2}^2 dt + M\int_0^t \|E_{R,t}^\lambda\|_{L_x^2}^2 dt \tag{127}$$

and

$$\int_0^t\int_0^1 f_t^\lambda z_{R,t}^\lambda dxdt \le \epsilon\int_0^t \|z_{R,t}^\lambda\|_{L_x^2}^2 dt + M(\epsilon)\lambda^3, \tag{128}$$

since f^λ does not depend upon the fast time scale. Here we also used $\mathcal{Z}_{tt}^0, \mathcal{E}_{tt}^0 \in C^0([0,1]\times[0,T])$.

Similarly, since H_{Inn} and H_B^λ do not depend upon the fast time scale, $H_{Inn,t}$ and $H_{B,t}^\lambda$ have the same structures as H_{Inn} and $H_{B,t}$, respectively. Hence, using (106) one obtains in a similar way as for (117) that

$$\int_0^t\int_0^1 H_{Inn,t} z_{R,xt}^\lambda dxdt \le \epsilon\int_0^t \|z_{R,xt}^\lambda\|_{L_x^2}^2 dt + M\lambda^4 \tag{129}$$

and

$$\int_0^t\int_0^1 H_{B,t}^\lambda z_{R,xt}^\lambda dxdt \le \epsilon\int_0^t \|z_{R,xt}^\lambda\|_{L_x^2}^2 dt + M\lambda. \tag{130}$$

Here we used the regularity of the inner solutions $\mathcal{E}^0$, $\mathcal{E}^0_x$, $\mathcal{E}^0_t$, $\mathcal{E}^0_{xt}$, $\mathcal{Z}^0$, $\mathcal{Z}^0_x$, $\mathcal{Z}^0_t$, $\mathcal{Z}^0_{xt} \in C^0([0,1]\times[0,T])$.

Owing to the strong singularity of time derivatives of the initial layers, we must estimate the integrals involving initial layers carefully.

First, by the Cauchy–Schwarz inequality, we have

$$\int_0^t \int_0^1 H^\lambda_{I,t} z^\lambda_{R,xt} dxdt \le \epsilon \int_0^t \|z^\lambda_{R,xt}\|^2_{L^2_x} dt + M(\epsilon) \int_0^t \int_0^1 |H^\lambda_{I,t}|^2 dxdt.$$

But, by (101) and (102), one gets

$$\begin{aligned}
&\int_0^t \int_0^1 |H^\lambda_{I,t}|^2 dxdt \\
&\quad = \int_0^t \int_0^1 |z^2_{I,xs} - \lambda^2\big(\mathcal{E}^0_t(x,t)E^0_{I,x} + E^0_I \mathcal{E}^0_{xt}(x,t)\big) \\
&\qquad - \big(\mathcal{E}^0(x,t)E^0_{I,xs} + E^0_{I,s}(\mathcal{E}^0_x(x,t) + E^0_{I,x}) + E^0_I E^0_{I,xs}\big)|^2 dxdt \\
&\quad \le M \int_0^t \int_0^1 \Big(|z^2_{I,xs}|^2 + |E^0_{I,x}|^2 + |E^0_I|^2 + |E^0_{I,xs}|^2 + |E^0_{I,s}|^2\Big) dxdt \\
&\quad \le M\lambda^2.
\end{aligned}$$

Hence

$$\int_0^t \int_0^1 H^\lambda_{I,t} z^\lambda_{R,xt} dxdt \le \epsilon \int_0^t \|z^\lambda_{R,xt}\|^2_{L^2_x} + M\lambda^2. \tag{131}$$

Then, using the definition of H^λ_{IB}, we have

$$\begin{aligned}
&\int_0^t \int_0^1 H^\lambda_{IB,t} z^\lambda_{R,tx} dxdt \\
&\quad = \int_0^t \int_0^1 -\lambda\Big(E^0_I(f(x)E^0_{+,\xi} - g(x)E^0_{-,\eta})\Big)_t z^\lambda_{R,tx} dxdt \\
&\qquad + \int_0^t \int_0^1 \lambda^2 \partial_t \Big(\cdots\Big)^R_{HIB} z^\lambda_{R,xt} dxdt,
\end{aligned} \tag{132}$$

where $(\cdots)^R_{HIB}$ represents the remaining higher order term $O(\lambda^2)$ of H^λ_{IB}. By (101) and (102), one easily gets

$$\int_0^t \int_0^1 |\lambda^2 \partial_t \Big(\cdots\Big)^R_{HIB}|^2 dxdt \le M\lambda^3,$$

which leads to

$$\begin{aligned}
&\int_0^t \int_0^1 \lambda^2 \partial_t \Big(\cdots\Big)^R_{HIB} z^\lambda_{R,tx} dxdt \\
&\quad \le \epsilon \int_0^t \|z^\lambda_{R,xt}\|^2_{L^2_x} dt + M(\epsilon) \int_0^t \int_0^1 |\lambda^2 \partial_t \Big(\cdots\Big)^R_{HIB}|^2 dxdt \\
&\quad \le \epsilon \int_0^t \|z^\lambda_{R,xt}\|^2_{L^2_x} dt + M\lambda^3.
\end{aligned} \tag{133}$$

It remains to control the first term on the right-hand side of (132). Note that this singular integration is caused by the interactions between the boundary layer and the initial layer. So, to control it, we must use twofold integrals in the time and space directions to cancel the oscillation of the electric field. Indeed, it can be treated as follows:

$$
\begin{aligned}
&-\int_0^t\int_0^1 \lambda\Big(E_I^0(f(x)E^0_{+,\xi}-g(x)E^0_{-,\eta})\Big)_t z^\lambda_{R,tx}dxdt\\
&\quad= -\frac{1}{\lambda}\int_0^t\int_0^1 \Big(E^0_{I,s}(f(x)E^0_{+,\xi}-g(x)E^0_{-,\eta})\Big) z^\lambda_{R,tx}dxdt\\
&\qquad-\int_0^t\int_0^1 \lambda\Big(E_I^0(f(x)E^0_{+,\xi t}-g(x)E^0_{-,\eta t})\Big) z^\lambda_{R,tx}dxdt\\
&\quad\le \epsilon\int_0^t \|z^\lambda_{R,xt}\|^2_{L^2_x}dt+\frac{1}{\lambda^2}\int_0^t\int_0^1 |E^0_{I,s}(f(x)E^0_{+,\xi}-g(x)E^0_{-,\eta})|^2dxdt+M\lambda^5\\
&\quad\le \epsilon\int_0^t \|z^\lambda_{R,xt}\|^2_{L^2_x}dt+M\lambda, \qquad (134)
\end{aligned}
$$

where we have used

$$
\begin{aligned}
&\frac{1}{\lambda^2}\int_0^t\int_0^1 |E^0_{I,s}(f(x)E^0_{+,\xi}+g(x)E^0_{-,\eta})|^2dxdt\\
&\quad\le \frac{M}{\lambda^2}\int_0^t\int_0^1 |E^0_{I,s}|^2(|E^0_{+,\xi}|^2+|E^0_{-,\eta}|^2)dxdt\\
&\quad\le \frac{M}{\lambda^2}\int_0^t \max_{0\le x\le 1}|E^0_{I,s}|^2dt\Big(\int_0^1\max_{0\le t\le T}|E^0_{+,\xi}|^2dx+\int_0^1\max_{0\le t\le T}|E^0_{+,\xi}|^2dx\Big)\\
&\quad\le M\lambda.
\end{aligned}
$$

Combining (132) with (133) and (134), one gets

$$
\int_0^t\int_0^1 H^\lambda_{IB,t}z^\lambda_{R,xt}dxdt\le\int_0^t\|z^\lambda_{R,xt}\|^2_{L^2_x}dt+M\lambda. \qquad (135)
$$

Finally, we estimate the last integral on the right-hand side of (126). We split it into five parts:

$$
\int_0^t\int_0^1 H^\lambda_{R,t}z^\lambda_{R,xt}dxdt = I_1+I_2+I_3+I_4+I_5, \qquad (136)
$$

where

$$
\begin{aligned}
I_1=-\lambda\int_0^t\int_0^1\Big\{&\big(f(x)E^0_{+,\xi}-g(x)E^0_{-,\eta}\big)E^\lambda_{R,t}\\
&+\big(f(x)E^0_{+,\xi t}-g(x)E^0_{-,\eta t}\big)E^\lambda_R\Big\}z^\lambda_{R,xt}dxdt,
\end{aligned}
$$

$$
\begin{aligned}
I_2=-\lambda^2\int_0^t\int_0^1&\Big\{\big(\mathcal{E}^0+f(x)E^0_++g(x)E^0_-\big)E^\lambda_{R,xt}\\
&+\big(\mathcal{E}^0_t+f(x)E^0_{+,t}+g(x)E^0_{-,t}\big)E^\lambda_{R,x}+\big(\mathcal{E}^0_x+f'(x)E^0_++g'(x)E^0_-\big)E^\lambda_{R,t}\\
&+\big(\mathcal{E}^0_{xt}+f'(x)E^0_{+,t}+g'(x)E^0_{-,t}\big)E^\lambda_R\Big\}z^\lambda_{R,xt}dxdt,
\end{aligned}
$$

$$I_3 = -\lambda^2 \int_0^t \int_0^1 (E_I^0 E_{R,xt}^\lambda + E_{I,x}^0 E_{R,t}^\lambda) z_{R,xt}^\lambda dxdt,$$

$$I_4 = -\lambda^2 \int_0^t \int_0^1 (E_{R,t}^\lambda E_{R,x}^\lambda + E_R^\lambda E_{R,xt}^\lambda) z_{R,xt}^\lambda dxdt,$$

$$I_5 = -\int_0^t \int_0^1 (E_{I,s}^0 E_{R,x}^\lambda + E_{I,xs}^0 E_R^\lambda) z_{R,xt}^\lambda dxdt.$$

Noting that there are an λ factor in the first term I_1 of (136) and an λ^2 factor in the second and third terms I_2, I_3 of (136), by the Cauchy–Schwarz inequality, (99), (101), and the fact that $\mathcal{E}^0, \mathcal{E}_x^0, \mathcal{E}_t^0, \mathcal{E}_{xt}^0 \in C^0([0,1]\times[0,T])$, we have

$$I_1 \le \epsilon \int_0^t \|z_{R,xt}^\lambda\|_{L_x^2}^2 dt + M(\epsilon)\lambda^2 \int_0^t \|(E_R^\lambda, E_{R,t}^\lambda)\|_{L_x^2}^2 dt, \tag{137}$$

$$I_2 \le \epsilon \int_0^t \|z_{R,xt}^\lambda\|_{L_x^2}^2 dt + M(\epsilon)\lambda^4 \int_0^t \|(E_R^\lambda, E_{R,x}^\lambda, E_{R,t}^\lambda, E_{R,xt}^\lambda)\|_{L_x^2}^2 dt, \tag{138}$$

and

$$I_3 \le \epsilon \int_0^t \|z_{R,xt}^\lambda\|_{L_x^2}^2 dt + M(\epsilon)\lambda^4 \int_0^t \|(E_{R,t}^\lambda, E_{R,xt}^\lambda)\|_{L_x^2}^2 dt. \tag{139}$$

For the nonlinear term I_4 of (136), one gets by Sobolev's lemma that

$$\begin{aligned} I_4 &\le \epsilon \int_0^t \|z_{R,xt}^\lambda\|_{L_x^2}^2 + M(\epsilon)\lambda^4 \int_0^t \int_0^1 (|E_{R,t}^\lambda E_{R,x}^\lambda|^2 + |E_R^\lambda E_{R,xt}^\lambda|^2) dxdt \\ &\le \epsilon \int_0^t \|z_{R,xt}^\lambda\|_{L_x^2}^2 + M\lambda^4 \int_0^t (\|E_{R,t}^\lambda\|_{L_x^\infty}^2 \|E_{R,x}^\lambda\|_{L_x^2}^2 + \|E_R^\lambda\|_{L_x^\infty}^2 \|E_{R,xt}^\lambda\|_{L_x^2}^2) dt \\ &\le \epsilon \int_0^t \|z_{R,xt}^\lambda\|_{L_x^2}^2 \\ &\quad + M\lambda^4 \int_0^t \big(\|(E_{R,t}^\lambda, E_{R,xt}^\lambda)\|_{L_x^2}^2 \|E_{R,x}^\lambda\|_{L_x^2}^2 + \|E_R^\lambda\|_{L_x^2}^2 \|E_{R,xt}^\lambda\|_{L_x^2}^2) dt. \end{aligned} \tag{140}$$

It remains to estimate I_5. This is more difficult due to the lack of the uniform L^2 estimate of $E_{R,x}^\lambda$. This will be achieved by using the uniform boundedness on $\|s(E_{I,s}^0, E_{I,xs}^0)\|_{L_{(x,t)}^\infty([0,1]\times[0,T])}$ and employing Hardy–Littlewood's inequality. In fact, by the Cauchy–Schwarz inequality, one gets

(141)

$$\begin{aligned} I_5 &\le \epsilon \int_0^t \|z_{R,xt}^\lambda\|_{L_x^2}^2 dt + M(\epsilon) \int_0^t \int_0^1 |E_{I,s}^0 E_{R,x}^\lambda + E_{I,xs}^0 E_R^\lambda|^2 dxdt \\ &= \epsilon \int_0^t \|z_{R,xt}^\lambda\|_{L_x^2}^2 dt + M(\epsilon) \int_0^t \int_0^1 \Bigg| t\Bigg(E_{I,s}^0 \frac{E_{R,x}^\lambda - E_{R,x}^\lambda(x,0)}{t} \\ &\qquad + E_{I,xs}^0 \frac{E_R^\lambda - E_R^\lambda(x,0)}{t} \Bigg) + \big(E_{I,s}^0 E_{R,x}^\lambda(x,0) + E_{I,xs}^0 E_R^\lambda(x,0)\big) \Bigg|^2 dxdt \end{aligned}$$

$$
\begin{aligned}
&= \epsilon \int_0^t \|z^\lambda_{R,xt}\|^2_{L^2_x} dt + M(\epsilon) \int_0^t \int_0^1 \Bigg| \lambda^2 s \Bigg(E^0_{I,s} \frac{E^\lambda_{R,x} - E^\lambda_{R,x}(x,0)}{t} \\
&\quad + E^0_{I,xs} \frac{E^\lambda_R - E^\lambda_R(x,0)}{t} \Bigg) + (E^0_{I,s} E^\lambda_{R,x}(x,0) + E^0_{I,xs} E^\lambda_R(x,0)) \Bigg|^2 dxdt \\
&\le \epsilon \int_0^t \|z^\lambda_{R,xt}\|^2_{L^2_x} dt \\
&\quad + M\lambda^4 \max_{0\le s\le\infty} \max_{0\le x\le 1} (s|E^0_{I,s}| + s|E^0_{I,xs}|)^2 \int_0^1 \int_0^t \Bigg(\Big|\frac{E^\lambda_{R,x} - E^\lambda_{R,x}(x,0)}{t}\Big|^2 \\
&\quad + \Big|\frac{E^\lambda_R - E^\lambda_R(x,0)}{t}\Big|^2 \Bigg) dtdx + M \int_0^t \int_0^1 (|E^0_{I,s} E^\lambda_{R,x}(x,0)|^2 + |E^0_{I,xs} E^\lambda_R(x,0)|^2) dxdt \\
&\le \epsilon \int_0^t \|z^\lambda_{R,xt}\|^2_{L^2_x} dt + M\lambda^4 \int_0^1 \left(\int_0^t |E^\lambda_{R,xt}|^2 dt + \int_0^t |E^\lambda_{R,t}|^2 dt \right) dx + M\lambda \\
&\le \epsilon \int_0^t \|z^\lambda_{R,xt}\|^2_{L^2_x} dt + M\lambda^4 \int_0^t \|(E^\lambda_{R,xt}, E^\lambda_{R,t})\|^2_{L^2_x} dt + M\lambda.
\end{aligned}
$$

Here we have used

$$
\begin{aligned}
&\int_0^t \int_0^1 (|E^0_{I,s} E^\lambda_{R,x}(x,0)|^2 + |E^0_{I,xs} E^\lambda_R(x,0)|^2) dxdt \\
&\qquad \le M\lambda^2 \left(\|E^\lambda_{0R,x}\|^2_{L^\infty_x} \int_0^t \int_0^1 |E^0_{I,s}|^2 dxdt + \|E^\lambda_{0R}\|^2_{L^\infty_x} \int_0^t \int_0^1 |E^0_{I,xs}|^2 dxdt \right) \\
&\qquad \le M\lambda^4 ((M\lambda^{\frac{1}{2}-2})^2 + (M\lambda^{\frac{1}{2}-1})^2) \\
&\qquad \le M\lambda,
\end{aligned}
$$

due to Sobolev's lemma; (102) and the assumption (50);

$$
\begin{aligned}
&\max_{0\le s\le\infty} \max_{0\le x\le 1} (s|E^0_{I,s}| + s|E^0_{I,xs}|) \\
&\qquad = \max_{0\le t\le T} \max_{0\le x\le 1} \left(\frac{t}{\lambda^2} \left(|E^0_{I,s}\left(x, \frac{t}{\lambda^2}\right)| + |E^0_{I,xs}\left(x, \frac{t}{\lambda^2}\right)| \right) \right) \\
&\qquad \le M \max_{0\le t\le T} \left(\left(\frac{t}{\lambda^2} \right) e^{-\delta \frac{t}{\lambda^2}} \right) \\
&\qquad \le M;
\end{aligned}
$$

and the fact that $(E^\lambda_R - E^\lambda_R(x,0))(x,0) = 0$ and hence $(E^\lambda_{R,x} - E^\lambda_{R,x}(x,0))(x,0) = 0$. We have also applied Hardy–Littlewood's inequality to control $\int_0^t |\frac{E^\lambda_{R,x} - E^\lambda_{R,x}(x,0)}{t}|^2 dt$ and $\int_0^t |\frac{E^\lambda_R - E^\lambda_R(x,0)}{t}|^2 dt$ by $\int_0^t |E^\lambda_{R,xt}|^2 dt$ and $\int_0^t |E^\lambda_{R,t}|^2 dt$, respectively.

Combining (136) with (137)–(142), one gets

$$
\begin{aligned}
&\int_0^t \int_0^1 H^\lambda_{R,t} z^\lambda_{R,xt} dxdt \\
&\quad \le \epsilon \int_0^t \|z^\lambda_{R,xt}\|^2_{L^2_x} dt + M\lambda^2 \int_0^t \|(E^\lambda_R, E^\lambda_{R,t})\|^2_{L^2_x} dt + M\lambda^4 \int_0^t \|(E^\lambda_{R,x}, E^\lambda_{R,xt})\|^2_{L^2_x} dt \\
&\quad + M\lambda^4 \int_0^t \left(\|(E^\lambda_{R,t}, E^\lambda_{R,xt})\|^2_{L^2_x} \|E^\lambda_{R,x}\|^2_{L^2_x} + \|E^\lambda_R\|^2_{L^2_x} \|E^\lambda_{R,xt}\|^2_{L^2_x} \right) dt + M\lambda^2.
\end{aligned} \tag{142}
$$

Therefore, putting (126) and estimates (127), (128), (129), (130), (131), (135), and (142) together and taking ϵ small enough, one shows that

$$\begin{aligned}&\|z^\lambda_{R,t}(t)\|^2_{L^2_x}+c_3\int_0^t\|z^\lambda_{R,xt}\|^2_{L^2_x}dt\\&\quad\le\|z^\lambda_{R,t}(x,0)\|^2_{L^2_x}+M\int_0^t\|(z^\lambda_{R,t},E^\lambda_{R,t})\|^2_{L^2_x}dt\\&\quad\quad+M\lambda^2\int_0^t\|E^\lambda_R\|^2_{L^2_x}dt+M\lambda^4\int_0^t\|(E^\lambda_{R,x},E^\lambda_{R,xt})\|^2_{L^2_x}dt\\&\quad\quad+M\lambda^4\int_0^t\big(\|(E^\lambda_{R,t},E^\lambda_{R,xt})\|^2_{L^2_x}\|E_{R,x}\|^2_{L^2_x}+\|E^\lambda_R\|^2_{L^2_x}\|E^\lambda_{R,xt}\|^2_{L^2_x}\big)dt+M\lambda.\end{aligned}\tag{143}$$

Note that $E^\lambda_{R,t}$ also satisfies the same boundary condition as in (109). Thus, differentiating (104) with respect to t, multiplying the resulting equations by $z^\lambda_{R,t}$, and then integrating it over $[0,1]\times[0,t]$, one gets by integrations by parts that

$$\begin{aligned}&\frac{\lambda^2}{2}\|E^\lambda_{R,t}(t)\|^2_{L^2_x}+\lambda^2\int_0^t\|E^\lambda_{R,xt}\|^2_{L^2_x}dt+\int_0^t\int_0^1\mathcal{Z}^0|E^\lambda_{R,t}|^2dxdt\\&\quad=\frac{\lambda^2}{2}\|E^\lambda_{R,t}(x,0)\|^2_{L^2_x}-\int_0^t\int_0^1\mathcal{Z}^0_tE^\lambda_RE^\lambda_{R,t}dxdt+\int_0^t\int_0^1g^\lambda_tE^\lambda_{R,t}dxdt.\end{aligned}\tag{144}$$

First, by the Cauchy–Schwarz inequality, one gets

$$-\int_0^t\int_0^1\mathcal{Z}^0_tE^\lambda_RE^\lambda_{R,t}dxdt\le\epsilon\int_0^t\|E^\lambda_{R,t}\|^2_{L^2_x}dt+M(\epsilon)\int_0^t\|E^\lambda_R\|^2_{L^2_x}dt\tag{145}$$

and

$$\begin{aligned}&\int_0^t\int_0^1g^\lambda_tE^\lambda_{R,t}dxdt\\&\quad\le\epsilon\int_0^t\|E^\lambda_{R,t}\|^2_{L^2_x}dt+M(\epsilon)\int_0^t\|g^\lambda\|^2_{L^2_x}dt\\&\quad\le\epsilon\int_0^t\|E^\lambda_{R,t}\|^2_{L^2_x}dt+M\Bigg(\int_0^t\int_0^1|G_{Inn,t}|^2dxdt+\int_0^t\int_0^1|G^\lambda_{B,t}|^2dxdt\\&\quad\quad+\int_0^t\int_0^1|G^\lambda_{I,t}|^2dxdt+\int_0^t\int_0^1|G^\lambda_{IB,t}|^2dxdt+\int_0^t\int_0^1|G^\lambda_{R,t}|^2dxdt\Bigg).\end{aligned}\tag{146}$$

Now we treat each term on the right-hand side of (146).

Using the structures of the inner solutions $\mathcal{E}^0_{tt}$, $\mathcal{E}^0_{xxt}\in C^0([0,1]\times[0,T])$, one can get

$$\int_0^t\int_0^1|G_{Inn,t}|^2dxdt\le M\lambda^4.\tag{147}$$

Since $G^\lambda_{B,t}$ has the same structure as G^λ_B, one can estimate the third term of (146) by (107) as

$$\int_0^t\int_0^1|G^\lambda_{B,t}|^2dxdt\le M\lambda.\tag{148}$$

Using the definition of G_I^λ, we have

$$\begin{aligned}
&\int_0^t\int_0^1 |G^\lambda_{I,t}|^2 dxdt\\
&\quad\le \int_0^t\int_0^1 |E^0_{I,xxs}|^2 dxdt + J_{IR}\\
&\quad\le M\lambda^2 + J_{IR}, \qquad (149)
\end{aligned}$$

where

$$\begin{aligned}
J_{IR} = &\int_0^t\int_0^1 |z^2_{I,s}(\mathcal{E}^0+E^0_I) + z^2_I E^0_{I,s}|^2 dxdt + \int_0^t\int_0^1 \big(|\lambda^2 z^2_I \mathcal{E}^0_t|^2 + |\mathcal{Z}^0_t E^0_I|^2\big)dxdt\\
&+\int_0^t\int_0^1 |\lambda^{-2}(\mathcal{Z}^0-\mathcal{Z}^0(x,0))E^0_{I,s}|^2 dxdt.
\end{aligned}$$

Using $\|z^2_I\|_{L^\infty_{x,t}} \le M$ and $\|(z^2_{I,s}, E^0_I, E^0_{I,s})\|_{L^2_t(L^\infty_x)} \le M\lambda$, one gets

$$J_{IR} \le M\lambda^2 + \int_0^t\int_0^1 |\lambda^{-2}(\mathcal{Z}^0-\mathcal{Z}^0(x,0))E^0_{I,s}|^2 dxdt. \qquad (150)$$

To estimate the remaining singular term on the right-hand side of (150), we will use the higher regularity of $\mathcal{Z}^0_t$. It will follow from the mean value theorem and (102) that

$$\begin{aligned}
&\int_0^t\int_0^1 |\lambda^{-2}(\mathcal{Z}^0-\mathcal{Z}^0(x,0))E^0_{I,s}|^2 dxdt\\
&\quad= \int_0^t\int_0^1 \Big|\int_0^1 \mathcal{Z}^0_t(x,t\theta)d\theta \frac{t}{\lambda^2} E^0_{I,s}\left(x,\frac{t}{\lambda^2}\right)\Big|^2 dxdt\\
&\quad\le M\int_0^t\int_0^1 \Big(\frac{t}{\lambda^2}\Big)^2 \Big|E^0_{I,s}\left(x,\frac{t}{\lambda^2}\right)\Big|^2 dxdt\\
&\quad\le M\lambda^2. \qquad (151)
\end{aligned}$$

Thus,

$$J_{IR} \le M\lambda^2. \qquad (152)$$

Therefore,

$$\int_0^t\int_0^1 |G^\lambda_{I,t}|^2 dxdt \le M\lambda^2. \qquad (153)$$

The fifth term of (146) can be treated as in (132) so that

$$\begin{aligned}
&\int_0^t\int_0^1 |G^\lambda_{IB,t}|^2 dxdt\\
&\quad= \int_0^t\int_0^1 \Big|\lambda\Big(\big(f(x)z^1_+ + g(x)z^1_-\big)E^0_I\Big)_t + \lambda^2\partial_t\Big(\cdots\Big)^R_{GIB}\Big|^2 dxdt\\
&\quad\le \frac{M}{\lambda^2}\int_0^t\int_0^1 (|z^1_+E^0_{I,s}|^2 + |z^1_-E^0_{I,s}|^2)dxdt + M\lambda^2\\
&\quad\le M\lambda. \qquad (154)
\end{aligned}$$

To estimate the sixth term of (146), we split it into four parts:

$$\int_0^t \int_0^1 G^\lambda_{R,t} E^\lambda_{R,t} dxdt = I_6 + I_7 + I_8 + I_9, \tag{155}$$

where

$$\begin{aligned} I_6 = -\int_0^t \int_0^1 \big\{ &(\mathcal{E}^0_t + f(x)E^0_{+,t} + g(x)E^0_{-,t}) z^\lambda_R \\ &+ (\mathcal{E}^0 + f(x)E^0_+ + g(x)E^0_- + E^0_I) z^\lambda_{R,t} \\ &+ (\mathcal{Z}^0_t + \lambda(f(x)z^1_{+,t} + g(x)z^1_{-,t})) E^\lambda_R \big\} E^\lambda_{R,t} dxdt, \end{aligned}$$

$$I_7 = -\int_0^t \int_0^1 (\mathcal{Z}^0 + \lambda(f(x)z^1_+ + g(x)z^1_-) + \lambda^2 z^2_I) E^\lambda_{R,t} E^\lambda_{R,t} dxdt,$$

$$I_8 = -\int_0^t \int_0^1 (z^\lambda_{R,t} E^\lambda_R + z^\lambda_R E^\lambda_{R,t}) E^\lambda_{R,t} dxdt,$$

and

$$I_9 = -\frac{1}{\lambda^2} \int_0^t \int_0^1 (E^0_{I,s} z^\lambda_R + \lambda^2 z^2_{I,s} E^\lambda_R) E^\lambda_{R,t} dxdt.$$

First I_6, I_7, and I_8 are treated as in (137)–(140) so that

$$I_6 + I_7 \le \epsilon \int_0^t \|E^\lambda_{R,t}\|^2_{L^2_x} dt + M \int_0^t \|(z^\lambda_R, z^\lambda_{R,t}, E^\lambda_R)\|^2_{L^2_x} dt + M\lambda^2 \int_0^t \|E^\lambda_{R,t}\|^2_{L^2_x} dt \tag{156}$$

and

$$\begin{aligned} I_8 &\le \epsilon \int_0^t \|E^\lambda_{R,t}\|^2_{L^2_x} dt + M(\epsilon) \int_0^t \int_0^1 (|z^\lambda_{R,t} E^\lambda_R|^2 + |z^\lambda_R E^\lambda_{R,t}|^2) dxdt \\ &\le \epsilon \int_0^t \|E^\lambda_{R,t}\|^2_{L^2_x} dt + M(\epsilon) \int_0^t (\|z^\lambda_{R,t}\|^2_{L^\infty} \|E^\lambda_R\|^2_{L^2_x} + \|z^\lambda_R\|^2_{L^\infty_x} \|E^\lambda_{R,t}\|^2_{L^2_x}) dt \\ &\le \epsilon \int_0^t \|E^\lambda_{R,t}\|^2_{L^2_x} dt \\ &\quad + M \int_0^t \big(\|(z^\lambda_{R,t}, z^\lambda_{R,xt})\|^2_{L^2_x} \|E^\lambda_R\|^2_{L^2_x} + \|(z^\lambda_R, z^\lambda_{R,x})\|^2_{L^2_x} \|E^\lambda_{R,t}\|^2_{L^2_x} \big) dt. \end{aligned} \tag{157}$$

Now we treat the most singular term I_9 by employing the Hardy–Littlewood inequality.

$$\begin{aligned} I_9 &= -\frac{1}{\lambda^2} \int_0^t \int_0^1 (E^0_{I,s} z^\lambda_R + \lambda^2 z^2_{I,s} E^\lambda_R) E^\lambda_{R,t} dxdt \\ &= -\frac{1}{\lambda^2} \int_0^t \int_0^1 t \left(E^0_{I,s} \frac{z^\lambda_R - z^\lambda_R(x,0)}{t} + \lambda^2 z^2_{I,s} \frac{E^\lambda_R - E^\lambda_R(x,0)}{t} \right) E^\lambda_{R,t} dxdt \\ &\quad - \frac{1}{\lambda^2} \int_0^t \int_0^1 (E^0_{I,s} z^\lambda_R(x,0) + \lambda^2 z^2_{I,s} E^\lambda_R(x,0)) E^\lambda_{R,t} dxdt \end{aligned} \tag{158}$$

$$
\begin{aligned}
&\leq \int_0^t \int_0^1 \Bigg(\|sE^0_{I,s}\|_{L^\infty_{(x,t)}} \Big|\frac{z^\lambda_R - z^\lambda_R(x,0)}{t}\Big| |E^\lambda_{R,t}| \\
&\quad + \lambda^2 \|sz^2_{I,s}\|_{L^\infty_{(x,t)}} \Big|\frac{E^\lambda_R - E^\lambda_R(x,0)}{t}\Big| |E^\lambda_{R,t}| \Bigg) dxdt \\
&\quad + \frac{1}{\lambda^2} \int_0^t \int_0^1 |(E^0_{I,s} z^\lambda_R(x,0) + \lambda^2 z^2_{I,s} E^\lambda_R(x,0)) E^\lambda_{R,t}| dxdt \\
&\leq M \int_0^1 \Big\|\frac{z^\lambda_R - z^\lambda_R(x,0)}{t}\Big\|_{L^2_t} \|E^\lambda_{R,t}\|_{L^2_t} dx + M\lambda^2 \int_0^1 \Big\|\frac{E^\lambda_R - E^\lambda_R(x,0)}{t}\Big\|_{L^2_t} \|E^\lambda_{R,t}\|_{L^2_t} dx \\
&\quad + \frac{\epsilon}{2} \int_0^t \|E^\lambda_{R,t}\|^2_{L^2_x} dt + \frac{M}{\lambda^4} \int_0^t \int_0^1 |E^0_{I,s} z^\lambda_R(x,0)|^2 dxdt + M \int_0^t \int_0^1 |z^2_{I,s} E^\lambda_R(x,0)|^2 dxdt \\
&\leq \frac{\epsilon}{2} \int_0^t \|E^\lambda_{R,t}\|^2_{L^2_x} dt + M \int_0^1 \|z^\lambda_{R,t}\|_{L^2_t} \|E^\lambda_{R,t}\|_{L^2_t} dx + M\lambda^2 \int_0^1 \|E^\lambda_{R,t}\|_{L^2_t} \|E^\lambda_{R,t}\|_{L^2_t} dx \\
&\quad + \frac{M}{\lambda^4} \|z^\lambda_R(x,0)\|^2_{L^\infty_x} \int_0^t \int_0^1 |E^0_{I,s}|^2 dxdt + M \|E^\lambda_R(x,0)\|^2_{L^\infty_x} \int_0^t \int_0^1 |z^2_{I,s}|^2 dxdt \\
&\leq (\epsilon + M\lambda^2) \int_0^t \|E^\lambda_{R,t}\|^2_{L^2_x} dt + M(\epsilon) \int_0^t \|z^\lambda_{R,t}\|^2_{L^2_x} dt + M\lambda.
\end{aligned}
$$

Here we used the fact that $\|z^\lambda_R(x,0)\|_{L^\infty_x} = \lambda \|z^\lambda_{0R}\|_{L^\infty_x} \leq M\lambda^{\frac{3}{2}}$ and $\|E^\lambda_R(x,0)\|_{L^\infty_x} = \lambda \|E^\lambda_{0R}\|_{L^\infty_x} \leq M$.

Hence, combining (155) with (156)–(158), one gets

$$
\begin{aligned}
&\int_0^t \int_0^1 G^\lambda_{R,t} E^\lambda_{R,t} dxdt \\
&\quad \leq (\epsilon + M\lambda^2) \int_0^t \|E^\lambda_{R,t}\|^2_{L^2_x} dt + M \int_0^t \|(z^\lambda_R, z^\lambda_{R,t}, E^\lambda_R)\|^2_{L^2_x} dt \\
&\qquad + M \int_0^t \big(\|(z^\lambda_{R,t}, z^\lambda_{R,xt})\|^2_{L^2_x} \|E^\lambda_R\|^2_{L^2_x} + \|(z^\lambda_R, z^\lambda_{R,x})\|^2_{L^2_x} \|E^\lambda_{R,t}\|^2_{L^2_x} \big) dt.
\end{aligned} \tag{159}
$$

Thus, putting (146), (147), (148), (153), (154), and (159) together and taking ϵ small enough shows

$$
\begin{aligned}
&\int_0^t \int_0^1 g^\lambda_t E^\lambda_{R,t} dxdt \\
&\quad \leq (\epsilon + M\lambda^2) \int_0^t \|E^\lambda_{R,t}\|^2_{L^2_x} dt + M \int_0^t \|(z^\lambda_R, z^\lambda_{R,t}, E^\lambda_R)\|^2_{L^2_x} dt \\
&\qquad + M \int_0^t \big(\|(z^\lambda_{R,t}, z^\lambda_{R,xt})\|^2_{L^2_x} \|E^\lambda_R\|^2_{L^2_x} + \|(z^\lambda_R, z^\lambda_{R,x})\|^2_{L^2_x} \|E^\lambda_{R,t}\|^2_{L^2_x} \big) dt + M\lambda.
\end{aligned} \tag{160}
$$

Therefore, for ϵ small enough, (144), together with (145) and (160), gives

$$
\begin{aligned}
&\lambda^2 \|E^\lambda_{R,t}(t)\|^2_{L^2_x} + \lambda^2 \int_0^t \|E^\lambda_{R,xt}\|^2_{L^2_x} dt + c_4 \int_0^t \|E^\lambda_{R,t}\|^2_{L^2_x} dt \\
&\quad \leq M\lambda^2 \|E^\lambda_{R,t}(x,0)\|^2_{L^2_x} + M\lambda^2 \int_0^t \|E^\lambda_{R,t}\|^2_{L^2_x} dt + M \int_0^t \|(z^\lambda_R, E^\lambda_R, z^\lambda_{R,t})\|^2_{L^2_x} dt \\
&\qquad + M \int_0^t \Big(\|(z^\lambda_{R,t}, z^\lambda_{R,xt})\|^2_{L^2_x} \|E^\lambda_R\|^2_{L^2_x} + \|(z^\lambda_R, z^\lambda_{R,x})\|^2_{L^2_x} \|E^\lambda_{R,t}\|^2_{L^2_x} \Big) dt + M\lambda.
\end{aligned} \tag{161}
$$

The desired estimate (126) follows from (143) and (161). This completes the proof of Lemma 7. ∎

Finally, we use the basic estimates and the time derivative estimates of $(z_R^\lambda, E_R^\lambda)$ to obtain these estimates of the space derivatives $\partial_x(z_R^\lambda, E_R^\lambda)$ of $(z_R^\lambda, E_R^\lambda)$.

LEMMA 8. *Under the assumptions of Theorem* 3*, we have*

$$\begin{aligned}
&\|(z_{R,x}^\lambda, E_R^\lambda)\|_{L_x^2}^2 + \lambda^2\|E_{R,x}^\lambda\|_{L_x^2}^2 \\
&\quad\le M\|(z_R^\lambda, z_{R,t}^\lambda)\|_{L_x^2}^2 + M\lambda^2\|(E_R^\lambda, E_{R,t}^\lambda)\|_{L_x^2}^2 + M\lambda^4\|E_{R,x}^\lambda\|_{L_x^2}^2 \\
&\qquad + M\lambda^4\|(E_R^\lambda, E_{R,x}^\lambda)\|_{L_x^2}^2\|E_{R,x}^\lambda\|_{L_x^2}^2 + M\|(z_R^\lambda, z_{R,x}^\lambda)\|_{L_x^2}^2\|E_R^\lambda\|_{L_x^2}^2 + M\lambda.
\end{aligned}\tag{162}$$

Proof of Lemma 8. It follows from (119) and the Cauchy–Schwarz inequality that

$$\begin{aligned}
c_1\|z_{R,x}^\lambda\|_{L_x^2}^2 &\le -\frac{d}{dt}\|z_R^\lambda\|_{L_x^2}^2 + M\|(z_R^\lambda, E_R^\lambda)\|_{L_x^2}^2 + M\lambda^4\|E_{R,x}^\lambda\|_{L_x^2}^2 \\
&\quad + M\lambda^4(\|(E_R^\lambda, E_{R,x}^\lambda)\|_{L_x^2}^2\|E_{R,x}^\lambda\|_{L_x^2}^2 + M\lambda \\
&\le M\|(z_R^\lambda, z_{R,t}^\lambda, E_R^\lambda)\|_{L_x^2}^2 + M\lambda^4\|E_{R,x}^\lambda\|_{L_x^2}^2 \\
&\quad + M\lambda^4\|(E_R^\lambda, E_{R,x}^\lambda)\|_{L_x^2}^2\|E_{R,x}^\lambda\|_{L_x^2}^2 + M\lambda.
\end{aligned}\tag{163}$$

Similarly, it follows from (123) and the Cauchy–Schwarz inequality that

$$\begin{aligned}
&\lambda^2\|E_{R,x}^\lambda\|_{L_x^2}^2 + c_2\|E_R^\lambda\|_{L_x^2}^2 \\
&\quad\le -\lambda^2\frac{d}{dt}\|E_R^\lambda\|_{L_x^2}^2 + M\|z_R^\lambda\|_{L_x^2}^2 + M\|(z_R^\lambda, z_{R,x}^\lambda)\|_{L_x^2}^2\|E_R^\lambda\|_{L_x^2}^2 + M\lambda \\
&\quad\le M\lambda^2\|(E_R^\lambda, E_{R,t}^\lambda)\|_{L_x^2}^2 + M\|z_R^\lambda\|_{L_x^2}^2 + M\|(z_R^\lambda, z_{R,x}^\lambda)\|_{L_x^2}^2\|E_R^\lambda\|_{L_x^2}^2 + M\lambda.
\end{aligned}\tag{164}$$

The desired estimate (162) follows from (163) and (164). This completes the proof of Lemma 8. ∎

The end of the proof of Theorem 3. Introduce the following λ-weighted functional for the remainder terms

$$\Gamma^\lambda(t) = \|(z_R^\lambda, z_{R,x}^\lambda, z_{R,t}^\lambda)\|_{L_x^2}^2 + \lambda^2\|(E_R^\lambda, E_{R,x}^\lambda, E_{R,t}^\lambda)\|_{L_x^2}^2 + \|E_R^\lambda\|_{L_x^2}^2.\tag{165}$$

Then it follows from $(113)+(126)+\delta(162)$ that, by taking δ small enough and then λ small enough, and hence absorbing $\delta\big(M\|(z_R^\lambda, z_{R,t}^\lambda)\|_{L_x^2}^2 + M\lambda^2\|(E_R^\lambda, E_{R,t}^\lambda)\|_{L_x^2}^2\big)$ and $M\delta\lambda^4\|E_{R,x}^\lambda\|_{L_x^2}^2 + M\lambda^4\int_0^t\|E_{R,xt}^\lambda\|_{L_x^2}^2dt$ of the right-hand side of $(113)+(126)+\delta(162)$ by $\|(z_R^\lambda, z_{R,t}^\lambda)\|_{L_x^2}^2 + \lambda^2\|(E_R^\lambda, E_{R,t}^\lambda)\|_{L_x^2}^2$ and $\delta\lambda^2\|E_{R,x}^\lambda\|_{L_x^2}^2 + \lambda^2\int_0^t\|E_{R,xt}^\lambda\|_{L_x^2}^2dt$ of the left-hand side of $(113)+(126)+\delta(162)$, and then using the definition of $\Gamma^\lambda(t)$,

$$\begin{aligned}
&\Gamma^\lambda(t) + \int_0^t\|(z_{R,x}^\lambda, z_{R,xt}^\lambda, E_R^\lambda, E_{R,t}^\lambda)\|_{L_x^2}^2dt + \lambda^2\int_0^t\|(E_{R,x}^\lambda, E_{R,xt}^\lambda)\|_{L_x^2}^2dt \\
&\quad\le M\Gamma^\lambda(t=0) + M\int_0^t(\Gamma^\lambda(t) + (\Gamma^\lambda(t))^2)dt + M\lambda^2\int_0^t\Gamma^\lambda(t)\|E_{R,xt}^\lambda\|_{L_x^2}^2dt \\
&\qquad + M\int_0^t\Gamma^\lambda(t)\|(z_{R,xt}^\lambda, E_{R,t}^\lambda)\|_{L_x^2}^2dt + M\lambda + M(\Gamma^\lambda(t))^2.
\end{aligned}\tag{166}$$

We claim that, for any $T \in [0, T_{\max})$, $T_{\max} \le \infty$, there exists an $\lambda_0 \ll 1$ such that, for any $\lambda \le \lambda_0$, if $\Gamma^\lambda(t=0) \le \tilde{M}\lambda^{\min\{\alpha,1\}}$ for some $\alpha > 0$, then

$$\Gamma^\lambda(t) \le \tilde{M}\lambda^{\min\{\alpha,1\}-\delta}\tag{167}$$

holds for any $\delta \in (0, \min\{\alpha, 1\})$ and $0 \le t \le T$.

Otherwise, there exists $T \in [0, T_{\max})$, $T_{\max} \leq \infty$, for any $\lambda_0 \ll 1$ such that, for some $\lambda \leq \lambda_0$,

$$\Gamma^\lambda(t_0^\lambda) > \tilde{M}\lambda^{\min\{\alpha,1\}-\delta}$$

holds for some $\delta \in (0, \min\{\alpha, 1\})$ and for some $0 < t_0^\lambda \leq T$.

Denote the first root of $\Gamma^\lambda(t) = \tilde{M}\lambda^{\min\{\alpha,1\}-\delta}$ in $[0, t_0^\lambda]$ by t_1^λ. Then we have

$$\Gamma^\lambda(t) \leq \tilde{M}\lambda^{\min\{\alpha,1\}-\delta}, \quad 0 < t \leq t_1^\lambda \leq t_0^\lambda \leq T, \quad \Gamma^\lambda(t_1^\lambda) = \tilde{M}\lambda^{\min\{\alpha,1\}-\delta}. \tag{168}$$

Using (166) and (168), one gets

$$\begin{aligned}
\Gamma^\lambda(t) &+ \int_0^t \|(z_{R,x}^\lambda, z_{R,xt}^\lambda, E_R^\lambda, E_{R,t}^\lambda)\|_{L_x^2}^2 dt + \lambda^2 \int_0^t \|(E_{R,x}^\lambda, E_{R,xt}^\lambda)\|_{L_x^2}^2 dt \\
&\leq M\tilde{M}\lambda^{\min\{\alpha,1\}} \\
&\quad + M\int_0^t (\Gamma^\lambda(t) + \tilde{M}\lambda^{\min\{\alpha,1\}-\delta}\Gamma^\lambda(t))dt + M\lambda^2 \int_0^t \tilde{M}\lambda^{\min\{\alpha,1\}-\delta}\|E_{R,xt}^\lambda\|_{L_x^2}^2 dt \\
&\quad + M\int_0^t \tilde{M}\lambda^{\min\{\alpha,1\}-\delta}\|(z_{R,xt}^\lambda, E_{R,t}^\lambda)\|_{L_x^2}^2 dt + M\lambda + M\tilde{M}\lambda^{\min\{\alpha,1\}-\delta}\Gamma^\lambda(t) \\
&\leq M\tilde{M}\lambda^{\min\{\alpha,1\}} + 2M\int_0^t \Gamma^\lambda(t)dt + \frac{\lambda^2}{2}\int_0^t \|E_{R,xt}^\lambda\|_{L_x^2}^2 dt \\
&\quad + \frac{1}{2}\int_0^t (\|z_{R,xt}^\lambda\|_{L_x^2}^2 + \|E_{R,t}^\lambda\|_{L_x^2}^2)dt + M\lambda + \frac{1}{2}\Gamma^\lambda(t),
\end{aligned} \tag{169}$$

since $\lambda \leq \lambda_0 \ll 1$ and λ_0 can be chosen to satisfy that

$$\tilde{M}\lambda_0^{\min\{\alpha,1\}-\delta} \leq 1, M\tilde{M}\lambda_0^{\min\{\alpha,1\}-\delta} \leq \frac{1}{2}.$$

Hence, it follows from (169) that

$$\Gamma^\lambda(t) \leq 2M\tilde{M}\lambda^{\min\{\alpha,1\}} + 4M\int_0^t \Gamma^\lambda(t)dt + 2M\lambda.$$

Gronwall's lemma gives

$$\begin{aligned}
\Gamma^\lambda(t) &\leq (4Me^{4MT}T + 1)\max\{2M\tilde{M}, 2M\}\lambda^{\min\{\alpha,1\}} \\
&\leq (4Me^{4MT}T + 1)\max\{2M\tilde{M}, 2M\}\lambda^\delta\lambda^{\min\{\alpha,1\}-\delta} \\
&\leq \frac{\tilde{M}}{2}\lambda^{\min\{\alpha,1\}-\delta},
\end{aligned}$$

which contradicts (168). This proves our claim (167).

What remains is to prove that there exist a positive constant $\tilde{M}$ and $\alpha > 0$ such that

$$\Gamma^\lambda(t = 0) \leq \tilde{M}\lambda^\alpha. \tag{170}$$

In fact, it follows from the assumptions (49) and (50) on the initial data and (165) that

$$\Gamma^\lambda(t = 0) = \|z_{R,t}^\lambda(x, 0)\|_{L_x^2}^2 + \lambda^2\|E_{R,t}^\lambda(x, 0)\|_{L_x^2}^2 + M\lambda. \tag{171}$$

First, note that (104) implies

$$\|z^\lambda_{R,t}(x,0)\|_{L^2_x} \le \lambda\|\partial_x^2 z^\lambda_{0R}\|_{L^2_x} + \lambda\|\partial_x(DE^\lambda_{0R})\|_{L^2_x} + \|(H_{Inn,x}, H^\lambda_{B,x}, H^\lambda_{I,x}, H^\lambda_{IB,x}, H^\lambda_{R,x})(t=0)\|_{L^2_x} + M\|f^\lambda(t=0)\|_{L^2_x}.$$

The assumptions (49) and (50) lead to

$$\lambda\|\partial_x^2 z^\lambda_{0R}\|_{L^2_x} + \lambda\|\partial_x(DE^\lambda_{0R})\|_{L^2_x} \le M\sqrt{\lambda},$$

while the definitions of H_{Inn}, H^λ_B, H^λ_I, H^λ_{IB}, and f^λ yield that

$$\|(H_{Inn,x}, H^\lambda_{I,x}, H^\lambda_{IB,x})(t=0)\|_{L^2_x} \le M\sqrt{\lambda},$$

$$\|f^\lambda(t=0)\|_{L^2_x} \le M\lambda^{\frac{3}{2}},$$

and

$$\begin{aligned}
&\|H^\lambda_{B,x}(t=0)\|_{L^2_x}\\
&\quad\le M\sqrt{\lambda} + \left\|\left((D(x)-D(0))\frac{1}{\lambda}f(x)E^0_{+,\xi} + (D(x)-D(1))\frac{1}{\lambda}g(x)E^0_{-,\eta}\right)(t=0)\right\|_{L^2_x}\\
&\quad= M\sqrt{\lambda} + \left\|\left(\int_0^1 D_x(\theta x)d\theta\frac{x}{\lambda}f(x)E^0_{+,\xi}\right.\right.\\
&\qquad\qquad \left.\left. - \int_0^1 D_x(1-\theta(1-x))d\theta(x)\frac{1-x}{\lambda}g(x)E^0_{-,\eta}\right)(t=0)\right\|_{L^2_x}\\
&\quad\le M\sqrt{\lambda}.
\end{aligned}$$

Here we have used the mean value theorem and the estimates $\|(\xi E^0_{+,\xi}, \eta E^0_{+,\eta})(t=0)\|_{L^2_x} \le M\sqrt{\lambda}$.

In addition, the definition of $H^\lambda_R(t=0)$ and assumption (50) imply that

$$\begin{aligned}
\|H^\lambda_{R,x}\|_{L^2_x} &\le M\lambda^2\|\lambda\partial_x^2 E^\lambda_{0R}(x)\|_{L^2_x} + M\lambda\|\lambda\partial_x E^\lambda_{0R}\|_{L^2_x} + M\|\lambda E^\lambda_{0R}(x)\|_{L^2_x}\\
&\quad + \lambda^2\|(\lambda^2 E^\lambda_{0R}(x)\partial_x^2 E^\lambda_{0R}(x), (\lambda\partial_x E^\lambda_{0R}(x))^2)\|_{L^2_x}\\
&\le M\lambda^{\frac{3}{2}}.
\end{aligned}$$

Hence,

$$\|z^\lambda_{R,t}(x,0)\|_{L^2_x} \le M\sqrt{\lambda}. \tag{172}$$

Next, (105) implies that

$$\|\lambda E^\lambda_{R,t}(x,0)\|_{L^2_x} \le \lambda^2\|\partial_x^2 E^\lambda_{0R}(x)\|_{L^2_x} + \|z^0_0(x)E^\lambda_{0R}(x)\|_{L^2_x} + \left\|\frac{1}{\lambda}(G_{Inn}, G^\lambda_B, G^\lambda_I, G^\lambda_{IB}, G^\lambda_R)(t=0)\right\|_{L^2_x}.$$

Note that the only singular term is

$$I_{10} = \|\frac{1}{\lambda}\big(-f(x)(\mathcal{Z}^0(x,t)-\mathcal{Z}^0(0,t))E^0_+ - g(x)(\mathcal{Z}^0(x,t)-\mathcal{Z}^0(1,t))E^0_-\big)(t=0)\|_{L^2_x},$$

while the other terms are easily controlled by $M\sqrt{\lambda}$. By the mean value theorem, one gets

$$\begin{aligned}
I_{10} &\le M \left\| \frac{1}{\lambda}(\mathcal{Z}^0(x,t) - \mathcal{Z}^0(0,t))E_+^0(t=0) \right\|_{L_x^2} \\
&\quad + M \left\| \frac{1}{\lambda}(\mathcal{Z}^0(x,t) - \mathcal{Z}^0(1,t))E_-^0(t=0) \right\|_{L_x^2} \\
&= M \left\| \frac{x}{\lambda} \int_0^1 \mathcal{Z}_x^0(\theta x, 0) d\theta E_+^0(t=0) \right\|_{L_x^2} \\
&\quad + M \left\| \frac{1-x}{\lambda} \int_0^1 \mathcal{Z}_x^0(1-\theta(1-x), 0) d\theta E_-^0(t=0) \right\|_{L_x^2} \\
&\le M \|(\xi E_+^0, \eta E_-^0)(t=0)\|_{L_x^2}^2 \\
&\le M\sqrt{\lambda}.
\end{aligned}$$

This gives

$$\|\lambda E_{R,t}^\lambda(x,0)\|_{L_x^2} \le M\sqrt{\lambda}. \tag{173}$$

Notice that here we have used the assumptions

$$\|\partial_x^2 E_{0R}^\lambda(x)\|_{L_x^2} \le M\lambda^{\frac{1}{2}-2}, \quad \|E_{0R}^\lambda(x)\|_{L_x^2} \le M\sqrt{\lambda}.$$

Thus, (171), together with (172) and (173), gives the desired result (170) with $\alpha = 1$.

By (167) with $\alpha = 1$, one gets (51). This completes the proof of Theorem 3. ◻

4.2. The proof of Theorem 4. In this subsection we prove Theorem 4 by pointing out some necessary modifications of the proof of Theorem 3. We want to proceed as in the proof of Theorem 3.

Now we assume that (54) and (55) hold. In this case, we must consider the effect of the nonzero limit (z_0^1, E_0^1) of the error terms $(z_{0R}^\lambda, E_{0R}^\lambda)$ of the initial data (46). In fact, $z_0^1(x)$ produces the extra initial layer functions (z_I^3, E_I^1), given by the solution to (57)–(60). Since (57)–(60) can be solved exactly, it is easy to see that (z_I^3, E_I^1) has exactly the same properties as (z_I^2, E_I^0). Hence we choose the ansatz as

$$\begin{aligned}
(\tilde{z}^\lambda, \tilde{E}^\lambda)_{app}^T = \Big(&\mathcal{Z}^0 + \lambda(f(x)z_+^1 + g(x)z_-^1) + \lambda z_0^1(x) + \lambda^2 z_I^2 + \lambda^3 z_I^3, \\
&\mathcal{E}^0 + f(x)E_+^0 + g(x)E_-^0 + E_I^0 + \lambda(E_0^1(x) + E_I^1)\Big)^T.
\end{aligned}$$

Set

$$\big(\tilde{z}_R^\lambda(x,t), \tilde{E}_R^\lambda(x,t)\big)^T = (z^\lambda, E^\lambda)^T - \big(\tilde{z}^\lambda, \tilde{E}^\lambda\big)_{app}^T. \tag{174}$$

Then

$$\begin{aligned}
\big(\tilde{z}_R^\lambda(x,t), \tilde{E}_R^\lambda(x,t)\big)^T(x,0) &= \lambda(z_{0R}^\lambda - z_0^1, E_{0R}^\lambda - E_0^1)^T \\
&= \lambda(\tilde{z}_{0R}^\lambda, \tilde{E}_{0R}^\lambda).
\end{aligned}$$

By assumptions (54) and (55), one gets that $(\tilde{z}^\lambda_{0R}, \tilde{E}^\lambda_{0R})$ satisfies assumptions (54) and (55). It remains to establish the energy estimates for the error function

$$\big(\tilde{z}^\lambda_R(x,t), \tilde{E}^\lambda_R(x,t)\big)^T.$$

First, replacing $(z^\lambda, E^\lambda)^T$ with

$$(z^\lambda, E^\lambda)^T = \big(\tilde{z}^\lambda, \tilde{E}^\lambda\big)^T_{app} + \big(\tilde{z}^\lambda_R(x,t), \tilde{E}^\lambda_R(x,t)\big)^T$$

in the system (14)–(15), we obtain (104) and (105) with $(z^\lambda_R, E^\lambda_R)$ replaced by $(\tilde{z}^\lambda_R(x,t), \tilde{E}^\lambda_R(x,t))^T$ and H, G replaced by $\tilde{H}, \tilde{G}$, where $\tilde{H}^\lambda_B = H^\lambda_B$, $\tilde{G}^\lambda_B = G^\lambda_B$, and $\tilde{H}_{Inn}$ $(\tilde{G}_{Inn})$, $\tilde{H}^\lambda_I$ $(\tilde{G}^\lambda_I)$, $\tilde{H}^\lambda_{IB}$ $(\tilde{G}^\lambda_{IB})$, $\tilde{H}^\lambda_R$ $(\tilde{G}^\lambda_R)$ are defined by the following:

$$\tilde{H}_{Inn}(x,t) = \lambda(z^1_{0x}(x) + D(x)E^1_0(x)) - \lambda^2\mathcal{E}^0\mathcal{E}^0_x - \lambda^3\mathcal{E}^0E^1_0(x),$$

$$\begin{aligned}
\tilde{H}^\lambda_I(x,s) &= \lambda^2 z^2_{I,x} + \lambda^3 z^3_{I,x}\\
&\quad -\lambda^2\big(\mathcal{E}^0(E^0_{I,x} + \lambda E^1_{I,x}) + (E^0_I + \lambda E^1_I)(\mathcal{E}^0_x + E^0_{I,x} + \lambda(E^1_{0x} + E^1_{I,x}))\big),\\
\tilde{H}^\lambda_{IB}(x,\xi,\eta,t,s) &= -\lambda\Big((E^0_I + \lambda E^1_I)(f(x)E^0_{+,\xi} - g(x)E^0_{-,\eta})\Big)\\
&\quad -\lambda^2\Big((E^0_I + \lambda E^1_I)(f'(x)E^0_+ + g'(x)E^0_-)\\
&\qquad + (f(x)E^0_+ + g(x)E^0_-)(E^0_{I,x} + \lambda E^1_{I,x})\Big),
\end{aligned}$$

$$\begin{aligned}
\tilde{H}^\lambda_R &= -\lambda\tilde{E}^\lambda_R(f(x)E^0_{+,\xi} - g(x)E^0_{-,\eta})\\
&\quad -\lambda^2\Big(((\mathcal{E}^0 + \lambda E^1_0) + f(x)E^0_+ + g(x)E^0_-)\tilde{E}^\lambda_{R,x}\\
&\qquad + ((\mathcal{E}^0_x + \lambda E^1_{0x}) + f'(x)E^0_+ + g'(x)E^0_-)\tilde{E}^\lambda_R\Big)\\
&\quad -\lambda^2\big((E^0_I + \lambda E^1_I)\tilde{E}^\lambda_{R,x} + (E^0_{I,x} + \lambda E^1_{I,x})\tilde{E}^\lambda_R\big) - \lambda^2\tilde{E}^\lambda_R\tilde{E}^\lambda_{R,x},
\end{aligned}$$

$$\tilde{G}_{Inn}(x,t) = -\lambda^2\big(\mathcal{E}^0_t - \mathcal{E}^0_{xx} - \lambda E^1_{0xx}\big) - \lambda(\mathcal{Z}^0E^1_0 + z^1_0\mathcal{E}^0) - \lambda^2 z^1_0E^1_0,$$

$$\begin{aligned}
\tilde{G}^\lambda_I &= -(\mathcal{Z}^0 - \mathcal{Z}^0(x,0))(E^0_I + \lambda E^1_I) + \lambda^2E^0_{I,xx} + \lambda^3E^1_{I,xx} - \lambda^2 z^1_0E^1_I\\
&\quad -\lambda^2(z^2_I + \lambda z^3_I)(\mathcal{E}^0 + \lambda E^1_0 + E^0_I + \lambda E^1_I),\\
\tilde{G}^\lambda_{IB} &= -\lambda(f(x)z^1_+ + g(x)z^1_-)(E^0_I + \lambda E^1_I) - \lambda^2(z^2_I + \lambda z^3_I)(f(x)E^0_+ + g(x)E^0_-),\\
\tilde{G}^\lambda_R &= -(\mathcal{E}^0 + \lambda E^1_0 + f(x)E^0_+ + g(x)E^0_- + E^0_I + \lambda E^1_I)\tilde{z}^\lambda_R\\
&\quad -\lambda(f(x)z^1_+ + g(x)z^1_- + z^1_0)\tilde{E}^\lambda_R - \lambda^2(z^2_I + \lambda z^3_I)\tilde{E}^\lambda_R - \tilde{z}^\lambda_R\tilde{E}^\lambda_R.
\end{aligned}$$

Next, we point out the difference at the boundary between $(\tilde{z}^\lambda_R, \tilde{E}^\lambda_R)$ and $(z^\lambda_R, E^\lambda_R)$. At present, $\tilde{E}^\lambda_R$ satisfies the nonhomogeneous boundary condition

$$(\tilde{E}^\lambda_R + \lambda E^1_0)(x=0,1;t) = 0, \quad t>0. \tag{175}$$

In fact, since $E^0_I(x=0,1;t)=0$, it follows from the system (57)–(58) that

$$E^0_I(x=0,1;t) = 0, \quad t>0, \tag{176}$$

which gives with (57) and (58) that

$$E_I^1(x=0,1;t)=0, \quad t>0. \tag{177}$$

Combining (174), the boundary condition (16), (68), (176), and (177), one gets (175).

But $\tilde{H}^\lambda$ still satisfies the homogeneous boundary condition

$$\tilde{H}^\lambda(x=0,1;t)=0, \quad t>0.$$

Thus

$$\tilde{H}_t^\lambda(x=0,1;t)=\tilde{E}_{R,t}^\lambda(x=0,1;t)=0, \quad t>0,$$

due to the fact that $E_0^1=E_0^1(x)$ does not depend upon time t.

Finally, notice that $\tilde{H}(\tilde{G})$ is the sum of $H(G)$ and the extra higher order $O(\lambda)$, and hence it has the very similar structure as $H(G)$, and that the only term to be affected by nonhomogeneous boundary condition (175) is $-\lambda^2\int_0^1 \tilde{E}_{R,xx}^\lambda \tilde{E}_R^\lambda dx$, which can be dealt with as follows:

$$\begin{aligned}
-\lambda^2 &\int_0^1 \tilde{E}_{R,xx}^\lambda \tilde{E}_R^\lambda dx \\
&= -\lambda^2\int_0^1 \tilde{E}_{R,xx}^\lambda(\tilde{E}_R^\lambda+\lambda E_0^1)dx+\lambda^3\int_0^1 \tilde{E}_{R,xx}^\lambda E_0^1 dx \\
&= \lambda^2\int_0^1 |\tilde{E}_{R,x}^\lambda|^2dx+\lambda^3\int_0^1 \tilde{E}_{R,x}^\lambda E_{0x}^1 dx+\lambda\int_0^1(\lambda^2\tilde{E}_{R,t}^\lambda+\mathcal{Z}^0\tilde{E}_R^\lambda-g^\lambda)E_0^1dx \\
&\geq \frac{\lambda^2}{2}\int_0^1 |\tilde{E}_{R,x}^\lambda|^2dx-M\lambda-M\lambda^5\int_0^1|\tilde{E}_{R,t}^\lambda|^2dx-\lambda\int_0^1(|\tilde{E}_R^\lambda|^2+|g^\lambda|^2)dx \\
&\geq \frac{\lambda^2}{2}\int_0^1 |\tilde{E}_{R,x}^\lambda|^2dx-M\lambda^5\int_0^1|\tilde{E}_{R,t}^\lambda|^2dx-\lambda\int_0^1|\tilde{E}_R^\lambda|^2dx-M\lambda.
\end{aligned}$$

Here we had used (105). Thus, we can proceed with the energy method as in the previous proof of Theorem 3. The proof of Theorem 4 is complete. □

Acknowledgment. The authors thank the referees for their constructive comments and suggestions.

REFERENCES

[1] A. Arnold, P. A. Markowich, and G. Toscani, *On large time asymptotics for drift-diffusion-Poisson systems*, Transport Theory Statist. Phys., 29 (2000), pp. 571–581.

[2] Y. Brenier, *Convergence of the Vlasov-Poisson system to the incompressible Euler equations*, Comm. Partial Differential Equations, 25 (2000), pp. 737–754.

[3] L. Caffarelli, J. Dolbeault, P. A. Markowich, and C. Schmeiser, *On Maxwellian equilibria of insulted semiconductors*, Interfaces Free Bound., 2 (2000), pp. 331–339.

[4] S. Cordier, P. Degond, P. Markowich, and C. Schmeiser, *A travelling wave analysis of a hydrodynamical model of quasineutral plasma*, Asymptot. Anal., 11 (1995), pp. 209–224.

[5] S. Cordier and E. Grenier, *Quasineutral limit of an Euler-Poisson system arising from plasma physics*, Comm. Partial Differential Equations, 23 (2000), pp. 1099–1113.

[6] S. Cordier and E. Grenier, *Quasineutral limit of two species Euler-Poisson system*, in Proceedings of the Workshop "Recent Progress in the Mathematical Theory on Vlasov-Maxwell Equations," Paris, 1997, pp. 95–122.

[7] J. Dolbeault, P. A. Markowich, and A. Unterreiter, *On singular limits of mean-field equations*, Arch. Ration. Mech. Anal., 158 (2001), pp. 319–351.

[8] H. Gajewski, *On existence, uniqueness and asymptotic behaviour of solutions of the basic equations for carrier transport in semiconductors*, ZAMM Z. Angew. Math. Mech., 65 (1985), pp. 101–108.

[9] H. Gajewski and K. Gröger, *On the basic equations for carrier transport in semiconductors*, J. Math. Anal. Appl., 113 (1986), pp. 12–35.

[10] I. Gasser, L. Hsiao, P. Markowich, and S. Wang, *Quasineutral limit of a nonlinear drift-diffusion model for semiconductor models*, J. Math. Anal. Appl., 268 (2002), pp. 184–199.

[11] I. Gasser, C. D. Levermore, P. Markowich, and C. Schmeiser, *The initial time layer problem and the quasineutral limit in the semiconductor drift-diffusion model*, Eur. J. Appl. Math., 12 (2001), pp. 497–512.

[12] E. Grenier, *Oscillations in quasineutral plasmas*, Comm. Partial Differential Equations, 21 (1996), pp. 363–394.

[13] E. Grenier, *Defect measures of the Vlasov-Poisson system*, Comm. Partial Differential Equations, 20 (1995), pp. 1189–1215.

[14] A. Jüngel, *Qualitative behavior of solutions of a degenerate nonlinear drift-diffusion model for semiconductors*, Math. Models Methods Appl. Sci., 5 (1995), pp. 497–518.

[15] A. Jüngel and Y.-J. Peng, *A hierarchy of hydrodynamic models for plasmas: Quasi-neutral limits in the drift-diffusion equations*, Asymptot. Anal., 28 (2001), pp. 49–73.

[16] A. Jüngel and S. Wang, *Convergence of nonlinear Schrödinger-Poisson systems to the compressible Euler equations*, Comm. Partial Differential Equations, 28 (2003), pp. 1005–1022.

[17] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, Trans. Math. Monogr., AMS, Providence, RI, 1968.

[18] J. G. Liu and Z. P. Xin, *Boundary layer behavior in the fluid-dynamic limit for a nonlinear model Boltzmann equation*, Arch. Ration. Mech. Anal., 135 (1996), pp. 61–105.

[19] P. A. Markowich and C. A. Ringhofer, *A singularly perturbed boundary value problem modelling a semiconductor device*, SIAM J. Appl. Math., 44 (1984), pp. 231–256.

[20] P. A. Markowich, C. Ringhofer, and C. Schmeiser, *Semiconductors Equations*, Springer-Verlag, New York, 1990.

[21] N. Masmoudi, *From Vlasov-Poisson system to the incompressible Euler system*, Comm. Partial Differential Equations, 26 (2001), pp. 1913–1928.

[22] M. S. Mock, *On equations describing steady-state carrier distributions in a semiconductor device*, Comm. Pure Appl. Math., 25 (1972), pp. 781–792.

[23] M. S. Mock, *An initial value problem from semiconductor device theory*, SIAM J. Math. Anal., 5 (1974), pp. 597–612.

[24] M. S. Mock, *Analysis of Mathematical Models of Semiconductor Devices*, Boole Press, Dublin, 1983.

[25] C. P. Please, *An analysis of semiconductor p-n junctions*, IMA J. Appl. Math., 28 (1982), pp. 301–318.

[26] M. Puel, *Etudes Variationnelles et Asymptotiques des Problèmes de la Mécanique des Fluides et des Plasmas*, Ph.D. thesis, Université Paris 6, 2000.

[27] C. Ringhofer, *An asymptotic analysis of a transient p-n-junction model*, SIAM J. Appl. Math., 47 (1987), pp. 624–642.

[28] W. V. Roosbroeck, *Theory of the flow of electrons and holes in Germanium and other semiconductors*, Bell Syst. Tech. J., 29 (1950), pp. 560–607.

[29] I. Rubinstein, *Electro-diffusion of Ions*, SIAM, Philadelphia, 1990.

[30] C. Schmeiser and S. Wang, *Quasineutral limit of the drift-diffusion model for semiconductors with general initial data*, Math. Model. Methods Appl. Sci., 13 (2003), pp. 463–470.

[31] W. Shockley, *The theory of p-n junctions in semiconductors and p-n junction transistors*, Bell. Syst. Tech. J., 28 (1949), pp. 435–489. (Reprinted in Ann. Mat. Pura Appl., 166 (1987), pp. 65–96.)

[32] A. Sitenko and V. Malnev, *Plasma Physics Theory*, Chapman and Hall, London, 1995.

[33] R. A. Smith, *Semiconductors*, 2nd ed., Cambridge University Press, Cambridge, UK, 1978.

[34] S. M. Sze, *Physics of Semiconductor Devices*, Wiley-Interscience, New York, 1969.

[35] S. Wang, *Quasineutral limit of Euler-Poisson system with and without viscosity*, Comm. Partial Differential Equations, 29 (2004), pp. 419–456.

[36] Z. P. Xin, *Viscous boundary layers and their stability* (I), J. Partial Differential Equations, 11 (1998), pp. 74–124.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1890–1902

ON THE STOLZ–ADAMS DECONVOLUTION MODEL FOR THE LARGE-EDDY SIMULATION OF TURBULENT FLOWS*

A. DUNCA† AND Y. EPSHTEYN†

Abstract. We consider a family of large-eddy simulation (LES) models with an arbitrarily high consistency error $O(\delta^{2N+2})$ for $N = 1, 2, 3, \dots$ that are based on the van Cittert deconvolution procedure. This family of models has been proposed and tested for LES with success by Adams and Stolz in a series of papers, e.g., [*Deconvolution methods for subgrid-scale approximation in large-eddy simulation*, in Modern Simulation Strategies for Turbulent Flow, R. T. Edwards, Philadelphia, 2001, pp. 21–41], [*Phys. Fluids*, 11 (1999), pp. 1699–1701]. We show that these models have an interesting and quite strong stability property. Using this property we prove an energy equality, existence, uniqueness, and regularity of strong solutions and give a rigorous bound on the modeling error $\|\overline{\mathbf{u}} - \mathbf{w}\|$, where $\mathbf{w}$ is the model's solution and $\overline{\mathbf{u}}$ is the true flow averages.

Key words. large-eddy simulation, scale similarity models, deconvolution, approximate deconvolution models

AMS subject classifications. 76F65, 76D03

DOI. 10.1137/S0036141003436302

1. Introduction.

We consider the problem of modeling the motion of large structures in a turbulent fluid. This involves the interaction of many complex decisions made in the simulation. To isolate some effects, we study herein the correctness of the approximate deconvolution modeling (ADM) approach to closure pioneered by Adams and Stolz; see, e.g., [1], [9].

The pointwise velocity and pressure, $\mathbf{u}, p$, in an incompressible viscous flow satisfy the Navier–Stokes equations

$$\begin{aligned} \mathbf{u}_t + \nabla \cdot (\mathbf{u}\mathbf{u}^T) - \nu\Delta\mathbf{u} + \nabla p &= \mathbf{f}, \\ \nabla \cdot \mathbf{u} &= 0, \\ \mathbf{u}(\mathbf{x}, 0) &= \mathbf{u}_0(\mathbf{x}). \end{aligned} \tag{1}$$

We study (1) subject to periodic boundary conditions (with zero mean)

$$\mathbf{u}(x + Le, t) = \mathbf{u}(x, t) \tag{2}$$

for $x \in \mathbb{R}^3, 0 < t \leq T$.

Periodic boundary conditions separate the hard problem of closure for the interior equations from another hard problem of wall laws and near wall models in turbulence.

Let an overbar denote a local spacial averaging associated with a length scale δ which commutes with differentiation. Averaging the Navier–Stokes equations gives the nonclosed equations for $\overline{\mathbf{u}}, \overline{p}$,

$$\begin{aligned} \overline{\mathbf{u}}_t + \nabla \cdot (\overline{\mathbf{u}\mathbf{u}^T}) - \nu\Delta\overline{\mathbf{u}} + \nabla\overline{p} &= \overline{\mathbf{f}}, \\ \nabla \cdot \overline{\mathbf{u}} &= 0. \end{aligned} \tag{3}$$

---

*Received by the editors October 21, 2003; accepted for publication (in revised form) June 11, 2004; published electronically March 3, 2006. This research was partially supported by NSF grant DMS 0207627.

http://www.siam.org/journals/sima/37-6/43630.html

†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (ardst21@math.pitt.edu, yee1@pitt.edu).

Let the averaging operation $\mathbf{u} \to \overline{\mathbf{u}}$ be denoted formally by G so $\overline{\mathbf{u}} = \mathrm{G}\mathbf{u}$. In the most interesting cases G is not invertible. Nevertheless, the closure problem (of replacing $\overline{\mathbf{u}\mathbf{u}^T}$ by a tensor depending only on $\overline{\mathbf{u}}$) is solved once the approximate deconvolution problem (of approximating the action of G^{-1}) is solved.

The van Cittert approximation to G^{-1} can be developed in various ways (see [2] and section 2 for a precise definition of it). The simplest is to find an approximation to $\mathbf{u}$ by extrapolating from the resolved scales of $\overline{\mathbf{u}}$ to those of $\mathbf{u}$. The first three examples are

$$
\begin{aligned}
&\mathbf{u} \approx \mathrm{G}_0\overline{\mathbf{u}} := \overline{\mathbf{u}} && \text{(constant extrapolation in } \delta\text{)},\\
&\mathbf{u} \approx \mathrm{G}_1\overline{\mathbf{u}} := \overline{\mathbf{u}} - \overline{\overline{\mathbf{u}}} && \text{(linear extrapolation in } \delta\text{)},\\
&\mathbf{u} \approx \mathrm{G}_2\overline{\mathbf{u}} := 3\overline{\mathbf{u}} - 3\overline{\overline{\mathbf{u}}} + \overline{\overline{\overline{\mathbf{u}}}} && \text{(quadratic extrapolation in } \delta\text{)}.
\end{aligned}
\tag{4}
$$

Let $\mathrm{G}_N\overline{\mathbf{u}}$ denote the analogous Nth degree accurate approximate inverse (section 2). Calling $(\mathbf{w}, q)$ the approximations that result when this is used in (3) to treat the closure problem, we are inevitably led to the fundamentally important question of how well the solution $\mathbf{w}$ of the resulting model,

$$
\begin{aligned}
\mathbf{w}_t + \nabla \cdot (\overline{\mathrm{G}_N\mathbf{w}(\mathrm{G}_N\mathbf{w})^T}) - \nu\Delta\mathbf{w} + \nabla\overline{q} &= \overline{\mathbf{f}},\\
\nabla \cdot \mathbf{w} &= 0,
\end{aligned}
\tag{5}
$$

matches the behavior of the true flow averages $\overline{\mathbf{u}}$. This question has obvious theoretical and experimental components. We consider herein the theoretical parts of the question for the whole family of models. Our analysis is based on a delicate skew symmetry property that the model's nonlinear interaction terms have when the averaging operator is the differential filter $\varphi \to \overline{\varphi}$ (as studied by Germano [4]). Here for given $\varphi \in L^2(Q)$, $\overline{\varphi}$ is defined to be the unique periodic solution of

$$
-\delta^2\Delta\overline{\varphi} + \overline{\varphi} = \varphi \tag{6}
$$

in Q, where Q denotes the d-dimensional cube of size $L > 0, Q = (0, L)^d$.

Our analysis is for periodic boundary conditions. We believe that many of the results presented in this work can be extended to nonperiodic boundary conditions with further research. Indeed, the basic model (5) does not increase the order of the differential operator, so the model makes perfect sense coupled with any of the well-posed boundary conditions used for the Navier–Stokes equations.

REMARK 1.1. *The model* (5) *using* G_0 *was considered recently in* [6] *and* [7]*. On the other hand, practical calculations of Adams and Stolz in* [1] *and* [9] *have stressed the superiority of models of order* 4, 5 *and higher in practical tests.*

Herein we show that a single, unified mathematical theory is possible for the entire family of models building on the analysis in [6] *and* [7].

2. Deconvolution models. It has been pointed out by Germano (presented well in [5]) that with the differential filter $\overline{\varphi} := (-\delta^2\Delta + \mathrm{I})^{-1}\varphi$ it seems that no deconvolution is necessary; one can write exactly $\varphi := (-\delta^2\Delta + \mathrm{I})\overline{\varphi}$.

This leads to the exact model for $\overline{\mathbf{u}}$ given by

$$
\overline{\mathbf{u}}_t + \nabla \cdot (\overline{(-\delta^2\Delta + \mathrm{I})\overline{\mathbf{u}}[(-\delta^2\Delta + \mathrm{I})\overline{\mathbf{u}}]^T}) - \nu\Delta\overline{\mathbf{u}} + \nabla\overline{p} = \overline{\mathbf{f}} \tag{7}
$$

subject to the periodic boundary conditions. One criticism with using the exact deconvolution model (7) to predict $\overline{\mathbf{u}}$ is that going from the Navier–Stokes equations to (7), no information is lost.

Thus there is no reason to believe that (7) can be approximated with fewer degrees of freedom than the Navier–Stokes equation itself. Another difficulty with (7) is that any model that increases the order of the differential equation must be supplied with extra boundary conditions. Thus for nonperiodic problems, models such as (7) shift the essential difficulty from interior closure to the harder problem of specifying as boundary conditions the higher derivatives of turbulent velocities at walls. Thus approximate deconvolution which will lose information is necessary. The van Cittert method of approximate deconvolution (see [2]) constructs a family G_N of inverses to G as follows: writing $\mathrm{G} = \mathrm{I} - (\mathrm{I} - \mathrm{G})$, a formal inverse to G can be written as the nonconvergent power series,

$$\mathrm{G}^{-1} = \sum_{n=0}^{\infty} (\mathrm{I} - \mathrm{G})^n.$$

Truncating this series gives

$$\mathrm{G}_N = \sum_{n=0}^{N} (\mathrm{I} - \mathrm{G})^n. \tag{8}$$

The first three approximations are given in (4).

LEMMA 2.1. *The operator* $\mathrm{G}_N : L^2(Q) \to L^2(Q)$ *is compact, self-adjoint, and positive.*

Proof. The operator $\mathrm{G} : L^2(Q) \to L^2(Q)$ is compact and self-adjoint. Multiplying (6) by $\overline{\varphi}$ and integrating over Q gives

$$\delta^2 \|\nabla \overline{\varphi}\|^2 + \|\overline{\varphi}\|^2 = (\varphi, \overline{\varphi}) \leq \frac{1}{2}\|\varphi\|^2 + \frac{1}{2}\|\overline{\varphi}\|^2.$$

It follows that G is positive and $\|\mathrm{G}\| \leq 1$. Let $h_N(x) = \sum_{k=0}^{N} (1-x)^k$. By the definition of G_N,

$$\mathrm{G}_N = h_N(\mathrm{G})$$

and, consequently, G_N is also a compact self-adjoint operator. Because h_N is positive on $[0, 1]$, which contains the spectrum of G, it also follows that G_N is positive. ∎

REMARK 2.1. *The operators* $\{\mathrm{G}_N\}_N$ *satisfy the following recursion:*

$$(\mathrm{I} - \delta^2 \Delta)\mathrm{G}_N \mathbf{u} = -\delta^2 \Delta \mathrm{G}_{N-1} \mathbf{u} + (\mathrm{I} - \delta^2 \Delta)\mathbf{u}. \tag{9}$$

The following lemma, which is a consequence of the identity

$$G_N G = I - (I - G)^{N+1},$$

was proved in [10].

LEMMA 2.2. *For smooth* $\mathbf{u}$ *the approximate deconvolution* (8) *has the consistency error* $O(\delta^{2N+2})$,

$$\mathbf{u} - \mathrm{G}_N \overline{\mathbf{u}} = (-1)^{N+1} \delta^{2N+2} \Delta^{N+1} G^{N+1} \overline{\mathbf{u}}, \tag{10}$$

locally in Q *and also*

$$\|\mathbf{u} - \mathrm{G}_N \overline{\mathbf{u}}\| \leq \delta^{2N+2} \|\overline{\mathbf{u}}\|_{H^{2N+2}(Q)}.$$

Lemma 2.2 shows that $\mathrm{G}_N\overline{\mathbf{u}}$ gives an approximation to $\mathbf{u}$ to the accuracy $O(\delta^{2N+2})$ in the smooth flow regions. Thus it is justified to use it for the closure approximation

$$\nabla \cdot \overline{(\mathbf{u}\mathbf{u}^T)} \approx \nabla \cdot \overline{(\mathrm{G}_N\overline{\mathbf{u}}(\mathrm{G}_N\overline{\mathbf{u}})^T)} + O(\delta^{2N+2}).$$

If μ denotes the usual subfilter scale stress tensor $\mu(\mathbf{u},\mathbf{u}) := \overline{\mathbf{u}\mathbf{u}^T} - \bar{\mathbf{u}}\bar{\mathbf{u}}^T$, then the closure approximation is equivalent to the closure model

$$\mu(\mathbf{u},\mathbf{u}) \approx \mu_N(\overline{\mathbf{u}},\overline{\mathbf{u}}) := \overline{\mathrm{G}_N\overline{\mathbf{u}}(\mathrm{G}_N\overline{\mathbf{u}})^T} - \bar{\mathbf{u}}\bar{\mathbf{u}}^T. \tag{11}$$

The true subgrid stress tensor $\mu(\mathbf{u},\mathbf{u})$ is both reversible and Galilean invariant (Sagaut [8]). Thus many feel that appropriate closure models should at least, to leading order effects, share these two properties. We next show that the model (11) is both reversible and Galilean invariant.

LEMMA 2.3. *For each $N = 0, 1, 2, \ldots$ the closure model* (11) *is reversible and Galilean invariant.*

Proof. Reversibility is immediate. Galilean invariance also follows easily once it is noted that $\overline{\mathrm{U}\mathbf{w}^T} = \mathrm{U}\overline{\mathbf{w}}^T$ so $\mathrm{G}_N(\mathrm{U}\overline{\mathbf{u}}^T = \mathrm{U}\mathrm{G}_N(\overline{\mathbf{u}})^T$. Using these and other analogous properties gives

$$\begin{aligned}
\nabla \cdot \mu(\overline{\mathbf{u}} + \mathrm{U}, \overline{\mathbf{u}} + \mathrm{U}) &= \nabla \cdot [\overline{\mathrm{G}_N(\overline{\mathbf{u}})\mathrm{G}_N(\overline{\mathbf{u}})^T} + \mathrm{U}\overline{\mathrm{G}_N(\overline{\mathbf{u}})}^T \\
&\quad + \overline{\mathrm{G}_N(\overline{\mathbf{u}})}\mathrm{U}^T + \overline{\mathrm{U}\mathrm{U}^T} - (\overline{\mathbf{u}} + \mathrm{U})(\overline{\mathbf{u}} + \mathrm{U})^T] \\
&= \nabla \cdot [\overline{\mathrm{G}_N(\overline{\mathbf{u}})\mathrm{G}_N(\overline{\mathbf{u}})^T} - \bar{\mathbf{u}}\bar{\mathbf{u}}^T] + \nabla \cdot \overline{\mathrm{G}_N(\overline{\mathbf{u}})}\mathrm{U} \\
&\quad + \mathrm{U}\nabla \cdot (\overline{\mathrm{G}_N(\overline{\mathbf{u}})} - \nabla \cdot (\overline{\mathbf{u}})\mathrm{U} - \mathrm{U}\nabla \cdot (\overline{\mathbf{u}}) \\
&= \nabla \cdot [\overline{\mathrm{G}_N(\overline{\mathbf{u}})\mathrm{G}_N(\overline{\mathbf{u}})^T} - \bar{\mathbf{u}}\bar{\mathbf{u}}^T] = \nabla \cdot \mu_N(\overline{\mathbf{u}},\overline{\mathbf{u}}),
\end{aligned}$$

since $\nabla \cdot \overline{\mathbf{u}} = \nabla \cdot \mathrm{G}_N(\overline{\mathbf{u}}) = \nabla \cdot \overline{\mathrm{G}_N(\overline{\mathbf{u}})} = 0$ and $\overline{\mathrm{U}\mathrm{U}^T} = \mathrm{U}\mathrm{U}^T$. ∎

3. Variational spaces. Q denotes a d-dimensional cube of size $L > 0$,

$$Q = (0, L)^d.$$

Let

$$H^m(Q) = \{\mathbf{u} \in H^m_{loc}(\mathbb{R}^n) | \mathbf{u} \text{ periodic with period } Q\}$$

and

$$\overline{H}^m(Q) = \left\{\mathbf{u} \in H^m(Q) | \int_Q \mathbf{u}\,\mathrm{d}x = 0\right\}.$$

For the variational formulation of the scale similarity model with periodic boundary conditions, we consider the spaces of divergence-free functions

$$V = \{\mathbf{u} \in H^1(Q), \nabla \cdot \mathbf{u} = 0 \text{ in } \mathbb{R}^d\}$$

and

$$H = \{\mathbf{u} \in L^2(Q), \nabla \cdot \mathbf{u} = 0 \text{ in } \mathbb{R}^d\}$$

as in Temam [12].

$D(Q)$ is defined as

$$D(Q) = \{\psi \in C^\infty(\mathbb{R}^d) | \psi \text{ is periodic with period } Q\}$$

and

$$D(Q_T) = \{\psi \in C^\infty([0,T) \times \mathbb{R}^d) | \text{ for } t \in [0,T), \psi(\cdot,t) \text{ is periodic with period } Q \text{ and } \psi \text{ has compact support in variable } t \in [0,T)\}.$$

The space of vector valued functions $\mathbb{D}(Q)$ is defined as

$$\mathbb{D}(Q) = D(Q)^d. \tag{12}$$

The other spaces $\mathbb{D}(Q_T), \mathbb{H}, \overline{\mathbb{H}}^p(Q)$, and $\mathbb{V}, \mathbb{L}^2(Q)$ are defined accordingly.

REMARK 3.1. *Because the inclusion $\overline{H}^2(Q) \to H$ is compact, the inverse of the Laplacian operator $(-\Delta)^{-1} : H \to \overline{H}^2(Q) \subset H$ is a bounded, self-adjoint, and compact operator. This implies that there exists an orthonormal basis $(\mathbf{w}_j)_{j\in\mathbb{N}}$ of H consisting of eigenfunctions of the Laplacian operator.*

4. The models and the existence of weak solutions.

DEFINITION 4.1. *The strong form of the Stolz–Adams model that we analyze is as follows: Find $(\mathbf{w}, q)$ such that*

$$\begin{aligned} &\mathbf{w} \in (\overline{H}^2(Q) \cap H)^d && \text{for a.e. } t \in [0.T], \\ &\mathbf{w} \in (H^1(0,T))^d && \text{for a.e. } \mathbf{x} \in \overline{Q}, \\ &q \in H^1(Q) \cap L^2_0(Q) && \text{if } t \in (0,T] \end{aligned} \tag{13}$$

and

$$\begin{aligned} \mathbf{w}_t - \nu\Delta\mathbf{w} + \nabla \cdot (\overline{(\mathrm{G}_N\mathbf{w})(\mathrm{G}_N\mathbf{w})^T}) + \nabla\overline{q} &= \overline{\mathbf{f}} && \text{in } (0,T) \times Q, \\ \nabla \cdot \mathbf{w} &= 0 && \text{in } (0,T] \times Q, \\ \mathbf{w}|_{t=0} &= \overline{\mathbf{u}_0} && \text{in } Q, \\ \int_Q q \, \mathrm{d}x &= 0 && \text{in } (0,T]. \end{aligned} \tag{14}$$

DEFINITION 4.2. *Let $\mathbf{f} \in L^2(0,T;\mathbb{V}')$ and $\mathbf{w}_0 \in \overline{\mathbb{H}}^2(Q)$. A measurable function $\mathbf{w} : [0,T] \times Q \to \mathbb{R}^d$ is a weak solution of* (14) *if*

$$\mathbf{w} \in L^2(0,T,\overline{\mathbb{H}}^1(Q)) \cap L^\infty(0,T;\mathbb{H}) \tag{15}$$

and

$$\begin{aligned} &\int_0^\infty \left[\left(\mathbf{w}, \frac{\partial\varphi}{\partial t} \right) - \nu(\nabla\mathbf{w}, \nabla\phi) - (\nabla \cdot (\overline{(\mathrm{G}_N\mathbf{w})(\mathrm{G}_N\mathbf{w})^T}), \varphi) \right] \mathrm{d}t \\ &\quad = -\int_0^\infty (\mathbf{f}, \varphi) \, \mathrm{d}t - (\mathbf{w}_0, \varphi(0)) \end{aligned} \tag{16}$$

for all $\varphi \in \mathbb{D}(Q)$.

The following lemma gives an energy inequality satisfied by the strong solutions of the Stolz–Adams models. We mention here that the same argument is used to derive an energy inequality for the approximate solutions in the proof of existence of weak solutions to the Stolz–Adams models.

LEMMA 4.1. *If* $\mathbf{w}$ *is a strong solution of* (14) *as in Definition* 4.1, *then* $\mathbf{w}$ *satisfies the following energy inequality:*

$$\begin{aligned}\tag{17} &\frac{1}{2}(\|\mathbf{w}(t)\|^2+\delta^2\|\nabla\mathbf{w}(t)\|^2)+\frac{\nu}{2}\int_0^t\|\nabla\mathbf{w}(s)\|^2+\delta^2\|\Delta\mathbf{w}(s)\|^2\,\mathrm{d}s\\ &\quad\le K\left(\int_0^T\|\mathbf{f}(s)\|^2_{V'}\mathrm{d}s+\|\mathbf{w}_0\|^2+\delta^2\|\nabla\mathbf{w}_0\|^2\right)\end{aligned}$$

for all $t\in[0,T]$ *with* $K=\max\{\frac{2\|\mathrm{G}_N\|^2_{L^2(Q)}}{\nu},\frac{1}{2}\delta^2,\frac{1}{2},\frac{\delta^2}{2}\|\mathrm{G}_{N-1}\|_{L^2(Q)}\}$.

Proof. We multiply (14) by the test function $\varphi=(-\delta^2\Delta+\mathrm{I})\mathrm{G}_N\mathbf{w}$ and integrate on Q. Because the weak form of the nonlinear term will vanish,

$$\tag{18} (\nabla\cdot\overline{((\mathrm{G}_N\mathbf{w})(\mathrm{G}_N\mathbf{w})^T)},(-\delta^2\Delta+\mathrm{I})\mathrm{G}_N\mathbf{w})$$

$$\tag{19} =(\nabla\cdot((\mathrm{G}_N\mathbf{w})(\mathrm{G}_N\mathbf{w})^T),\overline{(-\delta^2\Delta+\mathrm{I})\mathrm{G}_N\mathbf{w}})=(\nabla\cdot((\mathrm{G}_N\mathbf{w})(\mathrm{G}_N\mathbf{w})^T),\mathrm{G}_N\mathbf{w})=0,$$

we obtain the following energy equality:

$$\tag{20} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{w},(-\delta^2\Delta+\mathrm{I})\mathrm{G}_N\mathbf{w})+\nu(\Delta\mathbf{w},(-\delta^2\Delta+\mathrm{I})\mathrm{G}_N\mathbf{w})=(\overline{\mathbf{f}},(-\delta^2\Delta+\mathrm{I})\mathrm{G}_N\mathbf{w}).$$

In the above equality all terms $(-\delta^2\Delta+\mathrm{I})\mathrm{G}_N\mathbf{w}$ are replaced using Remark 2.1, leading to

$$\begin{aligned}&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{w},(-\delta^2\Delta+\mathrm{I})\mathbf{w})+\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{w},-\delta^2\Delta\mathrm{G}_{N-1}\mathbf{w})-\nu(\Delta\mathbf{w},(-\delta^2\Delta+\mathrm{I})\mathbf{w})\\ &\quad+\nu\delta^2(\Delta\mathbf{w},\delta^2\Delta\mathrm{G}_{N-1}\mathbf{w})=(\mathbf{f},\mathrm{G}_N\mathbf{w}).\end{aligned}$$

Using integration by parts and the commutation property of the operator G_{N-1} with differentiation gives

$$\begin{aligned}&\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{w}\|^2+\frac{1}{2}\delta^2\frac{\mathrm{d}}{\mathrm{d}t}\|\nabla\mathbf{w}\|^2+\frac{\delta^2}{2}\frac{\mathrm{d}}{\mathrm{d}t}(\nabla\mathbf{w},\mathrm{G}_{N-1}\nabla\mathbf{w})\\ &\quad+\nu\|\nabla\mathbf{w}\|^2+\nu\delta^2\|\Delta\mathbf{w}\|+\nu\delta^4(\Delta\mathbf{w},\mathrm{G}_{N-1}\Delta\mathbf{w})=(\mathbf{f},\mathrm{G}_N\mathbf{w}).\end{aligned}\tag{21}$$

We then integrate on $[0,t]$ and obtain

$$\begin{aligned}&\frac{1}{2}\|\mathbf{w}(t)\|^2+\frac{1}{2}\delta^2\|\nabla\mathbf{w}(t)\|^2+\frac{\delta^2}{2}(\nabla\mathbf{w}(t),\mathrm{G}_{N-1}\nabla\mathbf{w}(t))+\nu\int_0^t\|\nabla\mathbf{w}(s)\|^2\,\mathrm{d}s\\ &\quad+\nu\delta^2\int_0^t\|\Delta\mathbf{w}(s)\|^2\,\mathrm{d}s+\nu\delta^4\int_0^t(\Delta\mathbf{w}(s),\mathrm{G}_{N-1}\Delta\mathbf{w}(s))\,\mathrm{d}s\\ &=\int_0^t(\mathbf{f}(\mathbf{s}),\mathrm{G}_N\mathbf{w}(s))\,\mathrm{d}s+\frac{1}{2}\|\mathbf{w}_0\|^2+\frac{1}{2}\delta^2\|\nabla\mathbf{w}_0\|^2+\frac{\delta^2}{2}(\nabla\mathbf{w}_0,\mathrm{G}_{N-1}\nabla\mathbf{w}_0).\end{aligned}$$

We use the positivity of the operators $(\mathrm{G}_N)_N$ in the above inequality to get

$$\begin{aligned}\tag{22} &\frac{1}{2}\|\mathbf{w}(t)\|^2+\frac{1}{2}\delta^2\|\nabla\mathbf{w}(t)\|^2+\nu\int_0^t\|\nabla\mathbf{w}(s)\|^2\,\mathrm{d}s+\nu\delta^2\int_0^t\|\Delta\mathbf{w}(s)\|^2\,\mathrm{d}s\\ &\quad\le\int_0^t(\mathbf{f}(\mathbf{s}),\mathrm{G}_N\mathbf{w}(s))\,\mathrm{d}s+\frac{1}{2}\delta^2\|\mathbf{w}_0\|^2+\frac{1}{2}\|\nabla\mathbf{w}_0\|^2+\frac{\delta^2}{2}(\nabla\mathbf{w}_0,\mathrm{G}_{N-1}\nabla\mathbf{w}_0).\end{aligned}$$

An application of Cauchy's inequality on the first term on the right-hand side above gives

$$\begin{aligned}\int_0^t (\mathbf{f}(\mathbf{s}), \mathrm{G}_N \mathbf{w}(s))\,\mathrm{d}s &\leq \int_0^t \|\mathbf{f}(s)\|_{V'} \|\mathrm{G}_N\|_{L^2(Q)} \|\nabla \mathbf{w}(s)\|_{L^2(Q)}\,\mathrm{d}s \\ &\leq \frac{2\|\mathrm{G}_N\|^2_{L^2(Q)}}{\nu} \int_0^t \|\mathbf{f}(\mathbf{s})\|^2_{V'}\,\mathrm{d}s + \frac{\nu}{2}\int_0^T \|\nabla \mathbf{w}(s)\|^2_{L^2(Q)}\,\mathrm{d}s.\end{aligned}$$

We use this inequality in (22) to obtain

$$\begin{aligned}&\frac{1}{2}\|\mathbf{w}(t)\|^2 + \frac{1}{2}\delta^2\|\nabla \mathbf{w}(t)\|^2 + \nu\delta^2 \int_0^t \|\Delta \mathbf{w}(s)\|^2\,\mathrm{d}s \\ &\quad + \frac{\nu}{2}\int_0^t \|\nabla \mathbf{w}(s)\|^2\,\mathrm{d}s \leq \frac{2\|\mathrm{G}_N\|^2_{L^2(Q)}}{\nu}\int_0^t \|\mathbf{f}(\mathbf{s})\|^2_{V'}\,\mathrm{d}s \\ &\quad + \frac{1}{2}\delta^2\|\mathbf{w}_0\|^2 + \frac{1}{2}\|\nabla \mathbf{w}_0\|^2 + \frac{\delta^2}{2}\|\mathrm{G}_{N-1}\|_{L^2}\|\nabla \mathbf{w}_0\|^2_{L^2}. \qquad \square\end{aligned}$$

REMARK 4.1. *The turbulence model based on the approximate deconvolution procedure introduced by Stolz and Adams in* [9] *and tested by Stolz, Adams, and Kleizer in* [10] *and* [11] *contains a relaxation term added to the right-hand side of the first equation in* (14) *to drain energy near cutoff length scale. This term takes the form*

$$-\chi_\omega(I - G_N G)\mathbf{w}, \tag{23}$$

where $\chi_\omega > 0$ is a function of space and time. If χ_ω is smooth and bounded in space and time, the addition of the relaxation term (23) *does not change the mathematical results proved in this paper. For this model one can derive an energy estimate like Lemma* 4.1 *by treating the weak form of the relaxation term in the following way:*

$$\begin{aligned}&(-\chi_\omega(I - G_N G)\mathbf{w}, (-\delta^2\Delta + I)G_N\mathbf{w}) \\ &= (-\chi_\omega(I - G_N G)\mathbf{w}, G_N\mathbf{w}) + (\chi_\omega(I - G_N G)\mathbf{w}, \delta^2 G_N \Delta\mathbf{w}) \\ &\leq C(\varepsilon, \|\chi_\omega\|_\infty, N, \delta, \nu, \|G\|)\|\mathbf{w}\|^2 + \varepsilon\|\Delta\mathbf{w}\|^2\end{aligned}$$

for given $\varepsilon > 0$

One can then pick $\varepsilon = \frac{\nu\delta^2}{4}$, which results in the cancellation of the term $\frac{\nu\delta^2}{4}\|\Delta\mathbf{w}\|$ on the right-hand side (see (21)*) and apply the Gronwall lemma to get an energy inequality similar to* (17)*.*

Based on this energy inequality all other results proved here for the Stolz–Adams model without the relaxation term can be extended to the case where the relaxation term is incorporated into the equations.

For less regular functions $\chi_\omega > 0$ the same results cannot be proved with the same arguments; this case requires further investigation.

PROPOSITION 4.1. *Let $T > 0$. Then for $\mathbf{w}_0 \in \overline{\mathbb{H}}^2(Q) \cap \mathbb{H}$ and $\mathbf{f} \in L^2(0, T; \mathbb{V}')$, there exists a weak solution $\mathbf{w}$ of* (14) *in the sense of Definition* 4.2*. This solution $\mathbf{w}$ belongs to the space $L^2(0, T, \overline{\mathbb{H}}^2(Q)) \cap L^\infty(0, T; \mathbb{V})$; it is L^2-weakly continuous and satisfies the following energy inequality:*

$$\begin{aligned}&\frac{1}{2}(\|\mathbf{w}(t)\|^2 + \delta^2\|\nabla \mathbf{w}(t)\|^2) + \delta^2\nu\int_0^t \|\Delta\mathbf{w}(s)\|^2\,\mathrm{d}s \\ &\quad \leq K\bigg(\int_0^t \|\mathbf{f}(\mathbf{s})\|^2_{V'}\,\mathrm{d}s + \|\mathbf{w}_0\|^2 + \|\nabla \mathbf{w}_0\|^2\bigg)\end{aligned} \tag{24}$$

for all $t \in [0,T]$ *with* $K = \max\{\frac{2\|G_N\|^2_{L^2(Q)}}{\nu}, \frac{1}{2}\delta^2, \frac{1}{2}, \frac{\delta^2}{2}\|G_{N-1}\|_{L^2(Q)}\}$.

Proof. The proof uses the Faedo–Galerkin method. We will use Galdi [3] as a reference and only point out the differences between the proof of existence of the weak solution of the Navier–Stokes equations and the proof of existence for our models. We pick an orthonormal basis $\{\psi_j\}_j \in \mathbb{D}(Q)$ of $\mathbb{H}$ consisting of eigenfunctions of the Laplacian operator as in Remark (3.1). Let

$$\mathbf{w}_k(x,t) = \sum_{r=1}^{k} \eta_{kr}(t)\psi_r(x) \tag{25}$$

for $k \in \mathbb{N}$ be the solution of the following ODE system:

$$\left(\frac{\partial \mathbf{w}_k}{\partial t}, \psi_r\right) + \nu(\nabla \mathbf{w}_k, \nabla \psi_r) + (\nabla \cdot \overline{(G_N \mathbf{w}_k)(G_N \mathbf{w}_k)^T}, \psi_r) = (\mathbf{f}, \psi_r) \tag{26}$$

for all $r = 1, \ldots, k$ with the initial condition

$$(\mathbf{w}_k(0), \psi_r) = (\mathbf{w}_0, \psi_r)$$

for all $r = 1, \ldots, k$. It follows that the coefficients η_{kr} satisfy the following ODE system:

$$\frac{\mathrm{d}\eta_{kr}}{\mathrm{d}t} + \sum_{i=1}^{k} a_{ir}\eta_{ki} + \sum_{i,j=1}^{k} a_{ijr}\eta_{ki}\eta_{kj} = f_r \tag{27}$$

for all $r = 1, \ldots, k$ with the initial condition

$$\eta_{kr}(0) = C_{0r} \quad \text{for all } r = 1, \ldots, k,$$

where $a_{ir} = \nu(\nabla\psi_i, \nabla\psi_r), a_{ijr} = (\nabla \cdot \overline{((G_N\psi_i)(G_N\psi_j)^T)}, \psi_r), f_r = (\mathbf{f}, \psi_r)$, and $C_{0r} = (bw_0, \psi_r)$.

The function f_r belongs to $L^2[0,T)$ for any r, and consequently (27) has a unique solution near 0,

$$\eta_{kr} \in W^{1,2}(0, T_k),$$

where $T_k \leq T$. Because $\mathbf{w}_0 \in \overline{\mathbb{H}}^2(Q) \cap \mathbb{H}$ there exists $\mathbf{u}_0 \in \mathbb{H}$ such that $\overline{\mathbf{u}}_0 = \mathbf{w}_0$. For the ODE defined above we have $(\mathbf{w}_{k0}, \psi_r) = (\mathbf{w}_0, \psi_r)$ for all $r = 1, \ldots, k$. This gives

$$(\mathbf{w}_{k0}, \psi_r) = (\overline{\mathbf{u}}_0, \psi_r) \tag{28}$$

for all $r = 1, \ldots, k$. But $\mathbf{w}_{k,0} \in G_k = \mathrm{span}\{\psi_j\}_{j=1,\ldots,k}$ and G_k is an invariant subspace of the Laplacian operator. Consequently, we can replace ψ_r in formula (28) with $(I - \delta^2\Delta)\mathbf{w}_{k,0}$ to get

$$(\mathbf{w}_{k0}, (I - \delta^2\Delta)w_{k,0}) = (\overline{\mathbf{u}}_0, (I - \delta^2\Delta)\mathbf{w}_{k,0}) = (\mathbf{u}_0, \mathbf{w}_{k,0}). \tag{29}$$

Integrating by parts the first term above and using Cauchy's inequality in the second, we get

$$\|\mathbf{w}_{k0}\|^2 + \delta^2\|\nabla \mathbf{w}_{k0}\|^2 = (\mathbf{u}_0, \mathbf{w}_{k,0}) \leq \frac{1}{2}(\|\mathbf{u}_0\|^2 + \|\mathbf{w}_{k0}\|^2), \tag{30}$$

which gives the following estimate:

$$\frac{1}{2}\|\mathbf{w}_{k0}\|^2 + \delta^2 \|\nabla \mathbf{w}_{k0}\|^2 \le \frac{1}{2}\|\mathbf{u}_0\|^2 . \tag{31}$$

We want to prove that we can pick $T_k = T$. In (26) we replace ψ_r with $(\mathrm{I} - \delta^2\Delta)\mathrm{G}_N\mathbf{w}_k$. We can do this since $(\mathrm{I}-\delta^2\Delta)\mathrm{G}_N\mathbf{w}_k(t) \in \mathrm{G}_k = \mathrm{span}\{\psi_j\}_{j=1,\ldots,k}$ for any $t \in [0,T)$. In the same way in which the energy inequality (17) for strong solutions was derived, we obtain

$$\frac{1}{2}(\|\mathbf{w}_k(t)\|^2 + \delta^2\|\nabla\mathbf{w}_k(t)\|^2) + \delta^2\nu\int_0^t \|\Delta\mathbf{w}_k(s)\|^2\,\mathrm{d}s + \frac{\nu}{2}\int_0^t \|\Delta\mathbf{w}_k(s)\|^2\,\mathrm{d}s \le M, \tag{32}$$

where

$$M := K\left(\int_0^t \|\mathbf{f}(\mathbf{s})\|_{V'}^2\,\mathrm{d}s + \|\mathbf{w}_{k0}\|^2 + \delta^2\|\nabla\mathbf{w}_{k0}\|^2\right) \tag{33}$$

with $K = \max\{\frac{2\|\mathrm{G}_N\|^2_{L^2(Q)}}{\nu}, \frac{1}{2}\delta^2, \frac{1}{2}, \frac{\delta^2}{2}\|\mathrm{G}_{N-1}\|_{L^2(Q)}\}$.

M does not depend on t and using (31) M also does not depend on k. Due to orthonormality of the family $\{\psi_j\}_j$ in H we get that a priori the coefficients η_{kr} satisfy

$$|\eta_{kr}|^2 \le 2M^{\frac{1}{2}}$$

for any $t \in [0,T), r = 1,\ldots,k$, and $k \in \mathbb{N}$. This implies that for any k there exists a global solution (that is, on $[0,T)$)

$$\eta_{kr} \in W^{1,2}[0,T),$$

$r = 1,\ldots,k$, of the ODE system (26).

In the same way as in Galdi [3] one can show, using estimate (32), that there exists a subsequence of $\mathbf{w}_k$ (which is redenoted by $\mathbf{w}_k$) which converges weakly in V uniformly in t to a function $\mathbf{w} \in L^\infty(0,T,\mathbb{V})$. From estimate (32) we infer that the sequence $\mathbf{w}_k$ is bounded in $L^2(0,T,\overline{\mathbb{H}}^2(Q))$; consequently, it contains a subsequence (which is redenoted by $\mathbf{w}_k$) which is weakly convergent to a function $\mathbf{w}' \in L^2(0,T,\overline{\mathbb{H}}^2(Q))$. One can show, taking limits of $\mathbf{w}_k$ in the space $L^2(0,T,\mathbb{L}^2(Q))$, that $\mathbf{w} = \mathbf{w}'$. It follows that $\mathbf{w} \in (\overline{\mathbb{H}}^2(Q) \cap \mathbb{H})^d$.

We can show that $\mathbf{w}$ satisfies the variational equality (16) in the same way as in Galdi [3] taking the limits of $\mathbf{w}_k$ in equality (26). In the case of Stolz–Adams models, when taking limits, the nonlinear term is handled in the following way: one needs to show that for a given eigenfunction ψ_r,

$$\int_0^t (\overline{\mathrm{G}_N\mathbf{w}_k \cdot \nabla\mathrm{G}_N\mathbf{w}_k}, (\mathrm{I}-\delta^2\Delta)^{-1}\psi_r) - (\overline{\mathrm{G}_N\mathbf{w}\cdot\nabla\mathrm{G}_N\mathbf{w}}, (\mathrm{I}-\delta^2\Delta)^{-1}\psi_r)\,\mathrm{d}s \to 0.$$

However,

$$\begin{aligned}&\left|\int_0^t (\overline{\mathrm{G}_N\mathbf{w}_k \cdot \nabla\mathrm{G}_N\mathbf{w}_k}, (\mathrm{I}-\delta^2\Delta)^{-1}\psi_r) - (\overline{\mathrm{G}_N\mathbf{w}\cdot\nabla\mathrm{G}_N\mathbf{w}}, (\mathrm{I}-\delta^2\Delta)^{-1}\psi_r)\,\mathrm{d}s\right|\\ &\quad = \left|\int_0^t (\mathrm{G}_N\mathbf{w}_k\cdot\nabla\mathrm{G}_N\mathbf{w}_k, \psi_r) - (\mathrm{G}_N\mathbf{w}\cdot\nabla\mathrm{G}_N\mathbf{w}, \psi_r)\,\mathrm{d}s\right|\end{aligned}$$

$$\begin{aligned}&\le \left|\int_0^t (\mathrm{G}_N(\mathbf{w}_k-\mathbf{w})\cdot\nabla \mathrm{G}_N\mathbf{w}_k,\psi_r)\,\mathrm{d}s\right| + \left|\int_0^t (\mathrm{G}_N\mathbf{w}\cdot\nabla \mathrm{G}_N(\mathbf{w}_k-\mathbf{w}),\psi_r)\,\mathrm{d}s\right|\\
&\le \|\mathrm{G}_N\|^2_{L^2(Q)}\|\mathbf{w}_k-\mathbf{w}\|_{L^2(0,T,L^2)}\|\psi_r\|_\infty\|\nabla\mathbf{w}_k\|_{L^2(0,T,L^2)}\\
&\quad+\left|\int_0^t (\mathrm{G}_N\mathbf{w}\cdot \mathrm{G}_N(\nabla(\mathbf{w}_k-\mathbf{w})),\psi_r)\,\mathrm{d}s\right|.\end{aligned}$$

The first term on the right-hand side above converges to 0 since $\mathbf{w}_k \to \mathbf{w}$ in $L^2(0,T,\mathbb{L}^2(Q))$, and the second converges to 0 because $\nabla\mathbf{w}_k \to \nabla\mathbf{w}$ weakly in $L^2(0,T,\mathbb{L}^2(Q))$ and the operator G_N is self-adjoint. The energy inequality (24) is obtained in the same way as in the case of the Navier–Stokes equations taking limits in (32). ∎

LEMMA 4.2. *The weak solution* $\mathbf{w}$ *that was constructed in the previous theorem is also a strong solution of* (14).

Proof. This follows directly from definition (16), the regularity proven for the solution, and an integration by parts. ∎

LEMMA 4.3. *The weak solution* $\mathbf{w}$ *of* (14) *constructed in Proposition* 4.1 *is the unique weak solution of* (14).

Proof. This is a consequence of the regularity of $\mathbf{w}$. The proof is the same as in the case of the Navier–Stokes equations. ∎

5. An a priori estimate of the modeling error. Our goal here is to give an a priori estimate of the modeling error $\|\overline{\mathbf{u}}-\mathbf{w}\|$. In this direction there are several fundamental problems. First, in three dimensions there is no proof of uniqueness of weak solutions $\mathbf{u}$ of the Navier–Stokes equations. Thus for $\mathbf{u}$ a general weak solution of the Navier–Stokes equations, the best result attainable in the usual norms with the present technique seems to be the following.

PROPOSITION 5.1. *Let* $\mathbf{w}=\mathbf{w}(\delta)$ *be the unique strong solution of the model* (14). *Then there is a subsequence* $\delta_j\to 0$ *as* $j\to\infty$ *and a weak solution* $\mathbf{u}$ *of the Navier–Stokes equations such that* $\mathbf{w}(\delta_j)\to\mathbf{u}$ *in* $L^\infty(0,T,\mathbb{L}^2(Q))\cap L^2(0,T,\mathbb{H}^1(Q))$.

Proof. This proof follows that of Theorem 3.1 of Layton and Lewandowski [6]. ∎

The second question concerns the right norm. Obviously if we are restricting our attention to general weak solutions, the right norm must be a very weak norm for which the modeling residual $\left\|\mathbf{u}\mathbf{u}^T-\mathrm{G}_N\overline{\mathbf{u}}(\mathrm{G}_N\overline{\mathbf{u}})^T\right\|$ is not only well defined but also vanishes as $\delta\to 0$. The answer to this question is still unknown; see, e.g., Layton and Lewandowski [6] for first steps. The third question concerns extracting a rate of convergence for $\|\overline{\mathbf{u}}-\mathbf{w}\|$ which gives some insight into the model's accuracy on the laminar regions. This problem is much simpler. It reduces to proving the highest possible rate of convergence for $\|\overline{\mathbf{u}}-\mathbf{w}\|\to 0$ for very smooth solution $\mathbf{u}$.

In the remainder of this subsection we give the answer: the modeling error is a priori $O(\delta^{2N+2})$ for smooth $\mathbf{u}$.

PROPOSITION 5.2. *Assume* $\mathbf{u}$ *is a weak solution of the Navier–Stokes equations and* $\nabla\mathbf{u}\in L^4(0,T,\mathbb{L}^2(Q))$. *For* $\mathbf{w}\in L^2(0,T,\overline{\mathbb{H}}^2(Q))\cap L^\infty(0,T;\mathbb{V})$ *a weak solution of* (14) *and* $\tau := \mathbf{u}\mathbf{u}^T-\mathrm{G}_N\overline{\mathbf{u}}(\mathrm{G}_N\overline{\mathbf{u}})^T$ *there exists a positive constant* $P=P(\nu,N,\|\nabla\mathbf{u}\|_{L^4(0,T,\mathbb{L}^2)})\ge 0$ *such that*

(34)
$$\|\overline{\mathbf{u}}-\mathbf{w}\|^2_{L^\infty(0,T,\mathbb{L}^2)}+\|\nabla(\overline{\mathbf{u}}-\mathbf{w})\|^2_{L^2(0,T,\mathbb{L}^2)}\le P(\nu,N,\|\nabla\mathbf{u}\|_{L^4(0,T,\mathbb{L}^2)})\,\|\tau\|^2_{L^2(0,T,\mathbb{L}^2)}\,.$$

Proof. To begin we derive an equation for $\phi := \overline{\mathbf{u}} - \mathbf{w}$. First we note that $\mathbf{w}$ is a unique strong solution of the model and under stated regularity assumptions on $\mathbf{u}, \mathbf{u}$ is a unique strong solution of the Navier–Stokes equations; see [12, Remark 3.3]. Thus there are no subtleties in the derivation of the error equation. Equality (3) can be rewritten as

$$\begin{aligned} \overline{\mathbf{u}}_t + \nabla \cdot (\overline{\mathrm{G}_N \overline{\mathbf{u}} \mathrm{G}_N \overline{\mathbf{u}}^T}) - \nu \Delta \overline{\mathbf{u}} + \nabla \overline{p} &= \overline{\mathbf{f}} + \nabla \cdot (\overline{\mathrm{G}_N \overline{\mathbf{u}} \mathrm{G}_N \overline{\mathbf{u}}^T - \mathbf{u}\mathbf{u}^T}), \\ \nabla \cdot \overline{\mathbf{u}} &= 0. \end{aligned} \tag{35}$$

Subtraction gives the equation for $\varphi := \overline{\mathbf{u}} - \mathbf{w}$,

$$\begin{aligned} \overline{\varphi}_t + \nabla \cdot \overline{(\mathrm{G}_N \overline{\mathbf{u}} \mathrm{G}_N \overline{\mathbf{u}}^T - \mathrm{G}_N \mathbf{w} \mathrm{G}_N \mathbf{w}^T)} - \nu \Delta \varphi + \nabla \overline{(p-q)} &= -\nabla \cdot \overline{\tau} && \text{in } (0,T) \times \mathbb{R}^d, \\ \nabla \cdot \varphi &= 0 && \text{in } (0,T] \times \mathbb{R}^d, \\ \varphi|_{t=0} &= 0 && \text{in } \mathbb{R}^d, \\ \int_Q p - q \,\mathrm{d}x &= 0 && \text{in } (0,T]. \end{aligned} \tag{36}$$

We multiply the first equation in (36) by $(\mathrm{I} - \delta^2 \Delta)^{-1} \mathrm{G}_N \varphi$ and then integrate on Q. Following exactly the same computations as in Lemma 4.1 gives

$$\begin{aligned} &\frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|\varphi\|^2 + \frac{1}{2} \delta^2 \frac{\mathrm{d}}{\mathrm{d}t} \|\nabla \varphi\|^2 + \frac{\delta^2}{2} \frac{\mathrm{d}}{\mathrm{d}t} (\nabla \varphi, \mathrm{G}_{N-1} \nabla \varphi) + \nu \|\nabla \varphi\|^2 + \nu \delta^2 \|\Delta \varphi\| \\ &\quad + \nu \delta^4 (\Delta \phi, \mathrm{G}_{N-1} \Delta \varphi) = -(\nabla \cdot \tau, \mathrm{G}_N \varphi) + b(\mathrm{G}_N \varphi, \mathrm{G}_N \overline{\mathbf{u}}, \mathrm{G}_N \varphi), \end{aligned} \tag{37}$$

where b is the standard trilinear form

$$b(\mathbf{u}, \mathbf{v}, \mathbf{w}) = ((\mathbf{u} \cdot \nabla) \mathbf{v}, \mathbf{w}).$$

The first term on the right-hand side is bounded as follows:

$$|(\nabla \cdot \tau, \mathrm{G}_N \varphi)| = |(\tau, \mathrm{G}_N \nabla \varphi)| \leq \|\tau\| \, \|\mathrm{G}_N\|_{L^2} \|\nabla \varphi\| \leq \frac{2 \|\mathrm{G}_N\|_{L^2}^2}{\nu} \|\tau\|^2 + \frac{1}{2} \nu \|\nabla \varphi\|^2.$$

To bound the second term we use Young's inequality

$$ab \leq \epsilon a^4 + \frac{3}{4} (4\epsilon)^{-\frac{1}{3}} b^{\frac{4}{3}},$$

together with the standard estimate for the trilinear form

$$|b(\mathrm{G}_N \varphi, \mathrm{G}_N \overline{\mathbf{u}}, \mathrm{G}_N \varphi)| \leq C(Q) \|\nabla \overline{\mathbf{u}}\| \, \|\varphi\|^{\frac{1}{2}} \|\nabla \varphi\|^{\frac{3}{2}},$$

to obtain that for any $\epsilon > 0$,

$$|b(\mathrm{G}_N \varphi, \mathrm{G}_N \overline{\mathbf{u}}, \mathrm{G}_N \varphi)| \leq \epsilon \|\mathrm{G}_N\|^2 \|\nabla \varphi\|^2 + \frac{3}{4} (4\epsilon)^{-\frac{1}{3}} \|\mathrm{G}_N \nabla \overline{\mathbf{u}}\|^4 \|\mathrm{G}_N \varphi\|^2.$$

Plugging $\epsilon = \frac{\nu}{2\|\mathrm{G}_N\|^2}$ into the above inequality, we get that

$$|b(\mathrm{G}_N \varphi, \mathrm{G}_N \overline{\mathbf{u}}, \mathrm{G}_N \varphi)| \leq \frac{\nu}{2} \|\nabla \varphi\|^2 + \frac{3}{4} \left(\frac{2\nu}{\|\mathrm{G}_N\|^2} \right)^{-\frac{1}{3}} \|\mathrm{G}_N\|^4 \|\nabla \overline{\mathbf{u}}\|^4 \|\varphi\|^2.$$

Using the last two inequalities in (37) gives

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\varphi\|^2 + \frac{1}{2}\delta^2\frac{\mathrm{d}}{\mathrm{d}t}\|\nabla\varphi\|^2 + \frac{\delta^2}{2}\frac{\mathrm{d}}{\mathrm{d}t}(\nabla\varphi, \mathrm{G}_{N-1}\nabla\varphi) + \nu\delta^2\|\Delta\varphi\|$$
$$+\nu\delta^4(\Delta\phi, \mathrm{G}_{N-1}\Delta\varphi) \le \frac{2\|\mathrm{G}_N\|^2_{L^2}}{\nu}\|\tau\|^2 + \frac{3}{4}(2\nu)^{-\frac{1}{3}}\|\mathrm{G}_N\|^{\frac{10}{3}}\|\nabla\overline{\mathbf{u}}\|^4\|\mathbf{w}\|^2.$$

Gronwall's inequality and positivity of the operators $(\mathrm{G}_N)_N$ give

$$\|\varphi\|^2 \le \int_0^t \exp\left(-2\int_s^t \frac{3}{4}(2\nu)^{-\frac{1}{3}}\|\mathrm{G}_N\|^{\frac{10}{3}}\|\nabla\overline{\mathbf{u}}\|^4\,\mathrm{d}s'\right)\frac{2\|\mathrm{G}_N\|^2_{L^2}}{\nu}\|\tau\|^2_{L^2}\,\mathrm{d}s.$$

For fixed N we have that

$$\|\mathrm{G}_N\| \le 1 + (1+\|\mathrm{G}\|) + (1+\|\mathrm{G}\|)^2 + \cdots + (1+\|\mathrm{G}\|)^N,$$

and since for every δ, $\|\mathrm{G}\| \le 1$ it follows that

$$\|\mathrm{G}_N\| \le 2^{N+1} - 1$$

uniformly in δ. Under the assumption that $\nabla\mathbf{u} \in L^4(0,T,L^2)$ we infer the existence of a constant $M = M(\nu, N, \|\nabla\mathbf{u}\|_{L^4(0,T,L^2)})$ such that

$$\|\varphi\|^2_{L^\infty(0,T,L^2)} \le M\big(\nu, N, \|\nabla\mathbf{u}\|_{L^4(0,T,L^2)}\big)\int_0^T \|\tau\|^2_{0,T,L^2}. \tag{38}$$

To estimate $\|\nabla\varphi\|^2_{L^2(0,T,L^2)}$ we integrate (37) from 0 to t and, using inequality (38), we obtain

$$\|\nabla\phi\|^2_{L^2(0,T,L^2)} \le R(\nu, N, \|\nabla\mathbf{u}\|_{L^4(0,T,L^2)})\int_0^T \|\tau\|^2_{0,T,L^2}$$

for positive constant $R = R(\nu, N, \|\nabla\mathbf{u}\|_{L^4(0,T,L^2)})$. Consequently, there exists a constant $P = P(\nu, N, \|\nabla\mathbf{u}\|_{L^4(0,T,L^2)})$ such that

$$\|\varphi\|^2_{L^\infty(0,T,L^2)} + \|\nabla\phi\|^2_{L^2(L^2)} \le P(\nu, N, \|\nabla\mathbf{u}\|_{L^4(0,T,L^2)})\|\tau\|^2_{L^2(0,T,L^2)}. \tag{39}$$

PROPOSITION 5.3. *Under the conditions of the previous theorem, if* $\mathbf{u} \in \mathbb{H}^{N+1}(Q)$, *there exists* $P = P(\nu, N, \mathbf{u}) \ge 0$ *such that*

$$\|\overline{\mathbf{u}} - \mathbf{w}\|^2_{L^\infty(0,T,L^2)} + \|\nabla(\overline{\mathbf{u}} - \mathbf{w})\|^2_{L^2(0,T,L^2)} \le P(\nu, N, \mathbf{u})\delta^{2N+2}. \tag{40}$$

Proof. An application of Lemma 2.2 gives

$$\|\tau\|^2_{L^2(0,T,L^2)} \le C(\mathbf{u})\delta^{2N+2};$$

(40) will then follow from (39). ∎

6. Conclusions. The Stolz–Adams deconvolution models analyzed herein are shown to have very good mathematical properties, better than any other large-eddy simulation model for turbulent flows that is currently used.

There exists a weak solution of these models; that solution is unique, and further it is shown that it belongs to higher order Sobolev spaces and that it is also the strong solution of the models.

We proved that the Stolz–Adams models give a good description of the local spatial averages of fluid velocities, the modeling error converges to 0, and the rate of convergence is also derived.

This paper provides the mathematical foundations of the Stolz–Adams models, giving guidance for practical computations with these models.

REFERENCES

[1] N.A. Adams and S. Stolz, *Deconvolution methods for subgrid-scale approximation in large-eddy simulation*, in Modern Simulation Strategies for Turbulent Flow, R.T. Edwards, Philadelphia, 2001, pp. 21–41.

[2] M. Bertero and P. Boccacci, *Introduction to Inverse Problems in Imaging*, IOP Publishing, Bristol, UK, 1998.

[3] G.P. Galdi, *Lectures in Mathematical Fluid Dynamics*, Birkhäuser-Verlag, Basel, Switzerland, 2000.

[4] M. Germano, *Differential filters of elliptic type*, Phys. Fluids, 29 (1986), pp. 1757–1758.

[5] J.L. Guermond, J. Tinsley Oden, and S. Prudhomme, *Mathematical perspectives on large eddy simulation models for turbulent flows*, J. Math. Fluid Mech., 6 (2004), pp. 194–248.

[6] W. Layton and R. Lewandowski, *Analysis of the Zeroth order model for Large Eddy Simulation of Turbulence*, technical report, 2003.

[7] W. Layton and R. Lewandowski, *A simple and stable scale similarity model for large eddy simulation: Energy balance and existence of weak solutions*, Appl. Math. Lett., 16 (2003), pp. 1205–1209.

[8] P. Sagaut, *Large Eddy Simulation for Incompressible Flows*, Springer-Verlag, Berlin, Heidelberg, New York, 2001.

[9] S. Stolz and N.A. Adams, *An approximate deconvolution procedure for large eddy simulation*, Phys. Fluids, 11 (1999), pp. 1699–1701.

[10] S. Stolz, N.A. Adams, and D. Kleiser, *An approximate deconvolution model for large-eddy simulation with application to incompressible wall-bounded flows*, Phys. Fluids, 13 (2001), pp. 997–1015.

[11] S. Stolz, N.A. Adams, and D. Kleiser, *The approximate deconvolution model for large-eddy simulations of compressible flows and its applications to shock-turbulent-boundary-layer interaction*, Phys. Fluids, 13 (2001), pp. 2985–3001.

[12] R. Temam, *Navier–Stokes Equations and Nonlinear Functional Analysis*, SIAM, Philadelphia, 1995.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1903–1922

LATTICE POINTS ON CIRCLES AND DISCRETE VELOCITY MODELS FOR THE BOLTZMANN EQUATION*

LAURA FAINSILBER†, PÄR KURLBERG‡, AND BERNT WENNBERG†

Abstract. The construction of discrete velocity models or numerical methods for the Boltzmann equation, may lead to the necessity of computing the collision operator as a sum over lattice points. The collision operator involves an integral over a sphere, which corresponds to the conservation of energy and momentum. In dimension two there are difficulties even in proving the convergence of such an approximation since many circles contain very few lattice points, and some circles contain many badly distributed lattice points. However, by showing that lattice points on most circles are equidistributed we find that the collision operator can indeed be approximated as a sum over lattice points in the two-dimensional case. The proof uses a weak form of the Halberstam–Richert inequality for multiplicative functions (a proof is given in the paper), and estimates for the angular distribution of Gaussian primes. For higher dimensions, this result has already been obtained by Palczewski, Schneider, and Bobylev [*SIAM J. Numer. Anal.*, 34 (1997), pp. 1865–1883].

Key words. Boltzmann equation, discrete velocity model, multiplicative functions, distribution of Gaussian primes

AMS subject classifications. 82C40, 11L07, 11N64, 11E25

DOI. 10.1137/040618916

1. Introduction.

The phase space density f of a dilute gas evolves according to the Boltzmann equation. In the physically relevant case, the gas would be confined to a subset $\Omega \subset \mathbb{R}^3$, and then $f(x,v,t){:}\Omega\times\mathbb{R}^3\times\mathbb{R}^+ \to \mathbb{R}^+$, where x denotes a position in space, $v \in \mathbb{R}^3$ is a velocity, and t denotes the time. From a mathematical point of view, it is equally natural to consider the Boltzmann equation in any spatial dimension, and in some cases because of symmetries of Ω, it is also relevant to consider $\Omega \subset \mathbb{R}^{d_1}$ and $v \in \mathbb{R}^{d_2}$ with $d_1 < d_2$.

By a dilute gas we mean one where the particles interact with each other essentially only by *pairwise* interactions. Moreover, the Boltzmann equation assumes that the particles are so small compared to other distances, that they can be considered to be points.

Under these hypotheses, one can formally derive the Boltzmann equation (see [7])

$$\partial_t f(x,v,t) + v\cdot\nabla_x f(x,v,t) = Q(f,f)(x,v,t). \tag{1}$$

The left-hand side describes the evolution of the density by free transport, and the right-hand side describes the impact of collisions. Per definition, a collision is a pairwise interaction that takes place instantaneously and at one single point in space.

*Received by the editors November 15, 2004; accepted for publication (in revised form) August 15, 2005; published electronically March 3, 2006. The research of the second and third authors was supported in part by the Swedish Research Council.

http://www.siam.org/journals/sima/37-6/61891.html

†Department of Mathematics, Chalmers University of Technology, SE-41296 Gothenburg, Sweden (laura@math.chalmers.se, wennberg@math.chalmers.se). The third author was partially supported by the EC-funded RTN network HYKE, contract HPRN-CT-2002-00282.

‡Department of Mathematics, Royal Institute of Technology, SE-10044 Stockholm, Sweden (kurlberg@math.kth.se). This author was partially supported by the National Science Foundation (DMS 0071503) and the Royal Swedish Academy of Sciences.

Hence x and t appear only as parameters in $Q(f,f)$, and we can write

$$Q(f,f)(v) = \int_{\mathbb{R}^d}\int_{S^{d-1}} \left(f(v')f(v'_*) - f(v)f(v_*)\right) q(|w|,\cos\theta)\, dS(u)dv_*, \tag{2}$$

where the velocities "before and after a collision" are related by

$$\begin{aligned} v' &= \tfrac{1}{2}(v+v_*) + |w|\,u \\ v'_* &= \tfrac{1}{2}(v+v_*) - |w|\,u, \end{aligned} \tag{3}$$

with $w = (v_* - v)/2$, and with $\cos\theta = \frac{u\cdot w}{|w|}$; dv_* is the Lebesgue measure in $\mathbb{R}^d$, and $dS(u)$ is the surface measure on S^{d-1}. Note that the pair of velocities before a collision, v and v_*, and the pair of velocities after the collision, v' and v'_*, are the endpoints of a diameter on the sphere which has its center at $\frac{v+v_*}{2}$ and diameter $|v_* - v|$. This is exactly the condition needed in order that the collisions preserve the momentum and energy of the pair of particles. For $d = 2$, the sphere becomes a circle, and this motivates the title of the paper.

In a discrete velocity model (DVM), the velocities are concentrated on a (usually finite) set of points $v_j \in \mathbb{R}^d$ in the velocity space:

$$f(x,v,t) = \sum_j f_j(x,t)\delta_{v=v_j}.$$

The Boltzmann equation (1) is then changed into a nonlinear system of conservation laws,

$$\partial_t f_j + v_j \cdot \nabla_x f_j = \sum_{k,k',j'} \Gamma_{j,k}^{j',k'}\left(f_{j'}f_{k'} - f_j f_k\right), \tag{4}$$

where the constants $\Gamma_{j,k}^{j',k'} \geq 0$ must be chosen so that (4) makes sense from a physical point of view. In particular, we require that (v_j, v_k) and $(v_{j'}, v_{k'})$ define two diameters on the same sphere, just as for the usual Boltzmann equation.

The first example of a discrete velocity model is that of Carleman [5], which has two velocities in $\mathbb{R}$. Many other models have been proposed, and there is vast literature on how to construct and analyze physically realistic models (i.e., that satisfy the right conservation laws and an entropy principle); see, e.g., [4, 24, 25].

In this paper we are not mainly concerned with the question of whether the model we study is correct in this sense (this has actually been demonstrated in [4]), and so we defer our discussion on this matter to the last section of the paper.

Besides offering many interesting mathematical challenges (for example, there is no general theory of global existence of solutions to systems like (4)), the DVMs are also candidates for the numerical approximation of the real Boltzmann equation (1). This leads naturally to the following question, which is the subject matter of the paper.

Suppose that we choose the discrete set of velocities to be $h\mathbb{Z}^d$, i.e., the integer lattice in $\mathbb{R}^d$, scaled by a factor h, and that we take

$$f^h(v) = \sum_{\xi\in\mathbb{Z}^d} f_{\xi,h}\delta_{v=h\xi},$$

so that $f^h \to f$, in some suitable sense, where $f \in L^1(\mathbb{R}^d)$. Is it then true that $Q(f^h, f^h)(v) \to Q(f,f)(v)$ for all $v \in h\mathbb{Z}^d$ when $h \to 0$? (This property, which is called *consistency*, together with *stability* are the main ingredients when proving that a numerical method converges.) The answer is yes. This was proven by Palczewski, Schneider, and Bobylev [3] for dimensions $d \geq 3$ (see also [2]). In this paper, we prove that it is also true for $d = 2$, and hence for all relevant cases.

Results of this kind are interesting, because, together with the corresponding existence results [21], they provide examples that are relevant to previous results of Desvillettes and Mischler [8], who proved that solutions to families of DVMs can converge to DiPerna–Lions' solutions to (1) if certain conditions are satisfied.

Our result should not, however, be considered as relevant for numerical analysis, because the rate of convergence is so slow that a numerical method based on the theory presented here would hardly ever become useful.

The family of models considered here can be seen as coming from a rather straightforward discretization of the collision integral (2). This integral should be interpreted as an average over the $(2d-1)$-dimensional manifold defined by

$$\mathcal{M}_v = \big\{(v_*, v', v'_*) \in \mathbb{R}^{3d} \text{ such that } v' + v'_* - v_* = v \\ |v'|^2 + |v'_*|^2 - |v_*|^2 = |v|^2\big\}, \tag{5}$$

and (2) is an iterated integral over this manifold. For a fixed v, we write $w = (v_* - v)/2$, and then (3) becomes

$$v' = v + w + |w|u$$
$$v'_* = v + w - |w|u,$$

and also $v_* = v + 2w$. We then write

$$g_v(w, u) = (f(v')f(v'_*) - f(v)f(v_*))\, q(|w|, \cos\theta), \tag{6}$$

and so (after changing variables in the integral),

$$Q(f,f)(v) = 2^d \int_{\mathbb{R}^d} \left(\int_{S^{d-1}} g_v(w, u)\, dS(u) \right) dw.$$

If g is sufficiently regular (continuous), and decays sufficiently rapidly for large w, then the Riemann sum for the outer integral converges:

$$(2h)^d \sum_{\zeta \in \mathbb{Z}^d} \int_{S^{d-1}} g_v(h\zeta, u)\, dS(u) \longrightarrow 2^d \int_{\mathbb{R}^d} \left(\int_{S^{d-1}} g_v(w, u)\, dS(u) \right) dw \tag{7}$$

when $h \to 0$. In order to construct a consistent DVM, it is then sufficient to evaluate the inner integral in terms of the values of g on the lattice points $h\mathbb{Z}^d$, in such a way that the result converges to $\int_{S^{d-1}} g(w, u)\, dS(u)$. While with the formula (3), the collision integral should be taken over all $u \in S^{d-1}$, we have here only access to those u for which v' and v'_* belong to $h\mathbb{Z}^d$. But this is automatically achieved if $\zeta \in \mathbb{Z}^d$, and if $u = \zeta'/|\zeta'|$, where $\zeta' \in \mathbb{Z}^d$ and $|\zeta'| = |\zeta|$; then for all $v \in h\mathbb{Z}^d$,

$$v + h\zeta \pm h|\zeta|u \in h\mathbb{Z}^d.$$

However, note that with this construction, the center of the sphere is restricted to lie on a lattice point, so it excludes cases like $v = (0, 0)$, $v_* = (h, h)$.

Giving all points on the sphere equal weight, one arrives at the expression

$$\frac{1}{r_d(|\zeta|^2)} \sum_{\substack{\zeta' \in \mathbb{Z}^d \\ |\zeta'|=|\zeta|}} \left(f(v')f(v'_*) - f(v)f(v_*)\right) q(|h\zeta|, \cos\theta), \tag{8}$$

for approximating the inner integral in (7). The function $r_d(n)$ denotes the number of points with integer coordinates on a sphere in $\mathbb{R}^d$ with center at the origin and radius $\sqrt{n}$, i.e., the number of integer solutions to $x_1^2 + \cdots + x_d^2 = n$.

We write, for all $v \in h\mathbb{Z}^d$,

$$Q^h(f,f)(v) = (2h)^d \sum_{\zeta \in \mathbb{Z}^d} \frac{1}{r_d(|\zeta|^2)} \sum_{\substack{\zeta' \in \mathbb{Z}^d \\ |\zeta'|=|\zeta|}} \left(f(v')f(v'_*) - f(v)f(v_*)\right) q(|h\zeta|, \cos\theta). \tag{9}$$

In the two-dimensional case, all the terms in the sum are 2π-periodic functions of θ, and assuming sufficient regularity, they can be expressed as a convergent Fourier series. It is then natural to introduce the exponential sum

$$S(n,k) = \sum_{u \in \mathbb{Z}^2 : |u|^2 = n} e^{ik\theta_u}, \tag{10}$$

where θ_u is defined by $u = |u| \cdot (\sin\theta_u, \cos\theta_u)$. We will see in section 4 that to prove that (8) converges to the angular integral in (7), it is enough to prove that for $k \neq 0$, there is sufficient cancellation in the exponential sums $S(n,k)$ on average. Similar exponential sums are relevant for any dimension, and the work of Bobylev et al. [2] and Palczewski et al. [3] also involves such estimates.

Here the needed estimate is given as Proposition 6 in section 3. Then in section 4 we put the estimates together to a proof of the main result.

THEOREM 1. *Consider the Boltzmann equation in two dimensions. Assume that f and q are so smooth that the function $g_v(w,u)$ defined in* (6) *is a C^2-function. Then for all $v \in h\mathbb{Z}^2$*

$$\left|Q(f,f)(v) - Q^h(f,f)(v)\right| \to 0$$

when $h \to 0$.

Section 5, finally, contains a discussion of spurious invariants and of numerical cost. We also illustrate the distribution of points such that the circles passing through them contain many lattice points.

A more general construction of discrete velocity models on scaled integer lattices $h\mathbb{Z}^2$ consists of finding sets of integer points on the manifold $\mathcal{M}$ defined in (5). In this way, mass and energy conservation are automatically satisfied, but one also needs to verify that these are the *only* conserved quantities. Finally, in order that the models converge to the continuous model when $h \to 0$, it is necessary that the integer points are more or less uniformly distributed on $\mathcal{M}$.

The models studied here are constructed by discretizing, one at a time, the iterated integrals (2). An alternative way of writing this integral was introduced by Carleman [5]. Using that $v' - v$ and $v'_* - v$ are orthogonal, one can write (here we specialize to $d = 3$)

$$Q(f,f)(v) = \int_{\mathbb{R}^3} \int_{E_{v,v'}} \left(f(v')f(v'_*) - f(v)f(v_*)\right) q(w, \cos\theta) \frac{1}{|v - v'|^2}\, dE(v'_*)\, dv',$$

where $E_{v,v'}$ is the plane that contains v and is orthogonal to $v'-v$, and where $dE(v'_*)$ is the Euclidean measure on this plane. Panferov and Heintz [17] have analyzed a DVM based on this iterated integral, and proved that the method is consistent with the continuous model. This is somehow easier, because on a given plane, one can find all integer points by solving linear Diophantine equations. However, the density of points depends strongly on $v'-v$, and so it is far from trivial to prove the consistency. Again, the two-dimensional situation is more difficult, and has not yet been studied.

Yet another approach was introduced by Rogier and Schneider [23], who used the theory of Farey series to discretize the angular variable in the collision integral. Some recent, related results on DVMs can be found in [1].

2. Number theoretic background.

2.1. Points on spheres; Asymptotics. To prove that (8) converges to the correct limit when $h \to 0$, one has to study the set

$$\{\zeta/|\zeta| : \zeta \in \mathbb{Z}^d, |\zeta|^2 = n\}$$

and show that the points of this set are sufficiently well distributed on S^{d-1} when n is large; it is here that the number theoretical issues enter the game. Indeed, we can view the set of points with integer coordinates on a sphere of squared radius n centered at the origin,

$$\{(x_1, \dots x_d) \in \mathbb{Z}^d, \sum_{i=1}^{d} x_i^2 = n\},$$

as the solution set for a quadratic form, and use the theory of integral quadratic forms to get estimates on the number of points (see, for instance, [13]). The expected number of points with integer coordinates on a sphere clearly depends on the dimension d. The naive approach to find the order of magnitude for a given dimension is to use the volume of a ball, divided by the number of spheres contained in the ball. The volume of a ball of radius $\sqrt{n}$ grows as $n^{d/2}$ while the number of spheres is n. For $d = 2$, this leads us to expect a constant number of lattice points on circles, for $d = 3$, a growth proportional to $\sqrt{n}$, etc. However, for small d this approach is misleading; the growth is quite irregular and depends on the divisor structure of n. For $d = 2$, we will see below that only values of n of the form $n = 2^s q^2 p_1^{\alpha_1} \dots p_r^{\alpha_r}$, where q is a product of primes of the form $4k+3$ and the p_i's are primes of the form $4k+1$ (see below), yield circles with lattice points, and thus most circles have no points at all. In fact, Landau proved in 1908 that the number of circles with at least one lattice point, of integer squared radius smaller than x, grows as $Cx/\sqrt{\log x}$. Moreover, there are also infinite families of circles with very few lattice points; radii that are a power of 2 yield 4 points for instance, and radii that are the square root of a prime of the form $p = 4k+1$ yield exactly 8 points. On the other hand, the number of lattice points on a circle is not bounded; for instance, a circle with $n = p_1 \dots p_r$ as above, where all the p_i are distinct from each other, has $4 \cdot 2^r$ points.

In dimension 3, all values of n not of the form $n = 4^s(8k+7)$ yield spheres containing points with integer coordinates. This still leaves a fairly large number of spheres having no points, but for our purposes this does not really matter, as such spheres do not appear in the summation formulas (there is no relevant value for ζ). Among the spheres with lattice points, multiplying the radius by a power of 2 does not increase the number of points, but if we correct for this fact, the ratio between the number of points and the naive estimate is bounded, up to constants only depending on ϵ, from above by n^ϵ, and below by $n^{-\epsilon}$ for all $\epsilon > 0$ (see [13, Chap. 4] for exact formulas involving class numbers or L-series.)

The higher-dimensional cases behave in a somewhat more regular fashion. Lagrange proved that every positive integer can be written as the sum of four squares, and thus for dimension $d \geq 4$, every sphere whose squared radius is an integer has lattice points. For $d = 4$, the number of points still oscillates rather wildly (e.g., spheres with radius a power of 2 only have 24 points) but for greater dimensions, the naive estimate gives the correct asymptotic growth of the number of points.

Getting circles (or spheres) with "sufficiently many" lattice points, however, is not quite enough for our purposes: we also require that the lattice points be sufficiently uniformly distributed when projected on the unit sphere. In dimensions 3 and higher, this follows from estimates on Fourier coefficients of modular forms. The case $d \geq 4$, with some restrictions on the set of numbers in which n tends to infinity when $d = 4$, is due to Pommerenke [22]. For $d = 3$, Duke [10] and Golubeva–Fomenko [14] used Iwaniec's [18] estimates on Fourier coefficients of half integral weight forms to obtain uniform distribution. Unfortunately, these techniques do not apply in dimension 2. Moreover, there are circles with a large number of lattice points that are poorly distributed.

Theorem 2 (see Cilleruelo [6]). *For any $\epsilon > 0$ and for any integer k, there exists a circle $x^2 + y^2 = n$ with more than k lattice points such that all the lattice points are on the arcs $\sqrt{n}e^{(\pi/2)(t+\theta)i}$ with $|\theta| < \epsilon$, $t \in \{0, 1, 2, 3\}$.*

On the other hand, we may use some other techniques from analytic number theory to show that lattice points on circles are equidistributed *on average*, and this is good enough for our purpose.

2.2. From points on circles to Gaussian integers. In the plane, we can view lattice points on a circle of radius $\sqrt{n}$, centered at the origin, as complex numbers with integer real and imaginary parts, and squared modulus n. It might seem as a trivial restatement, but doing so allows us to use use some techniques from algebraic number theory. The *Gaussian integers*, i.e., the set

$$\mathbb{Z}[i] = \{x + iy \in \mathbb{C}, (x, y) \in \mathbb{Z}^2\}$$

is the ring of integers of the field $\mathbb{Q}(i)$. It shares an important property with the ordinary integers, namely, unique factorization,[1] i.e., just as every integer in $\mathbb{Z}$ factors into prime numbers, and the factorization is unique up to ordering the primes and multiplying by -1, Gaussian integers factor into Gaussian primes, uniquely up to ordering and multiplication by $-1, i, -i$ (these and 1 are the units, i.e., the elements having a multiplicative inverse in $\mathbb{Z}[i]$). For a more thorough introduction to primes in quadratic number fields, see, for instance, [16, Chap. XV].

The Gaussian primes (i.e., the elements of $\mathbb{Z}[i]$ that cannot be written as a product of Gaussian integers with smaller modulus) are of three types:

- The prime numbers $q \in \mathbb{Z}$ such that $q \equiv 3 \mod 4$ remain prime in $\mathbb{Z}[i]$ (e.g., $3, 7, 11, 19, \ldots$).
- For prime numbers $p \in \mathbb{Z}$ such that $p \equiv 1 \mod 4$, there exists $x, y \in \mathbb{Z}$ such that $p = x^2 + y^2$. Hence p factors in $\mathbb{Z}[i]$ as a product of two Gaussian primes

  $$p = (x + iy)(x - iy).$$

  For example, 5 factors into $(2 + i)(2 - i)$ in $\mathbb{Z}[i]$.

[1] This is rather unusual; the ring of integers in most number fields will not have this property.

- Last (and least!), $1+i$ is prime. (Note that $(1+i)(1-i)=2$ and that $1-i=-i(1+i)$ is merely "another form of the same prime," just as 3 and -3 represent the same prime.)

If n is the sum of two squares, then it can be factored in $\mathbb{Z}[i]$:

$$n = X^2 + Y^2 = (X+iY)(X-iY).$$

If $z = x+iy$ is a prime factor of $X+iY$, then $\bar{z} = x-iy$ must be a prime factor of $X-iY$. It follows that prime factors $q \equiv 3 \mod 4$ of n must appear in even powers. In addition, multiplying n by an even power of a prime q that is congruent with 3 mod 4 changes neither the number of solutions to $n = X^2+Y^2$ nor the distribution of arguments of the solutions.

Suppose now that n contains a factor p^α, where $p \equiv 1 \mod 4$. The number p can be factored in $\mathbb{Z}[i]$ as $(x+iy)(x-iy)$, and hence the multiplicity of $x+iy$ as a factor of n is α, and the same is true for $x-iy$. It follows that the multiplicity of $x+iy$ in $X+iY$ can be any integer j, with $0 \le j \le \alpha$, and the multiplicity of $x-iy$ is then $\alpha - j$.

The same calculation can be done for powers of 2; however, the solutions given by different choices of j in that case differ by a multiplication by a power of i, and so the power of 2 does not influence the number of solutions.

All solutions to $n = X^2+Y^2$ can now be expressed as $X+iY = \sqrt{n}\exp(i\theta)$, where all possible values of the argument θ can be computed as sums of terms deriving from the different factors of n in the following way:

1. $X+iY$ can be multiplied by any unit, i.e., by ± 1 or $\pm i$. This gives a term $k\pi/2$ in the argument, $k = 0,1,2,3$.
2. If the multiplicity of 2 in n is odd, then the argument must contain $\pi/4$, the argument of $1+i$; the number of solutions does not change.
3. For each prime factor $p \equiv 1 \mod 4$ in n, let α_p be the multiplicity of p in n, let $p = x_p^2 + y_p^2$, and set $\theta_p = \arg(x_p + iy_p)$. For a particular choice of j, $0 \le j \le \alpha_p$, the argument added to $X+iY$ is $j\theta_p - (\alpha_p - j)\theta_p = (2j-\alpha_p)\theta_p$.

Since the choices of k and of the different $j's$ are independent the number of different solutions is $4\prod_{p\equiv 1 \bmod 4}(\alpha_p+1)$.

2.3. Results on the distribution of primes and on the angular distribution of points. We will need the results that follow.

Theorem 3 (Merten's theorem, see [16, Chap. 22.8]).

$$\prod_{\substack{p \le x \\ p \text{ prime}}} (1-1/p) \sim e^{-\gamma}/\log x,$$

where $\gamma \simeq 0.57$ is Euler's constant.

As for the angular distribution of Gaussian primes, a result by Kubilyus gives that the angles $\{\theta_p\}_{p\equiv 1 \mod 4}$ are equidistributed in $[0,\pi/4]$ as shown in the following theorem.

Theorem 4 (Kubilyus [20]). *The number of Gaussian primes ω in the sector $0 \le \alpha \le \arg(\omega) \le \beta \le 2\pi$, $|\omega|^2 \le u$ is equal to*

$$\frac{2}{\pi}(\beta-\alpha)\int_2^u \frac{dv}{\log v} + O\left(u\exp(-b\sqrt{\log u})\right),$$

where b is an absolute positive constant.

From Kubilyus' theorem, it is straightforward to deduce (see [11, p. 92]) the following corollary.

COROLLARY 5. *If* $k \in 4\mathbb{N}$ *and* $\log k \le b\sqrt{\log x}$, *then*

$$\sum_{\substack{p \le x \\ p \equiv 1 \mod 4}} \frac{|\cos(k\theta_p)|}{p} \le \frac{1}{\pi} \log\log x + (1 - 2/\pi) \log\log k + O(1).$$

3. Equidistribution of lattice points on circles. What is needed for the proof of consistency of the discrete velocity model are estimates on the equidistribution of lattice points on circles, and the aim of this section is to show that lattice points on circles are equidistributed *on average*.

Let $r(m) = r_2(m)$ be the number of integer solutions to $x^2 + y^2 = m$. We recall the definition of $S(m,k)$:

$$S(m,k) = \sum_{|w'|^2 = m} e^{ik\theta_{w'}}.$$

PROPOSITION 6. *If* $4 \nmid k$, *then* $|S(m,k)| = 0$. *If* $4|k$ *and* $k \neq 0$, *there exist* C *and* $b > 0$ *such that*

$$\log\left(\frac{1}{X}\sum_{m \le X} |S(m,k)|\right) \le C - (1 - 2/\pi)\log\left(\frac{\log X}{(\log|k|)^2}\right)$$

for X *sufficiently large and* $\log|k| \le b\sqrt{\log X}$.

Remark. The mean discrepancy of the distribution of angles of Gaussian integers was studied by Kátai and Környei in [19], and by Erdős and Hall in [11]. Our method is similar to theirs, except that they bound

$$\frac{1}{X/\sqrt{\log X}} \sum_{m \le X} \frac{|S(m,k)|}{r(m)}$$

instead of

$$\frac{1}{X}\sum_{m \le X} |S(m,k)|.$$

The proof is based on the observation that $|S(m,k)|/4$ is a *multiplicative* function, i.e., a function $f : \mathbb{N} \to \mathbb{C}$ such that $f(mn) = f(m)f(n)$ for all m, n such that $(m,n) = 1$. It turns out that the mean value of a multiplicative function, under fairly general circumstances, can be bounded in terms of an exponential of a sum over primes. To make the paper more self-contained, we include a weak form of the *Halberstam–Richert inequality* (cf. [15]).

THEOREM 7. *Let* f *be a nonnegative multiplicative function such that*

$$\sum_{n \le x} f(n) = O(x), \tag{11}$$

and $f(p^k) = O(k)$ *for all primes* p *and* $k \ge 1$. *Then there exists* $C > 0$ *such that*

$$\frac{1}{X}\sum_{m \le X} f(m) \le C \cdot \exp\left(\sum_{p \le X} \frac{f(p) - 1}{p}\right) + O\left(\frac{1}{\log X}\right)$$

for all sufficiently large X.

Proof. Following Wirsing [26], let

$$F(t) = \sum_{n \le t} f(n).$$

Then

$$\int_1^X \frac{F(t)}{t}\,dt = F(X)\log X + O(1) - \sum_{n \le X} f(n)\log n.$$

On the other hand, by assumption, we have $F(t) = O(t)$, thus

$$\int_1^X \frac{F(t)}{t}\,dt = O(X),$$

and hence

$$F(X)\log X \le O(1) + X + \sum_{n \le X} f(n)\log n.$$

Using $\log n = \sum_{d|n} \Lambda(d)$, where Λ is the von Mangoldt function,[2] we have

$$\begin{aligned}\sum_{n \le X} f(n)\log n &= \sum_{n \le X} f(n) \sum_{d|n} \Lambda(d) = \sum_{d \le X} \Lambda(d) \sum_{m \le X/d} f(dm) \\ &= \sum_{d \le X} \Lambda(d) \sum_{\substack{m \le X/d, \\ (m,d)=1}} f(dm) + \sum_{d \le X} \Lambda(d) \sum_{\substack{m \le X/d, \\ (m,d)>1}} f(dm). \qquad (12)\end{aligned}$$

Now, since $\Lambda(d) = 0$ unless d is a prime power, we have

$$\begin{aligned}\sum_{d \le X} \Lambda(d) \sum_{\substack{m \le X/d, \\ (m,d)>1}} f(dm) &= \sum_{\substack{p^{k+l} \le X \\ k,l \ge 1}} \log(p) \sum_{\substack{m \le X/p^{k+l} \\ (p,m)=1}} f(p^{k+l}m) \qquad (13) \\ &= \sum_{\substack{p^{k+l} \le X \\ k,l \ge 1}} \log(p) f(p^{k+l}) \sum_{\substack{m \le X/p^{k+l} \\ (p,m)=1}} f(m).\end{aligned}$$

By the assumptions on f,

$$f(p^{k+l}) \sum_{\substack{m \le X/p^{k+l} \\ (p,m)=1}} f(m) \le O(k+l) \sum_{m \le X/p^{k+l}} f(m) = O\left((k+l)\frac{X}{p^{k+l}}\right),$$

and thus the second term in (12) is

$$= O\left(\sum_{\substack{p^n \le X \\ n \ge 2}} \log(p) n^2 \frac{X}{p^n}\right) = O(X),$$

[2] That is, $\Lambda(d) = \log p$ if $d = p^k$ and $k \ge 1$; otherwise $\Lambda(d) = 0$.

since

$$\sum_p \sum_{n\geq 2} \log(p) n^2 p^{-n} \leq \sum_p \frac{\log(p)}{p^2} \sum_{m\geq 0} (2+m)^2 2^{-m} < \infty.$$

As for the first term in (12), we have (recall that f is multiplicative and nonnegative)

$$\begin{aligned} \sum_{d\leq X} \Lambda(d) \sum_{\substack{m\leq X/d,\\ (m,d)=1}} f(dm) &= \sum_{d\leq X} \Lambda(d) f(d) \sum_{\substack{m\leq X/d,\\ (m,d)=1}} f(m) \\ &\leq \sum_{m\leq X} f(m) \sum_{d\leq X/m} \Lambda(d) f(d). \end{aligned}$$

Now,

$$\sum_{d\leq X/m} \Lambda(d) f(d) = \sum_{\substack{p^k\leq X/m\\ k\geq 1}} \log(p) f(p^k) \leq \sum_{\substack{p^k\leq X/m\\ k\geq 1}} \log(p) O(k) = O(X/m)$$

since

$$\sum_{p\leq X/m} \log(p) = O(X/m)$$

by the prime number theorem, and

$$\sum_{\substack{p^k\leq X/m\\ k\geq 2}} k\log(p) = O\left((X/m)^{1/2}\log^3(X/m)\right) = O(X/m).$$

Thus,

$$\sum_{m\leq X} f(m) \sum_{d\leq X/m} \Lambda(d) f(d) = O\left(\sum_{m\leq X} f(m)\frac{X}{m}\right).$$

But since f is nonnegative and multiplicative, we have

$$\begin{aligned} \sum_{m\leq X} \frac{f(m)}{m} &\leq \prod_{p\leq X} \left(1 + f(p)/p + f(p^2)/p^2 + \dots\right) \\ &\leq \prod_{p\leq X} \left((1+f(p)/p)\cdot\left(1 + f(p^2)/p^2 + f(p^3)/p^3 + \dots\right)\right), \end{aligned}$$

and since

$$\sum_{p\leq X} \left(f(p^2)/p^2 + f(p^3)/p^3 + \dots\right) \leq \sum_p \sum_{k\geq 2} \frac{O(k)}{p^k} < \infty,$$

we find that

$$\sum_{m\leq X} \frac{f(m)}{m} = O\left(\prod_{p\leq X} (1+f(p)/p)\right).$$

Thus,

$$F(X)\log X = O\left(X + X\cdot\prod_{p\le X}(1+f(p)/p)\right),$$

and hence

$$\frac{F(X)}{X} = O\left(\frac{1}{\log X} + \frac{\prod_{p\le X}(1+f(p)/p)}{\log X}\right).$$

Now, by Merten's theorem, we have

$$\prod_{p\le X}(1-1/p) \sim \frac{e^{-\gamma}}{\log X},$$

and thus

$$\begin{aligned}\frac{F(X)}{X} &= O\left(\frac{1}{\log X} + \prod_{p\le X}\left(1+\frac{f(p)-1}{p}-\frac{f(p)}{p^2}\right)\right)\\ &= O\left(\frac{1}{\log X} + \exp\left(\sum_{p\le X}\frac{f(p)-1}{p}\right)\right). \qquad \square\end{aligned}$$

Proof of Proposition 6. To see that $|S(m,k)/4|$ is a multiplicative function, it is enough to recall the factorization of m into Gaussian primes. Namely, if $p_1^{\alpha_1},\dots,p_J^{\alpha_J}$ are all prime factors of m with $p\equiv 1 \mod 4$,

$$S(m,k) = \sum_{\ell=0}^{3} i^{k\ell}\sum_{j_1=1}^{\alpha_1}\cdots\sum_{j_J=1}^{\alpha_J} e^{ik(\theta_0+(\alpha_1-2j_1)\theta_{p_1}+\cdots+(\alpha_J-2j_J)\theta_{p_J})}.$$

Here θ_0 is a multiple of $\pi/4$ which comes from powers of 2 in m, and the θ_{p_j} can be computed from the Gaussian factorization as described in section 2.2. Also, because $\sum_{\ell=0}^{3} i^{k\ell} = 4$ if $4\,|\,k$ and zero otherwise,

$$\frac{|S(m,k)|}{4} = \left|\sum_{j_1=1}^{\alpha_1}\cdots\sum_{j_J=1}^{\alpha_J} e^{ik((\alpha_1-2j_1)\theta_{p_1}+\ldots+(\alpha_J-2j_J)\theta_{p_J})}\right|,$$

and this sum clearly factors, each factor containing a sum of terms corresponding to one of the prime factors p. Hence

$$f_k(m) = \frac{|S(m,k)|}{4}$$

is a nonnegative multiplicative function, as stated. In addition it satisfies $f_k(m) \le r(m)/4$ for all m. Thus, since

$$\sum_{n\le T} r(n) = |\{x,y\in\mathbb{Z} : x^2+y^2\le T\}| \sim \pi\left(\sqrt{T}\right)^2 = \pi T,$$

we have

$$\sum_{n\le T} f_k(n) = O(T).$$

Moreover, if $p \equiv 3 \mod 4$, then

$$f_k(p^l) = \begin{cases} 1 \text{ if } l \text{ is even,} \\ 0 \text{ if } l \text{ is odd,} \end{cases} \tag{14}$$

and if $p \equiv 1 \mod 4$, then

$$f_k(p^l) = \left| \sum_{j=0}^{l} e^{ik(l-2j)\theta_p} \right|, \tag{15}$$

and thus $f_k(p^l) \le l+1$ for all primes p and $l \ge 1$. The assumptions in Theorem 7 are thus satisfied, and we obtain

$$\frac{1}{X}\sum_{m\le X} |S(m,k)| = \frac{4}{X}\sum_{m\le X} f_k(m) \le C\exp\left(\sum_{p\le X}\frac{f_k(p)-1}{p}\right) + O\left(\frac{1}{\log X}\right).$$

Now, by (14) and (15), we have

$$f_k(p) = \begin{cases} 2|\cos(k\theta_p)| & \text{if } p \equiv 1 \mod 4, \\ 0 & \text{if } p \equiv 3 \mod 4. \end{cases}$$

Hence

$$\sum_{p\le X}\frac{f_k(p)-1}{p} = \sum_{\substack{p\le X\\ p\equiv 1 \mod 4}} \frac{2|\cos(k\theta_p)|}{p} - \sum_{p\le X}\frac{1}{p}.$$

By Corollary 5,

$$\sum_{\substack{p\le X\\ p\equiv 1 \mod 4}} \frac{2|\cos(k\theta_p)|}{p} \le \frac{2}{\pi}\log\log X + 2(1-2/\pi)\log\log k + O(1)$$

if $\log k \le b\sqrt{\log X}$. By Merten's theorem,

$$\sum_{p\le X}\frac{1}{p} = \log\log X + O(1),$$

and thus

$$\sum_{p\le X}\frac{f_k(p)-1}{p} \le (2/\pi - 1)\log\log x + 2(1-2/\pi)\log\log k + O(1). \qquad \square$$

4. Proof of Theorem 1. Here we carry out the steps of the proof as indicated in the introduction. First recall that the collision operator can be written

$$Q(f,f)(v) = 4\int_{\mathbb{R}^2}\left(\int_{-\pi}^{\pi} g_v(w,\theta)\,d\theta\right)dw, \tag{16}$$

where, if we identify $u\in S^1$ with $\theta\in[-\pi,\pi[$,

$$g_v(w,\theta) = q(|w|,\cos(\theta))\left(f(v')f(v'_*)-f(v)f(v_*)\right),$$

and

$$\begin{aligned} v' &= v+w+R_\theta w,\\ v'_* &= v+w-R_\theta w; \end{aligned}$$

as before, $w=(v_*-v)/2$, and R_θ denotes a rotation by an angle θ. Writing the Boltzmann equation for two-dimensional velocities, we have of course already stepped away from the physically realistic case. Disregarding this, a common assumption on q is that

$$q(|w|,\cos(\theta)) = q_1(|w|)q_2(\theta),$$

where $q_1(|w|)\sim|w|^\alpha$ for some $\alpha\in[0,1]$, and where $q_2(\theta)\sim|\theta|^{-\gamma}$ for some $\gamma\in]1,3[$. This corresponds to a molecular interaction by hard inverse power law forces. With the stronger assumption that q_1 is smooth and strictly positive, it is possible to prove that there is a smooth solution $f(v,t)$ to the Boltzmann equation (see [9]), and then this also gives some regularity to $g(w,\theta)$, in spite of the singularity of q_2.

However, much work on the Boltzmann equation has been done with the hypothesis that q is bounded or continuous with respect to θ. With that assumption, the solution $f(v,t)$ keeps exactly the regularity of the initial data.

Because of this, it is appropriate to assume whatever regularity of the solutions that is needed for the computations. With the aim of making the calculations easy, Theorem 1 has been written with unnecessarily strong hypotheses.

To simplify notation a little, let

$$G_v(w) = \int_{-\pi}^{\pi} g_v(w,\theta)\,d\theta,$$

in the continuous case, and for the discrete case (where we assume, of course, that $v\in h\mathbb{Z}^2$)

$$G_v^h(h\zeta) = \frac{1}{r(|\zeta|^2)}\sum_{\substack{\zeta'\in\mathbb{Z}^d\\|\zeta'|=|\zeta|}} g_v(h\zeta,\theta),$$

where θ is the angle between ζ' and ζ. As before, $r(|\zeta|^2)$ denotes the number of integer points on a sphere with radius $|\zeta|$.

Let

$$Z_{h,R} = \{z\in\mathbb{Z}^2 \text{ such that } |z|\le R/h\} \tag{17}$$

for some $R > 0$ (this is the most natural example, but other choices might be more efficient). We want to prove that

$$Q(f,f)(v) - (2h)^2 \sum_{\zeta \in Z_{h,R}} G_v^h(h\zeta) \to 0 \tag{18}$$

when $h \to 0$, and also make as precise a statement as possible about the rate of convergence, and thus give a more explicit version of Theorem 1.

THEOREM 8. *Suppose that $g_v(w,\theta)$ in* (16) *satisfies:*

(1) *$g_v(w,\theta)$ is a C^1-function w.r.t. w,*

(2) *$g_v(w,\theta)$ is a C^2-function w.r.t. θ, and*

(3) $\|g_v(\cdot,\theta)(1+|\cdot|^2)\|_{L^1(dw)} \le C$.

(This holds, e.g., if the function f and the cross section q are C^2.) For given $R > 0$ and $h > 0$, let $Z_{h,R}$ be as in (17)*. Then given $\varepsilon > 0$ there are reals $R > 0$ and $h > 0$ such that*

$$\left| Q(f,f)(v) - (2h)^2 \sum_{\zeta \in Z_{h,R}} G_v^h(h\zeta) \right| \le \varepsilon.$$

Proof. We still consider $Q(f,f)$ as an iterated integral, and write (for $v \in h\mathbb{Z}^2$)

$$\begin{aligned} Q(f,f)(v) - (2h)^2 \sum_{\zeta \in Z_{h,R}} G_v^h(h\zeta) &= \int_{\mathbb{R}^2} G_v(w)\,dw - (2h)^2 \sum_{\zeta \in Z_{h,R}} G_v(h\zeta) \\ &\quad + (2h)^2 \sum_{\zeta \in Z_{h,R}} \left(G_v(h\zeta) - G_v^h(h\zeta) \right). \end{aligned} \tag{19}$$

From the third part of the hypothesis on g (which is implied by a decay of $f(v)$ for large velocities), it follows that for all $R > 0$,

$$\int_{|w| \ge R} G_v(w)\,dw \le \frac{C_1}{R^2}. \tag{20}$$

Continuity of $G_v(w)$ would be enough to conclude that

$$\left| \int_{|w| < R} G_v(w)\,dw - (2h)^2 \sum_{\zeta \in Z_{h,R}} G_v(h\zeta) \right| \to 0$$

when $h \to 0$. The hypothesis on $g_v(w,\theta)$ implies that actually $G_v(\cdot) \in C^1$, and there is a constant C_2 such that the difference is smaller than

$$C_2 R^2 h = C \max_{w,j} |\partial_{w_j} G_v(w)|\, R^2 h. \tag{21}$$

Next we turn to the difference $G_v(h\zeta) - G_v^h(h\zeta)$, i.e.,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} g_v(h\zeta,\theta)\,d\theta - \frac{1}{r(|\zeta|^2)} \sum_{\substack{\zeta' \in \mathbb{Z}^2 \\ |\zeta'| = |\zeta|}} g_v(h\zeta,\theta), \tag{22}$$

(recall that in the second term, θ is the angle between ζ' and ζ). We first write the periodic function $g_v(h\zeta,\theta)$ as a Fourier series,

$$g_v(h\zeta,\theta) = \sum_{k\in\mathbb{Z}} \hat{g}_v(\zeta,k)e^{ik\theta},$$

where

$$\hat{g}_v(\zeta,k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} g_v(h\zeta,\theta)e^{-ik\theta}\,d\theta.$$

The assumptions on g imply the existence of a constant C_3 so that

$$|\hat{g}_v(\zeta,k)| \le \frac{C_3}{1+k^2}. \tag{23}$$

Then (22) becomes

$$\hat{g}_v(\zeta,0) - \frac{1}{r(|\zeta|^2)}\sum_{\substack{\zeta'\in\mathbb{Z}^2\\|\zeta'|=|\zeta|}} \hat{g}_v(\zeta,0) + \frac{1}{r(|\zeta|^2)}\sum_{\substack{\zeta'\in\mathbb{Z}^2\\|\zeta'|=|\zeta|}}\sum_{k\neq 0}\hat{g}_v(\zeta,k)e^{ik\theta},$$

where the first terms cancel out, and only last sum remains. We next split that sum into a part with $|k|\le M$, and a remainder, which can be made small by choosing M large, if g is sufficiently smooth with respect to θ. Using (23),

$$\left|\frac{1}{r(|\zeta|^2)}\sum_{\substack{\zeta'\in\mathbb{Z}^2\\|\zeta'|=|\zeta|}}\sum_{|k|\ge M}\hat{g}_v(\zeta,k)e^{ik\theta}\right| \le 2\frac{C_3}{M}.$$

To find the contribution of this term to (19), we multiply by $(2h)^2$ and sum over $\zeta\in Z_{h,R}$ to find a bound of the form

$$\frac{R^2C_4}{M}. \tag{24}$$

For the remaining part, using (23) again, we find a bound of the form

$$\left|\sum_{0<|k|<M}\frac{C_3}{1+k^2}\frac{1}{r(|\zeta|^2)}\sum_{\substack{\zeta'\in\mathbb{Z}^2\\|\zeta'|=|\zeta|}} e^{ik\theta}\right| \le \max_{0<|k|<M}\left|\frac{S(|\zeta|^2,k)}{r(|\zeta|^2)}\right|\cdot\sum_{0<|k|<M}\frac{C_3}{1+k^2}. \tag{25}$$

Adding the error terms (20), (21), (24) and (25) gives

$$\begin{aligned}&|Q(f,f)(v)-Q^h(f^h,f^h)(v)|\\ &\le \frac{C_1}{R^2}+C_2R^2h+\frac{R^2C_4}{M}+C_3(2h)^2\max_{0<|k|<M}\sum_{\zeta\in Z_{h,R}}\left|\frac{S(|\zeta|^2,k)}{r(|\zeta|^2)}\right|.\end{aligned} \tag{26}$$

In the sum on the right-hand side,

$$\sum_{\zeta\in Z_{h,R}}\left|\frac{S(|\zeta|^2,k)}{r(|\zeta|^2)}\right| = \sum_{n<(R/h)^2}\left|S(n,k)\right|,$$

and this can be estimated by using Proposition 6 with $X = (R/h)^2$. To do this, we must require that

$$R/h > \exp\left(\log(M)^2/b\right) \tag{27}$$

for some positive constant b. Then there is a constant C_5 such that

$$\sum_{n<(R/h)^2} \left|S(n,k)\right| \le C_5 \left(\frac{R}{h}\right)^2 \exp\left(-\left(1-\frac{2}{\pi}\right)\frac{\log\left((R/h)^2\right)}{(\log M)^2}\right).$$

The last term in (26) will always be the dominating one, and at this point, it does not give much to try to optimize the choices of R, M and h. Hence to achieve an error of magnitude ε we

1. Take $R = \sqrt{4C_1/\varepsilon}$,
2. Observe that we must have $h < \varepsilon/(4R^2C_2) = \varepsilon^2/(4C_1C_2)$, and
3. Choose $M = 4R^2C_4/\varepsilon = 64C_1C_4/\varepsilon^2$.

With these choices of R and M, the last term can then be bounded by

$$4C_3C_5\frac{4C_1}{\varepsilon}\exp\left(-\left(1-\frac{2}{\pi}\right)\log\frac{\log(4C_1/(\varepsilon h^2))}{\left(\log(64C_1C_4/\varepsilon^2)\right)^2}\right), \tag{28}$$

which converges to zero when $h \to 0$, and so there is an h so small that also the last term in (26) is smaller than $\varepsilon/4$. We see that in order to achieve an error of magnitude ε, one must take h very small:

$$h = o\left(\exp\left(-2\left(\log\varepsilon\right)^2\varepsilon^{-2/(1-\frac{2}{\pi})}\right)\right)$$

(note that (27) is then satisfied). $\square$

5. Some remarks. From a numerical point of view, the discretization discussed above would be far too costly. It is most common to express the rate of convergence by giving an estimate of the error as a function of h, which in our case is the lattice parameter. The proof of Theorem 8 gives an estimate of h needed to achieve an error of the order ε, and this corresponds to a rate of convergence of the order $(\log(1/h))^p$, where $p < (1-2/\pi)/2$. This can then be used to give an estimation of the computational cost. A discrete velocity model with N velocities would at least correspond to a computational cost of $O(N)$ per time step, because one needs to compute a value for each velocity. When the collision term is computed by the sum (18), the cost is $O(N^2)$ times some logarithmic factor of N (which comes from the summation over the points on the circles). Because $N \sim h^{-2}$, we can conclude that $N >> \exp(\varepsilon^{-c})$ for some positive constant c. It seems clear that this kind of DVM will rarely be useful for numerical computations.

However, disregarding the computational cost, there is also another interesting issue with this kind of discrete velocity model, namely, the possibility of spurious invariants. By a collision invariant, we mean a function $\Psi(v)$ that satisfies

$$\begin{gathered}\forall(j,k,j',k') \text{ such that } \Gamma_{j,k}^{j',k'} > 0,\\ \Psi(v_j)+\Psi(v_k) = \Psi(v_{j'})+\Psi(v_{k'}).\end{gathered} \tag{29}$$

The only invariants should be those corresponding to the conservation of mass, momentum and energy, i.e.,

$$\Psi(v) = 1, \quad \Psi(v) = b\cdot v \ \ (b\in\mathbb{R}^2), \quad \text{and } \Psi(v) = |v|^2.$$

All other functions satisfying (29) are called spurious invariants.

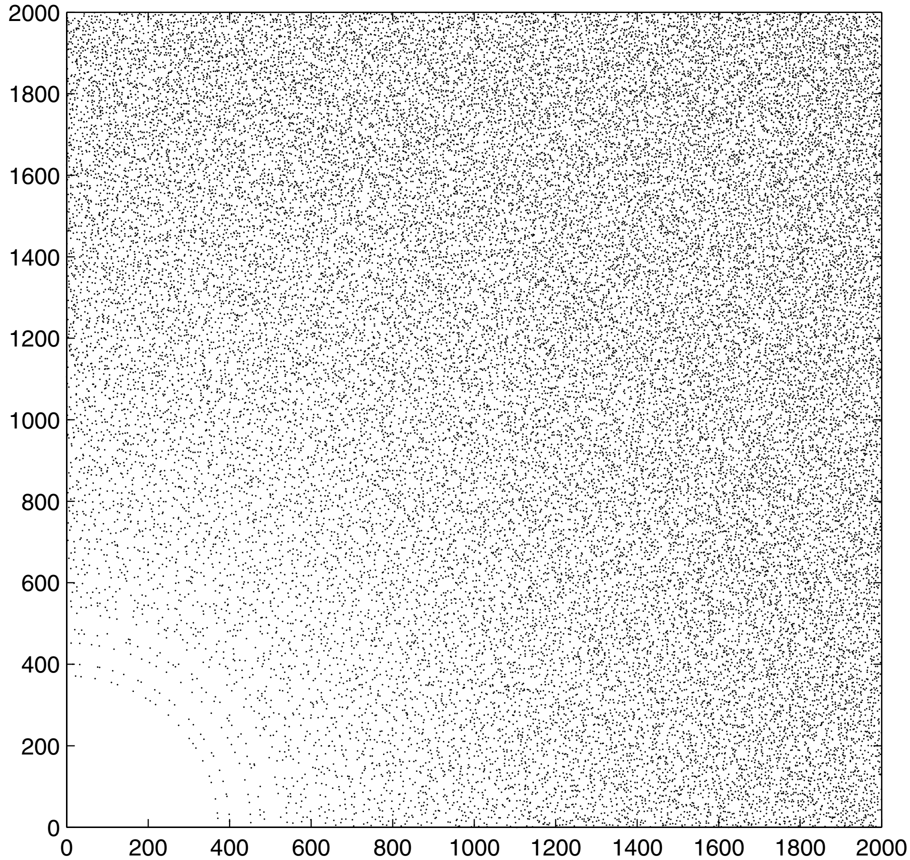


FIG. 1. *Lattice points such that circles through these points contain at least* 72 *lattice points.*

It is interesting to note that the present planar lattice model does not admit any spurious invariants, at least under some very modest requirements on the differential cross section (29). The proof, which can be found in [4], is constructive and basically works in the following way: Starting from a model which is known to possess the correct invariants, one adds one point at a time in such a way that the correct invariants are maintained. The planar lattice models that we study in this paper can thus be obtained by adding points to the Broadwell model, which consists of the velocities $(\pm 1, 0)$, $(0, \pm 1)$, extended with the point $(0, 0)$.

Another question one might ask is whether it is always essential to avoid spurious invariants. When considering whole families of models, one can at least imagine two different definitions of a spurious invariant. Taking into account the scaling of the lattice with respect to the total number of velocity points, N, the definition of a collision invariant would be

$$\begin{aligned}&\forall (j,k,j',k') \text{ such that } \Gamma_{j,k}^{j',k'} > 0,\\ &\Psi(h_N v_j) + \Psi(h_N v_k) = \Psi(h_N v_{j'}) + \Psi(h_N v_{k'}),\end{aligned} \tag{30}$$

where h_N is the lattice parameter corresponding to a given model with N velocities. A "strong invariant" would then be a function Ψ that satisfies (30), for all N sufficiently

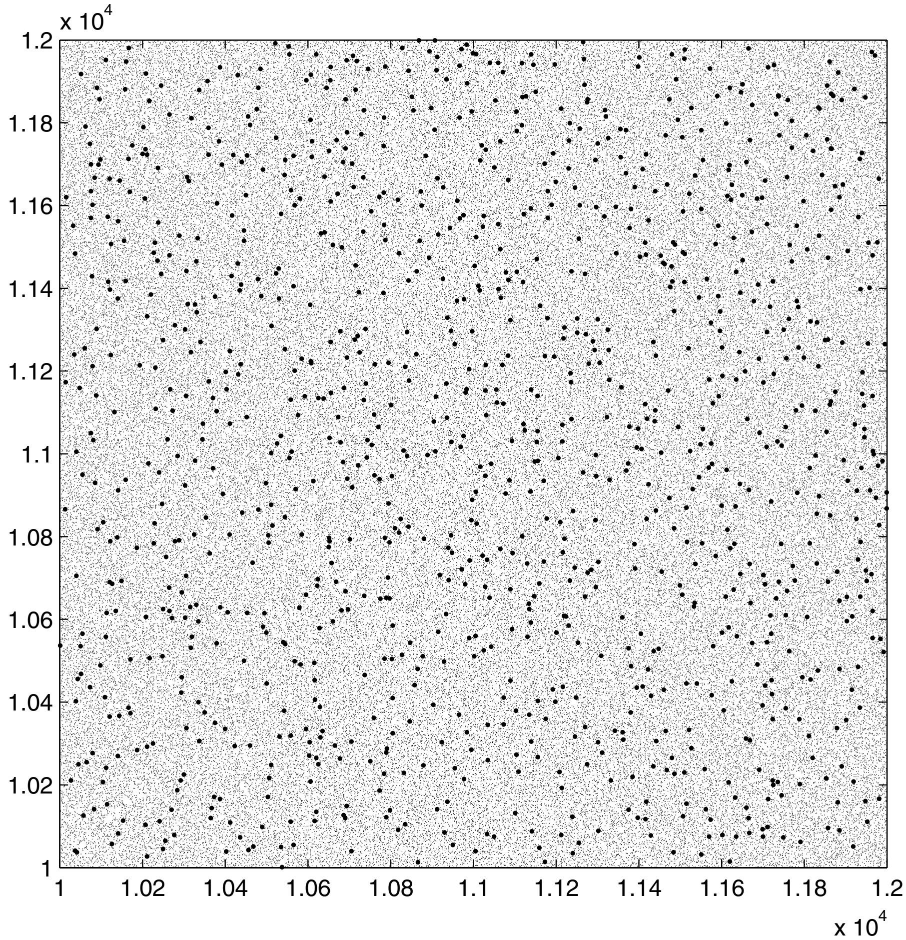


FIG. 2. *Lattice points such that circles through these points contain at least* 72 *lattice points (small dots), or at least* 192 *points (large dots).*

large, or for a large set of $N \in \mathbb{Z}$. However, besides the desired ones, such functions cannot exist in our case, because that would contradict the strong convergence result, namely, Theorem 8.

Although the estimated rate of convergence indicates that the computational cost would be prohibitive for a model with many velocities, it is still possible to carry out simulations. In a companion paper [12], a numerical calculation on a 400×400 grid is reported. This calculation could not have been done with a reasonable computational effort if all terms in (9) had been included. To reduce the number of terms, we chose to only include those ζ for which $r_2(|\zeta|^2)$ is larger than an (arbitrary) threshold. To justify this, one should check that the remaining points are well distributed, and one should verify that there are no spurious invariants in the reduced model.

Proposition 6, which gives an effective estimate on the equidistribution of points, is indirect, and it is interesting to see how the points $\zeta \in \mathbb{Z}^2$ ζ are distributed. We consider $\{\zeta = (\zeta_1, \zeta_2) \in \mathbb{Z}^2 \text{ such that } \zeta_1, \zeta_2 \geq 0, \;\; |\zeta| < 20000\}$. This is an extremely large set of points, which corresponds to a huge number of velocities (the $O(N^2)$ factor would in this case be of the order 10^{17}, which is, of course, absurd).

Among the circles with radii $|\zeta|$ in this set, the largest number of points on one circle is 384. In Figure 1, we show all points $\zeta = (\zeta_1, \zeta_2)$ with $0 < \zeta_i < 2000$ such that the circle passing through ζ has more than 72 points. There are 36163 points in this set. This is a small fraction of the total number of integer points, but they are seemingly well distributed, except near the origin.

Figure 2 shows points in the range $10000 \leq \zeta_i \leq 12000$. Here the small dots denote points on circles having at least 72 points, and the larger dots denote points on circles with at least 192 points (there are 141562 and 1120 points, respectively, in these sets).

As for the question of spurious invariants in this restricted model, this can be checked by an algorithm based on the Bobylev–Cercignani [4] construction. We refer to [12] for a more detailed description of this, but it is interesting to note that if admissibility of the model (i.e., the absence of spurious invariants) is all that one is interested in, then it is in many cases sufficient to use only points lying on circles with just one radius $|\zeta|$. This could be chosen to be a radius giving many points on the corresponding circle, and thus a good approximation of the angular integral. However, the integral over $\mathbb{R}^2$ would be poorly approximated.

Acknowledgment. We would like to thank A. Bobylev, J. Brzezinski, A. Heintz, and Z. Rudnick for useful discussions.

REFERENCES

[1] A. V. Bobylev and N. Bernhoff, *Discrete velocity models and dynamical systems*, in Lecture Notes on the Discretization of the Boltzmann Equation, N. Bellomo and R. Gatignol, eds., World Scientific, River Edge, NJ, 2003, pp. 203-222.

[2] A. V. Bobylev, A. Palczewski, and J. Schneider, *On approximation of the Boltzmann equation by discrete velocity models*, C. R. Acad. Sci. Paris Sér. I Math., 320 (1995), pp. 639–644.

[3] A. Palczewski, J. Schneider, A Bobylev, *A consistency result for a discrete-velocity model of the Boltzmann equation*, SIAM J. Numer. Anal., 34 (1997), pp. 1865–1883.

[4] A. Bobylev and C. Cercignani, *Discrete velocity models without nonphysical invariants*, J. Statist. Phys., 97 (1999), pp. 677–686.

[5] T. Carleman, *Problèmes mathématiques dans la théorie cinétique des gaz*, Almqvist & Wiksell, Uppsala, Sweden, 1957.

[6] J. Cilleruelo, *The distribution of lattice points on circles*, J. Number Theory, 43 (1993), pp. 198–202.

[7] C. Cercignani, R. Illner, M. Pulvirenti, *The Mathematical Theory of Dilute Gases*, Springer-Verlag, New York, 1994.

[8] L. Desvillettes and S. Mischler, *About the splitting algorithm for Boltzmann and B.G.K. equations*, Math. Models Methods Appl. Sci., 6 (1996), pp. 1079–1101.

[9] L. Desvillettes and B. Wennberg, *Regularity of solutions to the spatially homogeneous Boltzmann equation without cutoff*, Comm. Partial Differential Equations, 29 (2003), pp. 133–156.

[10] W. Duke, *Hyperbolic distribution problems and half-integral weight Maass forms*, Invent. Math., 92 (1988), pp. 73–90.

[11] P. Erdős and R. R. Hall, *On the angular distribution of Gaussian integers with fixed norm*, Discrete Math., 200 (1999), pp. 87–94. Paul Erdős memorial collection.

[12] L. Fainsilber, P. Kurlberg, and B. Wennberg, *The Boltzmann equation on a two dimensional lattice, theoretical and numerical results*, Transport. Theory Statist. Phys., to appear.

[13] E. Grosswald, *Representations of Integers as Sums of Squares*, Springer-Verlag, New York, 1985.

[14] E. P. Golubeva and O. M. Fomenko, *Asymptotic distribution of lattice points on the three-dimensional sphere*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI), 160 (1987), pp. 54–71, 297.

[15] H. Halberstam and H.-E. Richert, *On a result of R. R. Hall*, J. Number Theory, 11 (1979), pp. 76–89.

[16] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 3rd ed., Clarendon Press, Oxford, 1954.

[17] V. Panferov and A. Heintz, *A new consistent discrete-velocity model for the Boltzmann equation*, Math. Methods Appl. Sci., 25 (2002), pp. 571–593.

[18] H. Iwaniec, *Fourier coefficients of modular forms of half-integral weight*, Invent. Math., 87 (1987), pp. 385–401.

[19] I. Kátai and I. Környei, *On the distribution of lattice points on circles*, Ann. Univ. Sci. Budapest. Eötvös Sect. Math., 19 (1977), pp. 87–91.

[20] I. Kubilyus, *The distribution of Gaussian primes in sectors and contours*, Leningrad. Gos. Univ. Uč. Zap. Ser. Mat. Nauk, 137 (1950), pp. 40–52.

[21] A. Palczewski and J. Schneider, *Existence, stability, and convergence of solutions of discrete velocity models to the Boltzmann equation*, J. Statist. Phys., 91 (1998), pp. 307–326.

[22] C. Pommerenke, *Über die Gleichverteilung von Gitterpunkten auf m-dimensionalen Ellipsoiden*, Acta Arith., 5 (1959), pp. 227–257.

[23] F. Rogier and J. Schneider, *A direct method for solving the Boltzmann equation*, Transport Theory Statist. Phys., 23 (1994), pp. 313–338.

[24] V. V. Vedenyapin, *Velocity inductive construction for mixtures*, Transport Theory Statist. Phys., 28 (1999), pp. 727–742.

[25] V. V. Vedenyapin, S. A. Amosov, and L. Toscano, *Invariants for Hamiltonians and kinetic equations*, Russian Math. Surveys, 54 (1999), pp. 1056–1057.

[26] E. Wirsing, *Das asymptotische Verhalten von Summen über multiplikative Funktionen*, Math. Ann., 143 (1961), pp. 75–102.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1923–1946

MASS CONCENTRATION PHENOMENON FOR THE QUINTIC NONLINEAR SCHRÖDINGER EQUATION IN ONE DIMENSION*

NIKOLAOS TZIRAKIS†

Abstract. We consider the L^2-critical quintic focusing nonlinear Schrödinger equation (NLS) on $\mathbf{R}$. It is well known that H^1 solutions of the aforementioned equation blow up in finite time. In higher dimensions, for H^1 spherically symmetric blow-up solutions of the L^2-critical focusing NLS, there is a minimal amount of concentration of the L^2-norm (the mass of the ground state) at the origin. In this paper we prove the existence of a similar phenomenon for the one-dimensional case and rougher initial data, ($u_0 \in H^s,\ s < 1$), without any additional assumption.

Key words. nonlinear Schrödinger equation, blow-up, mass concentration

AMS subject classification. 35Q55

DOI. 10.1137/050626284

1. Introduction.

This paper continues the investigation of the quintic nonlinear Schrödinger equation (NLS) in one dimension that we started in [27]:

$$\begin{aligned} &iu_t + u_{xx} \pm |u|^4 u = 0, \\ &u(x,0) = u_0(x) \in H^s(\mathbf{R}), t \in \mathbf{R}. \end{aligned} \tag{1}$$

The (+) sign in front of the nonlinearity corresponds to the focusing NLS while the (−) sign corresponds to the defocusing NLS. The Cauchy problem for (1) is known to be locally well-posed in $H^s(\mathbf{R})$ for $s > 0$ (see [4]). A local result also exists for $s = 0$, but the time of existence depends on the profile of the data as well as the norm. NLS is an infinite-dimensional Hamiltonian system with energy space H^1. It also has a scaling property. Thus $u(x,t)$ is a solution of (1) with initial data u_0 if and only if

$$u^\lambda(x,t) = \frac{1}{\lambda^{1/2}} u\left(\frac{x}{\lambda}, \frac{t}{\lambda^2}\right)$$

is a solution to the same equation with initial data $u_0(\frac{x}{\lambda})$. In [27] we extend the local existence theorem for the defocusing NLS for all times. We do so by iterating the local result in the appropriate norms. To iterate the local result by standard limiting arguments we just need an a priori bound for our solutions in H^s. This bound comes from the next theorem that we proved in [27].

THEOREM 1. *Let u be a global H^1 solution to* (1) *with the* (−) *sign. Then for any $T > 0$ and $s > 4/9$ we have that*

$$\sup_{0 \le t \le T} \|u(t)\|_{H^s} \lesssim C_{(\|u_0\|_{H^s}, T)},$$

where the right-hand side does not depend on the H^1-norm of u.

*Received by the editors March 8, 2005; accepted for publication (in revised form) August 23, 2005; published electronically March 3, 2006. This material is based upon work supported by the National Science Foundation under agreement DMS-0111298. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not reflect the views of the National Science Foundation.

http://www.siam.org/journals/sima/37-6/62628.html

†School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540 (tzirakis@math.toronto.edu).

Remark 1. Note that in the focusing case (where in front of the nonlinearity we have the (+) instead of the (−) sign) we can also proved global well-posedness for $4/9 < s \leq 1/2$, but with the crucial assumption that $\|u_0\|_{L^2} < \|Q\|_{L^2}$, where Q is the unique positive solution (up to translations) of

$$Q_{xx} - Q + |Q|^4 Q = 0.$$

In [29], Q was solved explicitly as $Q(x) = \frac{3^{\frac{1}{4}}}{\sqrt{\cosh(2x)}}$ and then $\|Q\|_{L^2}^2 = \frac{\sqrt{3}\pi}{2}$. In the same paper the author also proved a result that we will use below, namely that $C = \|Q\|_{L^2}^{-4}$ is the best constant in the Gagliardo–Nirenberg inequality

$$\frac{1}{6}\|u\|_{L^6}^6 \leq \frac{C}{2}\|\nabla u\|_{L^2}^2 \|u\|_{L^2}^4.$$

We used the "I-method" that was recently introduced by Colliander et al. [5, 7, 8, 9]. This method allows us to define a modification of the energy functional that is "almost conserved"; that is, its time derivative decays with respect to a very large parameter. Since an implementation of this method also gives the main result of this paper, the details of the method are delayed until the next section. As we mentioned above for the focusing case, the solutions blow up in H^1, in finite time. An elementary proof of the existence of blow-up solutions has been known since the 1960s, but is based on energy constraints and is not constructive (see [26]). In particular, no qualitative information of any type is obtained for the blow-up dynamics. A lower estimate for the blow-up solutions in H^1 is given by Theorem 2 using the scaling and the local existence theorem (see [3]).

THEOREM 2. *Let $[0, T^\star)$ be the maximal interval of existence of the following Cauchy problem:*

$$\begin{aligned} &iu_t + u_{xx} + |u|^4 u = 0, \\ &u(x,0) = u_0(x) \in H^1(\mathbf{R}), t \in \mathbf{R}. \end{aligned} \tag{2}$$

If $u_0 \in H^1$ is such that $T^\star < \infty$, then there exists a C such that

$$\|u_x\|_{L^2} \geq \frac{C}{(T^\star - t)^{\frac{1}{2}}}$$

for $0 \leq t < T^\star$. This bound is often called the scaling bound. It is also fairly easy to show that $\|u(t)\|_{L^p}$ blows up for $p > 2$. In particular we have

$$\|u\|_{L^p} \geq \frac{C}{(T^\star - t)^{\frac{1}{4} - \frac{1}{2p}}}$$

for $p > 2$.

Because it is related to the scaling symmetry of the problem, the above lower bound has long been conjectured to be optimal. But in 1988, Landman, et al. [14] suggested that the correct and stable blow-up speed is a slight correction to the scaling bound:

$$\|u_x\|_{L^2} \sim \sqrt{\frac{\log|\log|T^\star - t||}{T^\star - t}}.$$

In this framework Perelman [24] has constructed a family of blow-up solutions for which

$$\left(\frac{\log|\log|T^\star - t||}{T^\star - t}\right)^{-\frac{1}{4}} \|u(t)\|_{L^\infty} \to c > 0$$

as $t \to T^\star$, which is very close to but different from the scaling bound. Moreover, for initial data in some special class, Merle and Raphael [18, 19] recently showed that for t close to $T^\star$, there is a universal constant $C^\star$ such that

$$\|u_x\|_{L^2} \le C^\star \sqrt{\frac{\log|\log|T^\star - t||}{T^\star - t}},$$

as suggested by the numerics in [14]. Finally it is worth noting that an easy application of the pseudoconformal transformation yields interesting information on the blow-up solutions. In particular we can show that some solutions blow up twice as fast as the scaling bound. For details see [2] and [28]. The above results show in particular that at least two different blow-up estimates are actually achieved.

Another property of the blow-up solutions in the critical case is the phenomenon of mass concentration [3, 26]. For H^1 solutions, there is a concentration of a finite amount of mass in a neighborhood of the focus of width slightly larger than $(T^\star - t)^{1/2}$. For radial initial data in dimension $d \ge 2$ there is a precise lower bound on the amount of concentrated mass in terms of the mass of the ground state Q (see [20]). More precisely we have the following.

Let $d \ge 2$ and let $\gamma : (0, \infty) \to (0, \infty)$ be any function such that $\gamma(s) \to \infty$ and $s^{1/2}\gamma(s) \to 0$ as $s \downarrow 0$. Finally, let $u_0 \in H^1(\mathbf{R^d})$ be radial symmetric. Then if $u(x,t)$ is the maximal solution of the equivalent of (2) in higher dimensions and $T^\star < \infty$, we have

$$\liminf_{t \uparrow T^\star} \|u(t)\|_{L^2_{\{|x| < |T^\star - t|^{1/2}\gamma(T^\star - t)\}}} \ge \|Q\|_{L^2},$$

where Q is the ground state solution of the elliptic equation $Q_{xx} - Q + |Q|^{\frac{4}{d}} Q = 0$.

In the nonradial case and in dimension $d = 1$ this was generalized by Nawa [22] using concentration compactness techniques [15, 16]. In addition to the scaling properties of the NLS, the main ingredients in the proof that H^1 blow-up solutions concentrate at least the mass of the ground state are

(i) the conservation of mass

$$\|u(t)\|_{L^2} = \|u_0\|_{L^2}$$

and the energy

$$E(u)(t) = E(u_0),$$

where

$$E(u) = \frac{1}{2}\int |u_x(t)|^2 dx - \frac{1}{6}\int |u(t)|^6 dx,$$

(ii) a precise Galiardo–Nirenberg inequality which implies that nonzero H^1 functions of nonpositive energy have at least ground state mass.

The purpose of this paper is to investigate the mass concentration phenomenon in H^s, for $s<1$, where the conservation of energy cannot be used. Using the *I-method* we show that solutions of (2) with a finite maximal (forward) existence interval are expected to concentrate at least the L^2-mass of the ground state in H^s for $s<1$. More precisely we have the following theorem.

THEOREM 3. *Suppose $H^s \ni u_0 \longmapsto u(t)$ with $s>0$ solves (2) on the maximal interval of existence $[0,T^\star)$ with $T^\star<\infty$. Then for any $1>s>\frac{10}{11}$ there exists a positive function $\gamma(x)\uparrow\infty$ arbitrarily slowly as $x\downarrow 0$ and a real function $z(t)$ such that*

$$\limsup_{t\uparrow T^\star} \|u(t)\|_{L^2_{\{|x-z(t)|<(T^\star-t)^{\frac{s}{2}}\gamma(T^\star-t)\}}} \geq \|Q\|_{L^2}.$$

Remark 2. In a recent preprint, Colliander, et al. [11] considered the two-dimensional (2D) focusing critical NLS and proved a similar theorem with the additional assumption of radial symmetry. The radial symmetry assumption is needed in order to pass from weak to strong convergence since the general embedding $H^1(\mathbf{R^d})\hookrightarrow L^2(\mathbf{R^d})$ is not compact. But as the four authors noted in [11], one can utilize the concentration compactness method of Lions [15, 16] and prove the analogous theorem in two dimensions. The one-dimensional case that we are dealing with has some similar features but also significant differences. First, in the one dimensional case, the radial assumption does not play a role. More precisely, in one dimension, radial symmetry is not enough for a bounded sequence in H^1 to have a strongly convergent subsequence in L^p for $2<p<\infty$, although the latter is true if we further assume that the sequence in question is a nonincreasing function of $|x|$ for every $n\geq 0$. (For the above discussion the reader can also consult [25]). So we have to prove Theorem 3 by implementing techniques different from those used in [11]. Second, the nonlinearity has a fifth power, and thus the correction terms in the "modified energy" have larger growth. We take advantage of the fact that at each step we work on $[0,\delta]$ and prove a stronger proposition about the decay of the "modified energy" and thus somehow balance the additional correction terms with the greater decay that we prove. Finally, the crucial Lemma 3 that we use in one dimension is true only if the frequencies of the two solutions are separated. In higher dimensions the analogous lemma holds in general [1], although we avoid this difficulty in one dimension by analyzing further the correction terms of the "modified energy"; see Proposition 5.

Remark 3. As we mentioned before, to prove the theorem we use a combination of the concentration compactness and the I-method. Since the energy is infinite for initial data in H^s we define a "modified energy," $E(Iu)$, which is finite, where $I: H^s\to H^1$ is a multiplier operator defined below. The crucial step is to prove that the modified total energy grows more slowly than the modified kinetic energy

$$\frac{1}{2}\int |Iu_x(t)|^2dx.$$

These two steps are shown in Propositions 5 and 3, respectively. Note that Proposition 5 relies on the local theory that we shall establish in Proposition 1.

Remark 4. Let $p(s)$ be a number that depends on s and for the range of s in Theorem 3, $(10/11<s<1)$, it is $p(s)<2$. The statement that the modified total energy grows more slowly than the modified kinetic energy is reflected exactly on $p(s)<2$ and is proven in Proposition 3 below. Note also that our concentration width $(T^\star - t)^{\frac{s}{2}}$ is larger than $(T^\star - t)^{\frac{1}{2}}$ with which ground state mass concentration is conjectured to occur.

Two quick by-products of the above theorem follow. The first is the conjecture that tiny L^2 mass concentration cannot occur when $u_0 \in L^2$, a question that was asked in [21]. See also the relevant result of Bourgain [1]. The second is the following lemma which, as we mention on the first page, is basically a result of the work in [27].

LEMMA 1. *If* $u_0 \in H^s$, $s > \frac{10}{11}$, *and* $\|u_0\|_{L^2} < \|Q\|_{L^2}$, *then the initial value problem* (2) *is globally well-posed.*

We end this section by introducing some useful notation. In what follows we use $A \lesssim B$ to denote an estimate of the form $A \leq CB$ for some constant C. If there exist constants C and D such that $DB \leq A \leq CB$, we say that $A \sim B$, and $A \ll B$ to denote an estimate of the form $A \leq cB$ for small constant $c > 0$. In addition, $\langle a \rangle := 1 + |a|$ and $a\pm := a \pm \epsilon$.

2. Linear and bilinear estimates. Before we state the linear and bilinear estimates that we will use throughout this paper, we recall some basic facts about the $X^{s,b}$ spaces. For an equation of the form

$$iu_t - \phi(-i\nabla)u = 0, \tag{3}$$

where ϕ is a measurable function, let $X^{s,b}$ be the completion of $\mathbb{S}(\mathbb{R}^{d+1})$ with respect to

$$\|u\|_{X^{s,b}} = \|\langle\xi\rangle^s \langle\tau + \phi(\xi)\rangle^b \hat{u}(\xi,\tau)\|_{L^2_\xi L^2_\tau}.$$

From the above definition it is clear that the dual space of $X^{s,b}_{\tau=\phi(\xi)}$ is $X^{-s,-b}_{\tau=-\phi(-\xi)}$. Furthermore for a given interval I, we define

$$\|f\|_{X^{s,b}(I)} = \inf_{\tilde{f}_{|I}=f} \|\tilde{f}\|_{X^{s,b}}.$$

In our case, the interval of existence of the local solutions will be $[0,\delta]$ and we write $X^{s,b}_\delta = X^{s,b}_{[0,\delta]}$. Since conjugate solutions will not play any role in our arguments from now on, we omit any reference to the difference between u and $\bar{u}$. We know that if u is a solution of (3) with $u(0) = f$ and if ψ is a cut-off function in C_0^∞ with support of $\psi \subset (-2,2)$, $\psi = 1$ on $[0,1]$, $\psi(-t) = \psi(t)$, $\psi(t) \geq 0$, $\psi_\delta(t) = \psi(\frac{t}{\delta})$, then if $0 < \delta \leq 1$, we have that for $b \geq 0$

$$\|\psi_1 u\|_{X^{s,b}} \leq C\|f\|_{H^s}. \tag{4}$$

In addition, if ν is a solution of

$$i\nu_t - \phi(-i\nabla)\nu = F$$

with $\nu(0) = 0$, then for $b' + 1 \geq b \geq 0 \geq b' > -\frac{1}{2}$,

$$\|\psi_\delta \nu\|_{X^{s,b}} \leq C\delta^{1+b'-b}\|F\|_{X^{s,b'}}. \tag{5}$$

The proofs of (4) and (5) can be found in [12]. The Strichartz estimates for the Schrödinger equation on $\mathbb{R}^d$ state that for $q, r \geq 2$ such that $(d,q) \neq (2,2)$ and $0 \leq \frac{2}{q} = d(\frac{1}{2} - \frac{1}{r}) < 1$, we have that

$$\|e^{it\Delta}u_0\|_{L^q_t L^r_x} \lesssim \|u_0\|_{L^2(\mathbb{R}^d)}. \tag{6}$$

In particular, in one dimension we have

$$\|e^{it\partial_x^2}u_0\|_{L_t^6L_x^6} \lesssim \|u_0\|_{L^2(\mathbb{R})}$$

and

$$\|e^{it\partial_x^2}u_0\|_{L_t^\infty L_x^2} \lesssim \|u_0\|_{L^2(\mathbb{R})},$$

which by a standard argument gives

$$\|u\|_{L_t^6L_x^6} \lesssim \|u\|_{X_\delta^{0,1/2+}} \tag{7}$$

and

$$\|u\|_{L_t^\infty L_x^2} \lesssim \|u\|_{X_\delta^{0,1/2+}}. \tag{8}$$

By the Sobolev embedding theorem in one dimension, (8) implies that

$$\|u\|_{L_t^\infty L_x^\infty} \lesssim \|u\|_{X_\delta^{1/2+,1/2+}}. \tag{9}$$

Also by interpolation between (7) and the trivial estimate

$$\|u\|_{L_t^2L_x^2} = \|u\|_{X_\delta^{0,0}}, \tag{10}$$

we get

$$\|u\|_{L_t^pL_x^p} \lesssim \|u\|_{X_\delta^{0,(1/2+)\cdot(3/2-3/p)}} \tag{11}$$

for any $2 \leq p \leq 6$.

The dual version of (6) gives

$$\|u\|_{X_\delta^{0,-1/2-}} \lesssim \|u\|_{L_t^{q'}L_x^{r'}}, \tag{12}$$

where r' and q' are the conjugate exponents of r and q, respectively. Interpolation with the trivial estimate

$$\|u\|_{X_\delta^{0,0}} = \|u\|_{L_t^2L_x^2} \tag{13}$$

gives that

$$\|u\|_{X_\delta^{0,-1/2+}} \lesssim \|u\|_{L_t^{q'+}L_x^{r'+}} \tag{14}$$

and also that

$$\|u\|_{X_\delta^{1,-1/2+}} \lesssim \|u\|_{L_t^{q'+}W_x^{1,r'+}} \tag{15}$$

for any $\frac{2}{q}+\frac{1}{r}=\frac{1}{2}$.

As we state in the introduction, we prove that the modified energy grows more slowly than the modified kinetic energy in Proposition 3. We can take advantage of the fact that we work on a small interval $[0,\delta]$ and improve the decay of the "modified energy." To do so we state the following lemma, which we can find in [12] and [23].

LEMMA 2. *If $1/2 > b > b' \geq 0$ and $s \in \mathbf{R}$, then the following embedding is true:*

$$\|f\|_{X_\delta^{s,b'}} \lesssim \delta^{b-b'} \|f\|_{X_\delta^{s,b}}.$$

The second lemma that we state in this section is an improved bilinear Strichartz-type estimate. It is due to Bourgain [1]. As we mentioned before, a general analogue of Lemma 3 holds for $d \geq 2$; see, for example, [10].

LEMMA 3. *Let u and v be any two Schwartz functions whose support of Fourier transform is in $|\xi| \sim M$ and $|\xi| \ll M$, respectively, and let $M \gg 1$. Then*

$$\|(D_x^{\frac{1}{2}} u)v\|_{L_t^2 L_x^2} = \|(D_x^{\frac{1}{2}} \bar{u})v\|_{L_t^2 L_x^2} \lesssim \|u\|_{X^{0,1/2+}} \|v\|_{X^{0,1/2+}}.$$

3. The I-method and the proof of Theorem 3. As we mentioned above, the basic step toward Theorem 3 is the fact that the "modified total energy" decays more slowly than the "modified kinetic energy." To prove the last statement we iterate the local solutions for the new modified system

$$\begin{aligned} &iIu_t + Iu_{xx} + I(|u|^4 u) = 0, \\ &Iu(x,0) = Iu_0(x) \in H^1(\mathbf{R}), t \in \mathbf{R}. \end{aligned} \tag{16}$$

Thus let us define the I-operator. We introduce as in [5, 8] a radial C^∞, monotone multiplier, taking values in [0,1], where

$$m(\xi) := \begin{cases} 1 & \text{if } |\xi| < N, \\ \left(\frac{|\xi|}{N}\right)^{s-1} & \text{if } |\xi| > 2N, \end{cases}$$

and we define $I : H^s \to H^1$ by $\widehat{Iu}(\xi) = m(\xi)\hat{u}(\xi)$. The operator I is smoothing of order $1-s$, and we have that

$$\|u\|_{X_\delta^{s_0,b_0}} \lesssim \|Iu\|_{X_\delta^{s_0+1-s,b_0}} \lesssim N^{1-s} \|u\|_{X_\delta^{s_0,b_0}} \tag{17}$$

for any $s_0, b_0 \in \mathbf{R}$.

Remark 5. It is shown in [6] that if

$$\|uv\|_{X^{s,b-1}} \lesssim \|u\|_{X^{s,b}} \|v\|_{X^{s,b}},$$

then

$$\|I(uv)\|_{X^{1,b-1}} \lesssim \|Iu\|_{X^{1,b}} \|Iv\|_{X^{1,b}},$$

where the constants in the above inequality are independent of N. From now on we use this fact and refer to it as the "interpolation lemma." For details see [8].

PROPOSITION 1. *Let $s > 10/11$ and consider the equation*

$$iIu_t + (Iu)_{xx} + I(|u|^4 u) = 0 \tag{18}$$

with initial data $Iu(x,0) = Iu_0$. Then there exists a

$$\delta \sim (\|Iu_0\|_{H^1})^{-4-\epsilon}$$

such that for all times in $[0,\delta]$, the above problem is locally well-posed and

$$\|Iu\|_{X_\delta^{1,1/2+}} \lesssim \|Iu_0\|_{H^1}.$$

Proof. By Duhamel's formula (18) is equivalent to

$$Iu(t) = \psi_1(t)e^{it\partial_x^2}(Iu_0) + i\psi_\delta(t)\int_0^t e^{i(t-s)\partial_x^2} I(|u|^4u)(s)ds.$$

By (4) and (5) and the fact that $\delta \leq 1$ we have

$$\|Iu\|_{X_\delta^{1,1/2+}} \lesssim \|Iu_0\|_{H^1} + \|I(|u|^4u)\|_{X_\delta^{1,-1/2+}}.$$

Now recall the dual Strichartz estimate (15)

$$\|u\|_{X_\delta^{1,-1/2+}} \lesssim \|u\|_{L_t^{q'+}W_x^{1,r'+}},$$

where $\frac{2}{q} = \frac{1}{2} - \frac{1}{r}$. Thus for $r' = 2-$ we have

$$\|I(|u|^4u)\|_{X_\delta^{1,-1/2+}} \lesssim \|I(|u|^4u)\|_{L_t^{1+\epsilon}H_x^1} \lesssim \delta^{1-\epsilon}\|I(|u|^4u)\|_{L_t^\infty H_x^1}.$$

Since for $s > 1/2$, H^s is a Banach algebra, we have that

$$\||u|^4u\|_{H_x^s} \lesssim \|u\|_{H_x^s}^5,$$

which by the interpolation lemma quickly translates to

$$\|I(|u|^4u)\|_{H_x^1} \lesssim \|Iu\|_{H_x^1}^5.$$

But then

$$\|Iu\|_{X_\delta^{1,1/2+}} \lesssim \|Iu_0\|_{H^1} + \delta^{1-\epsilon}\|Iu\|_{L_t^\infty H_x^1}^5 \lesssim \|Iu_0\|_{H^1} + \delta^{1-\epsilon}\|Iu\|_{X_\delta{}^{1,1/2+}}^5,$$

and by standard iteration arguments (see [13]), we have that the system is locally well-posed for

$$\delta^{1-\epsilon}\|Iu_0\|_{H^1}^4 < \frac{1}{2}. \qquad \square$$

We also need an analogue of Theorem 2 for the I-system (16). Let ∇^s denote the operator which on the Fourier side is given by $\widehat{\nabla^s u}(\xi) = |\xi|^s\hat{u}(\xi)$. It then follows by the definition of the Japanese bracket that on the Fourier side the $\langle\nabla\rangle$u is given by $(1+|\xi|)\hat{u}(\xi)$.

Proposition 2. *If $H^s \ni u_0 \longmapsto u(t)$ with $s > 10/11$ solves* (2) *for all t close enough to $T^\star$ in the maximal finite interval of existence* $[0, T^\star)$, *then*

$$\|I\langle\nabla\rangle u(t)\|_{L^2} \geq C(T^\star - t)^{-\frac{s}{2}}.$$

Proof. Since we know that

$$\|I\langle\nabla\rangle u(t)\|_{L^2} \geq \|u(t)\|_{H^s},$$

it suffices to show that

$$\|\nabla^s u(t)\|_{L^2} \geq C(T^\star - t)^{-\frac{s}{2}}.$$

We assume that $\|\nabla^s u(t)\|_{L^2} > 1$ since otherwise we can change variables to put the time origin near $T^\star$. Now fix $t \in [0, T^\star)$ and consider

$$v^t(\tau, x) = \lambda^{-1/2} u\left(t + \frac{\tau}{\lambda^2}, \frac{x}{\lambda}\right),$$

where $\lambda = \|\nabla^s u(t)\|_{L^2}^{\frac{1}{s}}$. By scaling, invariance $v^t(\tau, x)$ is a solution to (2). Moreover, an easy calculation shows that

$$\|v^t(0, x)\|_{L^2} = \|u_0\|_{L^2}$$

and that

$$\|\nabla^s v^t(0, x)\|_{L^2} = \lambda^{-s}\|u(t, x)\|_{\dot{H}^s} = 1.$$

Thus $\|v^t(t, x)\|_{H^s} < C$, and by the local theory, this means that there exists a $\tau_0 > 0$, independent of t, such that $v^t(t, x)$ is defined on $[0, \tau_0]$ and therefore

$$t + \frac{\tau_0}{\|\nabla^s u(t)\|_{L^2}^{\frac{2}{s}}} \leq T^\star \Longrightarrow \|\nabla^s u(t)\|_{L^2} \geq C(T^\star - t)^{-\frac{s}{2}}. \qquad \square$$

The last step for the proof of Theorem 3 is the following proposition, which for the moment we assume and prove later.

PROPOSITION 3. *For $s > \frac{10}{11}$ there exists $p(s) < 2$ such that the following hold true:*

If $H^s \ni u_0 \longmapsto u(t)$ solves (2) on $[0, T^\star)$, then for all $T < T^\star$ there exists $N = N(T)$ such that

$$|E[I_{N(T)} u(T)]| \leq C_0 \Lambda(T)^{p(s)}$$

with $C_0 = C_0(s, T^\star, \|u_0\|_{H^s})$, and $\Lambda(T)$ is given in terms of $N(T)$ by $N(T) = C(\Lambda(T))^{\frac{p(s)}{2(1-s)}}$.

We prove Theorem 3 by using the concentration compactness method that was developed by Lions in [15, 16]. We will need a series of lemmas. The proofs of the first two lemmas are easy and can be found on pages 21 and 24, respectively, of [3].

LEMMA 4. *Let $u \in L^2$ and let the concentration function be defined by*

$$\rho(u, t) = \sup_{y \in \mathbf{R}} \int_{\{|x-y|<t\}} |u(x)|^2 dx$$

for $t > 0$. Then ρ is a nondecreasing function of t, and there exists $y(u, t) \in \mathbf{R}$ such that

$$\rho(u, t) = \int_{\{|x-y(u,t)|<t\}} |u(x)|^2 dx.$$

Moreover, if $u \in L^r(\mathbb{R})$ for some $r > 2$, then for all $s, t > 0$ and $C = C(r)$ we have

$$|\rho(u, t) - \rho(u, s)| \leq C\|u\|_{L^r}^2 |t - s|^{\frac{r-2}{r}}.$$

LEMMA 5. *There exists a constant K such that for all $u \in H^1$, all $t > 0$, and ρ defined above we have*

$$\int |u|^6 \leq K\rho(u, t)^2 \left(\int |\nabla u|^2 + t^{-2} \int |u|^2\right).$$

LEMMA 6. *Let* $(u_n)_{n\geq 0} \subset H^1$ *be such that*

$$\|u_n\|_{L^2} \leq a < \infty,$$

$$\sup_{n\geq 0} \|\nabla u_n\|_{L^2} < \infty,$$

and let $\rho(u_n, t)$ *be defined as before. Set*

$$\mu = \lim_{t\to\infty} \liminf_{n\to\infty} \rho(u_n, t).$$

Then there exist a subsequence $(u_{n_k})_{n_k\geq 0}$*, a nondecreasing function* $\gamma(t)$*, and a sequence* $t_k \to \infty$ *with the following properties:*

(i) $\rho(u_{n_k}, .) \to \gamma(.) \in [0, a]$ *as* $k \to \infty$ *uniformly on bounded sets of* $[0, \infty)$.

(ii) $\mu = \lim_{t\to\infty} \gamma(t) = \lim_{k\to\infty} \rho(u_{n_k}, t_k) = \lim_{k\to\infty} \rho(u_{n_k}, t_k/2)$.

Proof. Since

$$\mu = \lim_{t\to\infty} \liminf_{n\to\infty} \rho(u_n, t),$$

there exists a $t_k \to \infty$ such that

$$\mu = \lim_{k\to\infty} \rho(u_{n_k}, t_k) \tag{19}$$

and thus one part of (ii) is evident. To prove the first part note that

$$\rho(u_n, t) \leq \|u_n\|_{L^2} \leq a < \infty.$$

In addition since $H^1(\mathbb{R}) \hookrightarrow L^r(\mathbb{R})$ for some r, by the last property of the previous lemma $\rho(u_n, \cdot)$ is Hölder continuous. Therefore (i) follows from Ascoli's theorem (after renaming the sequence n_k). Notice that property (19) is still true after passing to a subsequence. For the rest of (ii) by (19) and the fact that $\rho(u_n, \cdot)$ is nondecreasing we deduce that

$$\limsup_{k\to\infty} \rho\left(u_{n_k}, \frac{t_k}{2}\right) \leq \limsup_{k\to\infty} \rho(u_{n_k}, t_k) = \mu. \tag{20}$$

Next for every $t > 0$ we have

$$\liminf_{k\to\infty} \rho\left(u_{n_k}, t\right) \geq \liminf_{n\to\infty} \rho(u_n, t),$$

Now by letting $t \to \infty$ and using part (i) of the lemma and the definition of μ we get that

$$\lim_{t\to\infty} \gamma(t) \geq \mu. \tag{21}$$

Finally, given $t > 0$ we have $\frac{t_k}{2} > t$ for k large, so that

$$\rho\left(u_{n_k}, \frac{t_k}{2}\right) \geq \rho(u_{n_k}, t),$$

and by letting $k \to \infty$ by part (i) we get

$$\liminf_{k\to\infty} \rho\left(u_{n_k}, \frac{t_k}{2}\right) \geq \mu. \tag{22}$$

By (20) and (22) we have that

$$\mu = \lim_{k\to\infty} \rho(u_{n_k}, t_k/2).$$

Similarly

$$\rho\left(u_{n_k}, \frac{t_k}{2}\right) \geq \rho(u_{n_k}, t) \Rightarrow \sup \rho\left(u_{n_k}, \frac{t_k}{2}\right) \geq \rho(u_{n_k}, t),$$

and by taking $k \to \infty$ and using (20) we get

$$\mu \geq \lim_{t\to\infty} \gamma(t). \qquad \square \tag{23}$$

LEMMA 7. *Let $(u_n)_{n\geq 0} \subset H^1$ be such that*

$$\|u_n\|_{L^2} \leq a < \infty,$$

$$\lim_{n\to\infty} \|u_n\|_{L^2}^2 = b > 0,$$

and

$$\sup_{n\geq 0} \|\nabla u_n\|_{L^2} < \infty.$$

Then there exists a subsequence $(u_{n_k})_{k\geq 0}$ which satisfies the following.

There exist $(q_k)_{k\geq 0}, (w_k)_{k\geq 0} \subset H^1(\mathbb{R})$ such that

$$\operatorname{supp} q_k \cap \operatorname{supp} w_k = \emptyset, \tag{24}$$

$$|q_k| + |w_k| \leq |u_{n_k}|, \tag{25}$$

$$\|q_k\|_{H^1} + \|w_k\|_{H^1} \leq C\|u_{n_k}\|_{H^1}, \tag{26}$$

$$\lim_{k\to\infty} \|q_k\|_{L^2}^2 = \mu, \qquad \lim_{k\to\infty} \|w_k\|_{L^2}^2 = b - \mu, \tag{27}$$

$$\liminf_{k\to\infty} \left\{ \int |\nabla u_{n_k}|^2 - \int |\nabla q_k|^2 - \int |\nabla w_k|^2 \right\} \geq 0, \tag{28}$$

$$\lim_{k\to\infty} \left| \int |u_{n_k}|^p - \int |q_k|^p - \int |w_k|^p \right| = 0 \tag{29}$$

for all $2 \leq p < \infty$.

Proof. We use the sequences $(u_{n_k})_{k\geq 0}$ and $(t_k)_{k\geq 0}$ constructed in the previous lemma. We fix $\theta, \phi \in C^\infty([0,\infty))$ such that $0 \leq \theta, \phi \leq 1$ and

$$\theta(t) = 1 \text{ for } 0 \leq t \leq \frac{1}{2}, \quad \theta(t) = 0 \text{ for } t \geq \frac{3}{4},$$

$$\phi(t) = 0 \text{ for } 0 \leq t \leq \frac{3}{4}, \quad \phi(t) = 1 \text{ for } t \geq 1,$$

and we set

$$q_k = \theta_k u_{n_k}, \quad w_k = \phi_k u_{n_k},$$

where

$$\theta_k = \theta\left(\frac{|x - y(u_{n_k}, \frac{t_k}{2})|}{t_k}\right), \quad \phi_k = \phi\left(\frac{|x - y(u_{n_k}, \frac{t_k}{2})|}{t_k}\right).$$

Now (24), (25), and (26) are immediate. To prove (27) we estimate

$$\rho\left(u_{n_k}, \frac{t_k}{2}\right) = \int_{|x-y(u_{n_k},\frac{t_k}{2})|\leq \frac{t_k}{2}} |u_{n_k}|^2 \leq \int |q_k|^2 \leq \int_{|x-y(u_{n_k},\frac{t_k}{2})|\leq t_k} |u_{n_k}|^2$$

$$\leq \int_{|x-y(u_{n_k},t_k)|\leq t_k} |u_{n_k}|^2 \leq \rho(u_{n_k}, t_k).$$

Applying the second part of Lemma 6 we immediately get

$$\lim_{k\to\infty} \|q_k\|_{L^2}^2 = \mu. \tag{30}$$

We now set $z_k = u_{n_k} - q_k - w_k$. Note that in particular $|z_k| \leq |u_{n_k}|$. We have

$$\begin{aligned}\int |z_k|^2 &\leq \int_{\frac{t_k}{2}\leq|x-y(u_{n_k},\frac{t_k}{2})|\leq t_k} |u_{n_k}|^2 \\ &= \int_{|x-y(u_{n_k},\frac{t_k}{2})|\leq t_k} |u_{n_k}|^2 - \int_{|x-y(u_{n_k},\frac{t_k}{2})|\leq \frac{t_k}{2}} |u_{n_k}|^2 \\ &\leq \int_{|x-y(u_{n_k},t_k)|\leq t_k} |u_{n_k}|^2 - \int_{|x-y(u_{n_k},\frac{t_k}{2})|\leq \frac{t_k}{2}} |u_{n_k}|^2 = \rho(u_{n_k}, t_k) - \rho\left(u_{n_k}, \frac{t_k}{2}\right)\end{aligned}$$

and again by Lemma 6 we have

$$\lim_{k\to\infty} \|z_k\|_{L^2}^2 = 0. \tag{31}$$

By the Cauchy–Schwartz inequality and the above we have that

$$\lim_{k\to\infty} \int u_{n_k} \bar{z}_k = 0.$$

But now by (24), (30), (31), and some trivial algebra we get after integration that

$$\lim_{k\to\infty} \|w_k\|_{L^2}^2 = b - \mu$$

and (27) follows. Also note that z_k is bounded in H^1 and converges to 0 in L^2, and by the Gagliardo–Nirenberg inequality, in L^p for any $2 \leq p < \infty$. Moreover, one can easily verify that

$$\left||u_{n_k}|^p - |q_k|^p - |w_k|^p\right| \leq C|u_{n_k}|^{p-1}|z_k|$$

and by Cauchy–Schwartz since z_k tends to 0 in L^p, (29) follows. Finally, (28) follows easily from the initial assumptions, the Cauchy–Schwartz inequality and the easy calculation

$$|\nabla u_{n_k}|^2 - |\nabla q_k|^2 - |\nabla w_k|^2 = |\nabla u_{n_k}|^2(1-\theta_k^2-\phi_k^2) - |u_{n_k}|^2(|\nabla\theta_k|^2+|\nabla\phi_k|^2)$$

$$-Re(\bar{u}_{n_k}\nabla u_{n_k})\cdot\nabla(\theta_k^2+\phi_k^2) \geq -\frac{C}{t_k^2}|u_{n_k}|^2 - \frac{C}{t_k}|u_{n_k}|\ |\nabla u_{n_k}|. \qquad \square$$

Proof of Theorem 3. Define the blowup parameters:

$$\lambda(t) = \|u(t)\|_{H^s}, \quad \Lambda(t) = \sup_{0\leq\tau\leq t}\lambda(\tau),$$
$$\sigma(t) = \|I_N\langle\nabla\rangle u(t)\|_{L^2}, \quad \Sigma(t) = \sup_{0\leq\tau\leq t}\sigma(\tau).$$

Let $\{t_n\}_{n=1}^{\infty}$ be a sequence such that $t_n \uparrow T^\star$ and for each t_n we have

$$\|u(t_n)\|_{H^s} = \Lambda(t_n)$$

and with $u(t_n) = u_n$ we define

$$I_N u_n = I_{N(t_n)}u(t_n).$$

We rescale these as follows:

$$v_n(x) = \frac{1}{\sqrt{\sigma_n}} I_N u_n\left(\frac{x}{\sigma_n}\right),$$

where

$$\sigma_n = \|I_N\langle\nabla\rangle u_n\|_{L^2} = \sigma(t_n).$$

Note that for these sequences (let us call them maximizing) we have that

$$\Lambda(t_n) \leq \sigma_n,$$

where $\sigma_n \to \infty$ as $n \to \infty$. It is important to note that we are in the blow-up regime and thus

$$\|u_0\|_{L^2} \geq \|Q\|_{L^2}.$$

Moreover, the L^2 part of v_n is bounded uniformly in n. This is because

$$\|v_n\|_{L^2} = \|I_N u_n\|_{L^2} \leq \|u_n\|_{L^2} = \|u(t_n)\|_{L^2} = \|u_0\|_{L^2}.$$

Thus in the limit as $n \to \infty$ we have

$$\lim_{n\to\infty}\|\nabla v_n\|_{L^2} = 1.$$

Also, since $N(t_n)$ goes to infinity as $n \to \infty$,

$$\lim_{n\to\infty}\|v_n\|_{L^2} = \lim_{n\to\infty}\|I_N u_n\|_{L^2} = \lim_{n\to\infty}\|u(t_n)\|_{L^2} = \|u_0\|_{L^2} \geq \|Q\|_{L^2}.$$

In addition, by Proposition 3 we have that

$$|E(v_n)| = \frac{1}{\sigma_n^2}|E(I_N u_n)| \leq C\sigma_n^{-2}\Lambda^{p(s)}(t_n) \leq C\Lambda^{p(s)-2}(t_n)$$

and thus

$$\lim_{n\to\infty} E(v_n) = 0$$

since $p(s) < 2$. This allows another way to prove that

$$\lim_{n\to\infty} \|v_n\|_{L^2} \geq \|Q\|_{L^2}$$

since by the optimality of the Gagliardo–Nirenberg inequality we have

$$E(v_n) \geq \frac{1}{2}\left(1 - \frac{\|v_n\|_{L^2}^4}{\|Q\|_{L^2}^4}\right)\|\nabla v_n\|_{L^2}^2$$

and in the limit as $n \to \infty$ we get

$$\lim_{n\to\infty} \|v_n\|_{L^2} \geq \|Q\|_{L^2}.$$

We collect the three important relations that we have:

$$\lim_{n\to\infty} \|v_n\|_{L^2} = \|u_0\|_{L^2} \geq \|Q\|_{L^2}, \tag{32}$$

$$\lim_{n\to\infty} \|\nabla v_n\|_{L^2} = 1, \tag{33}$$

$$\lim_{n\to\infty} E(v_n) = 0. \tag{34}$$

With the help of (32), (33), and (34) we will conclude the following claim.

Claim 1.

$$\mu(\{v_n\}_{n\geq 0}) \geq \|Q\|_{L^2}^2.$$

First assuming the claim and revisiting the statement of the theorem, it is enough to prove that for any $\epsilon > 0$ we have

$$\lim_{n\to\infty} \|u(t_n)\|_{L^2_{\{|x-z_n|<(T^\star-t_n)^{\frac{s}{2}}\gamma(T^\star-t_n)\}}} \geq \|Q\|_{L^2} - \epsilon.$$

Note that since $N(t_n)$ goes to ∞ we have

$$\lim_{n\to\infty} \|u(t_n)\|_{L^2} = \lim_{n\to\infty} \|I_{N(t_n)}u(t_n)\|_{L^2}.$$

Now given $\epsilon > 0$, the relation

$$\mu(\{v_n\}_{n\geq 0}) \geq \|Q\|_{L^2}^2$$

by Lemma 6 implies that there exists a $T \in \mathbf{R}$ such that

$$\rho(v_n, T) \geq \|Q\|_{L^2}^2 - \epsilon$$

for large n. Note that ρ is a nondecreasing function of t and it is crucial to find a fixed T such that $\rho(v_n, T) \geq \|Q\|_{L^2}^2 - \epsilon$ holds for any large n, for the given ϵ. That this T is independent of all the n's after some n large, up to a subsequence, is guaranteed by the two parts of Lemma 6. Thus the same T works for all large n.

Now setting $y_n = y(v_n, T)$ defined by Lemma 4, we have that

$$\liminf_{n\to\infty} \|v(t_n)\|_{L^2_{\{|x-y_n|<T\}}} \geq \|Q\|^2_{L^2} - \epsilon$$

and up to a subsequence

$$\lim_{n\to\infty} \|v(t_n)\|_{L^2_{\{|x-y_n|<T\}}} \geq \|Q\|^2_{L^2} - \epsilon. \tag{35}$$

But

$$\lim_{n\to\infty} \|v(t_n)\|_{L^2_{\{|x-y_n|<T\}}} = \lim_{n\to\infty} \left\| \frac{1}{\sqrt{\sigma_n}} I_N u_n \left(\frac{x}{\sigma_n}\right) \right\|_{L^2_{\{|x-y_n|<T\}}}$$

$$= \lim_{n\to\infty} \|I_n u_n(x)\|_{L^2_{\{|x-\frac{y_n}{\sigma_n}|<\frac{T}{\sigma_n}\}}} = \lim_{n\to\infty} \|u_n(x)\|_{L^2_{\{|x-\frac{y_n}{\sigma_n}|<\frac{T}{\sigma_n}\}}} = \lim_{n\to\infty} \|u_n(x)\|_{L^2_{\{|x-z_n|<\frac{T}{\sigma_n}\}}},$$

where $z_n = \frac{y_n}{\sigma_n}$. Moreover, note that $\frac{T}{\sigma_n} \to 0$ and that σ_n goes to infinity at least as fast as $(T^\star - t)^{-\frac{s}{2}}$. Thus there exists a function $\gamma(x) \uparrow \infty$ as $x \downarrow 0$ such that

$$\begin{aligned}\lim_{n\to\infty} \|u(t_n)\|_{L^2_{\{|x-z_n|<(T^\star-t_n)^{\frac{s}{2}}\gamma(T^\star-t_n)\}}} &\geq \lim_{n\to\infty} \|u_n(x)\|_{L^2_{\{|x-z_n|<\frac{T}{\sigma_n}\}}} \\ &= \lim_{n\to\infty} \|v(t_n)\|_{L^2_{\{|x-y_n|<T\}}} \geq \|Q\|^2_{L^2} - \epsilon\end{aligned}$$

by equation (35), and the proof is complete. □

Proof of Claim 1. We prove the claim by contradiction. We claim that there exists $\delta > 0$ with the following property. If $(v_n)_{n\geq 0} \in H^1$ is such that

$$\lim_{n\to\infty} \|v_n\|^2_{L^2} = \|u_0\|^2_{L^2}, \tag{36}$$

$$0 < \liminf_{n\to\infty} \|\nabla v_n\|_{L^2} \leq \limsup_{n\to\infty} \|\nabla v_n\|_{L^2} < \infty, \tag{37}$$

$$\limsup_{n\to\infty} E(v_n) \leq 0, \tag{38}$$

and

$$\mu(\{v_n\}_{n\geq 0}) < \|Q\|^2_{L^2}, \tag{39}$$

then there exists a sequence $(\tilde{v}_n)_{n\geq 0} \in H^1$ satisfying (37), (38), and (39) and such that

$$\lim_{n\to\infty} \|\tilde{v}_n\|^2_{L^2} = \|u_0\|^2_{L^2} - \beta$$

for some $\beta > \delta$. Clearly the sequence v_n of Theorem 3 satisfies (36), (37), and (38). But then by Galiardo–Nirenberg and (37), (38), and (39) we have that

$$\mu(\{v_n\}_{n\geq 0}) < \|u_0\|^2_{L^2} - \delta.$$

If we apply the above procedure k times, we get $\mu(\{v_n\}_{n\geq 0}) < \|u_0\|^2_{L^2} - k\delta$, which for large k is absurd. Thus it suffices to prove the claim. We apply Lemmas 6 and 7 to the sequence $(v_n)_{n\geq 0}$ and we consider the corresponding sequences $(q_n)_{n\geq 0}$ and $(w_n)_{n\geq 0}$. We set

$$\delta = \left(\frac{3}{K}\right)^{\frac{1}{2}} > 0,$$

where K is given in Lemma 5. We first show that

$$\mu(\{v_n\}_{n\geq 0}) \geq \delta.$$

By Lemma 5 and the definition of δ we have that

$$E(v_{n_k}) \geq \frac{1}{2}\left(1-\left(\frac{\rho(v_{n_k},t_k)}{\delta}\right)^2\right)\int |\nabla v_{n_k}|^2 - \frac{K}{6t_k^2}\rho(v_{n_k},t_k)^2.$$

Now if we assume by contradiction that $\mu < \delta$, then we obtain by letting $k \to \infty$, applying the second part of Lemma 6, and using (37) that up to a subsequence

$$\limsup_{n\to\infty} E(v_n) \geq \frac{1}{2}\left(1-\left(\frac{\mu}{\delta}\right)^2\right)\liminf_{n\to\infty}\int |\nabla v_n|^2 > 0,$$

which is absurd. Now since by (25) we know that $|w_k| \leq |v_{n_k}|$, we have by the second part of Lemma 6 and (39)

$$\mu((w_k)_{k\geq 0}) \leq \mu < \|Q\|_{L^2}^2.$$

This proves that $(w_k)_{k\geq 0}$ satisfies (39). Also by (27), (39), and the Gagliardo–Nirenberg inequality we know that there exists a $\sigma > 0$ such that for k large

$$E(q_k) \geq \sigma \|\nabla q_k\|_{L^2}^2. \tag{40}$$

On the other hand, by (28) and (29) we have that

$$\liminf_{k\to\infty}\{E(v_{n_k}) - E(q_k) - E(w_k)\} \geq 0 \tag{41}$$

and thus

$$\limsup_{k\to\infty} E(w_k) \leq 0.$$

This proves that $(w_k)_{k\geq 0}$ satisfies (38). By (26) and (37) we easily get that

$$\|\nabla w_k\|_{L^2} \leq C\|v_{n_k}\|_{H^1} < \infty.$$

Finally we show the last property (37), namely, that

$$\liminf_{k\to\infty} \|\nabla w_k\|_{L^2} > 0.$$

We argue again by contradiction and assume that there exists a sequence which we still denote by $(w_k)_{k\geq 0}$ such that $\lim_{k\to\infty} \|\nabla w_k\|_{L^2} = 0$. But then it trivially follows that $E(w_k) \to 0$ as $k \to \infty$ and thus by (38), (40), and (41) we get that

$$\lim_{k\to\infty} \|\nabla q_k\|_{L^2} = 0.$$

But then by (29) and the fact that $\|w_k\|_{L^6}, \|q_k\|_{L^6} \to 0$ we deduce that $\lim_{k\to\infty} \|v_{n_k}\|_{L^6} = 0$ and thus

$$\limsup_{k\to\infty} E(v_{n_k}) > 0,$$

which contradicts (38). Now setting

$$\tilde{v}_k = \frac{\sqrt{\|u_0\|_{L^2}^2 - \mu}}{\|w_k\|_{L^2}} w_k,$$

we see that with the help of (27) the sequence $(\tilde{v}_n)_{n\geq 0}$ satisfies (37), (38), and (39) and that

$$\lim_{k\to\infty} \|\tilde{v}_k\|_{L^2}^2 = \|u_0\|_{L^2}^2 - \mu \leq \|u_0\|_{L^2}^2 - \delta,$$

and we are done. ∎

Remark 6. In dimensions $n \geq 2$ with the additional assumption of radial symmetry on the initial data, the solution of the equivalent L^2-critical Schrödinger equation satisfies the conclusion of Theorem 3 with $z(t) \equiv 0$.

We define the "modified energy" for the system (16) as

$$E(Iu)(t) = \frac{1}{2}\int |Iu_x(t)|^2 dx - \frac{1}{6}\int |Iu(t)|^6 dx.$$

This "energy" functional is not conserved, but we can show that its time derivative decays with respect to a large parameter N. The next proposition quantifies the increament of this functional on $[0,\delta]$.

PROPOSITION 4. *Let u be an H^1 solution of* (2). *Then*

$$\begin{aligned} E(Iu)(\delta) - E(Iu)(0) &= Im\left(\int_0^\delta \int I\bar{u}_{xx}\left(I(|u|^4u) - Iu|Iu|^4\right) dxdt\right) \\ &\quad + Im\left(\int_0^\delta \int I(|u|^4u)\left(I(|u|^4u) - Iu|Iu|^4\right) dxdt\right). \end{aligned}$$

Proof. The derivative of the "modified energy" is

$$\begin{aligned} \frac{dE}{dt}(Iu) &= Im\left(\int I\bar{u}_{xx}\left(I(|u|^4u) - Iu|Iu|^4\right) dx\right) \\ &\quad + Im\left(\int I(|u|^4u)\left(I(|u|^4u) - Iu|Iu|^4\right) dx\right). \end{aligned}$$

But then Proposition 4 follows immediately by applying the fundamental theorem of calculus. ∎

By the previous formal identity we can deduce the desired decay of the "modified energy."

PROPOSITION 5. *For any Schwartz function u we have that*

$$E(Iu)(\delta) - E(Iu)(0) \lesssim \delta^{\frac{1}{4}-} N^{-\frac{3}{2}+} \|Iu\|_{X_\delta^{1,1/2+}}^6 + \delta^{\frac{1}{2}-} N^{-2+} \|Iu\|_{X_\delta^{1,1/2+}}^{10}.$$

Proof. First we establish

$$\left|\int_0^\delta \int I\bar{u}_{xx}\left(I(|u|^4u) - Iu|Iu|^4\right) dxdt\right| \lesssim \delta^{\frac{1}{4}-} N^{-\frac{3}{2}+} \|Iu\|_{X_\delta^{1,1/2+}}^6$$

or, by Plancherel's theorem, that

$$\left|\int_0^\delta \int_{\Gamma_6} \frac{\xi_1^2}{\langle \xi_1 \rangle} \left(\frac{m(\xi_2 + \cdots + \xi_6) - m(\xi_2) \cdots m(\xi_6)}{m(\xi_2) \cdots m(\xi_6)} \right) \hat{u}(\xi_1, t) \cdots \hat{\bar{u}}(\xi_6, t) d\xi dt \right|$$

$$\lesssim \delta^{1/4-} N^{-\frac{3}{2}+} \|u\|^5_{X^{1,1/2+}} \|u_1\|_{X^{0,1/2+}}, \tag{42}$$

where Γ_6 denotes the hyperplane $\xi_1 + \xi_2 + \cdots + \xi_6 = 0$, and u_1 is the function that corresponds on the Fourier side to the frequency ξ_1.

Remark 7.

1. Let us denote $N_i \sim |\xi_i|$ and $N_{max} \sim |\xi|_{max}$, $N_{med} \sim |\xi|_{med}$, where $|\xi|_{max}$, $|\xi|_{med}$ are the largest and second largest of the $|\xi_i|$. If all $|\xi_i| \ll N$, then the big parenthesis above is zero and there is nothing to prove. Thus since the ξ_i are related by $\xi_1 + \xi_2 + \cdots + \xi_6 = 0$, we have that $|\xi|_{max} \sim |\xi|_{med} \gtrsim N$. We also write m_i for $m(\xi_i)$ and m_{ij} for $m(\xi_i + \xi_j)$.
2. Our strategy from now on is to break all the functions into a sum of dyadic constituents ψ_j, each with frequency support $\langle \xi \rangle \sim 2^j, j = 0, \cdots$. Then we pull the absolute value of the symbols out of the integral, estimating it pointwise. After bounding the multiplier, the remaining integrals involving the pieces ψ_j are estimated by reversing Plancherel's formula and using duality, Hölder's inequality, and the Strichartz estimates. We can sum over all the frequency pieces ψ_j as long as we always keep a factor $N_{max}^{-\epsilon}$ inside the summation.
3. Since in all of the estimates that we establish from now on, the right-hand side is in terms of the $X^{s,b}$ norms and the $X^{s,b}$ spaces depend only on the absolute value of the Fourier transform, we can assume without loss of generality that the Fourier transform of all the functions in the estimates are real and positive.
4. Note also that $\frac{N_1^2}{\langle N_1 \rangle} \leq N_1$.

Since, as we mentioned before, our analyses do not rely upon the complex conjugate structure of the left-hand side, there is a symmetry under the interchange of the indices and thus we can assume that

$$N_2 \geq N_3 \geq \cdots \geq N_6.$$

Case 1. Let $N \gg N_2$. Then

$$\frac{m(\xi_2 + \cdots + \xi_6) - m(\xi_2) \cdots m(\xi_6)}{m(\xi_2) \cdots m(\xi_6)} = 0$$

and there is nothing to prove.

Case 2. $N_2 \gtrsim N \gg N_3 \geq \cdots \geq N_6$. This forces $N_1 \sim N_2$ on Γ_6. But then by the mean value theorem we have

$$\left|\frac{m(\xi_2 + \cdots + \xi_6) - m(\xi_2) \cdots m(\xi_6)}{m(\xi_2) \cdots m(\xi_6)}\right| = \left|\frac{m(\xi_2) - m(\xi_1)}{m(\xi_2)}\right|$$

$$\lesssim \left|\frac{\nabla m(\xi_2) \cdot (\xi_3 + \cdots + \xi_6)}{m(\xi_2)}\right| \lesssim \frac{N_3}{N_2}.$$

Now by undoing Plancherel's theorem, using the Cauchy–Schwartz inequality, applying the Strichartz estimates, and using Lemmas 2 and 3, we have that the left-hand side of (42) is

$$\lesssim \frac{N_1 N_3}{N_2 N_1^{1/2}} \|(D^{1/2}u_1)u_3\|_{L_t^2 L_x^2} \|u_2 u_4 u_5 u_6\|_{L_t^2 L_x^2}$$

$$\lesssim \frac{N_3}{N_1^{1/2}} \|u_1\|_{X_\delta^{0,1/2+}} \|u_3\|_{X_\delta^{0,1/2+}} \prod_{j=4}^{6} \|u_j\|_{L_t^\infty L_x^\infty} \|u_2\|_{L_t^2 L_x^2}$$

$$\lesssim \delta^{1/2-} \frac{N_3}{N_1^{1/2}} \prod_{j=1}^{3} \|u_j\|_{X_\delta^{0,1/2+}} \prod_{j=4}^{6} \|u_j\|_{X_\delta^{1/2,1/2+}},$$

where in the last inequality we also used (9) in its dyadic form. Comparing with (42) we see that it is enough to have

$$\delta^{1/2-} \frac{N_3 (N_4 N_5 N_6)^{1/2}}{N_1^{1/2}} \lesssim \delta^{1/4-} N^{-\frac{3}{2}+} N_{max}^{-\epsilon} N_2 \cdots N_6,$$

which is true. Note that in the process we summed the Littlewood–Paley pieces, using the factor $N_{max}^{-\epsilon}$.

Case 3. $N_2 \geq N_3 \gtrsim N$. In this case we use the crude estimate

$$|1 - \frac{m_1}{m_2 \cdots m_6}| \lesssim \frac{m_1}{m_2 \cdots m_6}.$$

Since it is impossible to have $N_1 \gg N_{med} = N_2$, we can divide this case into two subcases.

(a) $N_1 \sim N_2 \geq N_3 \gtrsim N$. Now we start by comparing the different frequencies in order to be able to apply Lemma 3. Note that $m_1 \sim m_2$.

(i) Suppose first that $N_2 \gg N_3$. Without loss of generality we can assume that $N_4 \leq N$. This is because in the case that one of the N_4, N_5, N_5 is $\gtrsim N$, the estimate is even easier and the decay is greater. For the suspicious reader who might object to the previous argument because of the presence of $m_4 m_5 m_6$ in the denominator, we comment that for $N_j \gtrsim N$ we have that

$$\frac{1}{m_j N_j^{1/2}} \lesssim \frac{1}{N^{1/2}}$$

and indeed we can get a better decay. From now on we will use this heuristic without any comment. Thus we can apply Cauchy–Schwartz and Lemma 3, and the left-hand side of (42) is

$$\lesssim \frac{N_1}{m_3 N_1^{1/2}} \|(D^{1/2}u_1)u_3\|_{L_t^2 L_x^2} \|u_2\|_{L_t^2 L_x^2} \cdot \prod_{j=4}^{6} \|u_j\|_{L_t^\infty L_x^\infty}$$

$$\lesssim \delta^{1/2-} \frac{N_1 N_3^{1/2}}{m_3 N_3^{1/2} N_1^{1/2}} \prod_{j=1}^{3} \|u_j\|_{X_\delta^{0,1/2+}} \cdot \prod_{j=4}^{6} \|u_j\|_{X_\delta^{1/2,1/2+}}$$

$$\lesssim \delta^{1/2-} \frac{N_1 N_3^{1/2}}{N^{1/2} N_1^{1/2}} \prod_{j=1}^{3} \|u_j\|_{X_\delta^{0,1/2+}} \cdot \prod_{j=4}^{6} \|u_j\|_{X_\delta^{1/2,1/2+}}.$$

Comparing with (42) we see that it is enough to have

$$\delta^{1/2-}\frac{N_1 N_3^{1/2}}{N^{1/2}N_1^{1/2}} \lesssim \delta^{1/4-}N^{-\frac{3}{2}+}N_{max}^{-\epsilon}(N_4\cdots N_6)^{1/2}N_2N_3,$$

which is true.

(ii) Now assume that $N_2 \sim N_3$ and by the comment in case (i) the worst case is when $N_5, N_6 \leq N$ which we assume without loss of generality. In this case we compare N_3 with N_4. In case that $N_3 \sim N_4$, the estimate is easy since

$$\frac{N_1 m_1}{m_2\cdots m_6} \lesssim \frac{N_1 N_3^{1/2}}{m_3 m_4 N_3^{1/2}} \lesssim \frac{N_1 N_3^{1/2}}{N^{1/2}}$$

and thus the left-hand side of (42) is

$$\lesssim \frac{N_1 N_3^{1/2}}{N^{1/2}}\prod_{j=1}^{4}\|u_j\|_{L_t^4L_x^4}\cdot\|u_5\|_{L_t^\infty L_x^\infty}\|u_6\|_{L_t^\infty L_x^\infty}$$

$$\lesssim \delta^{1/2-}\frac{N_1 N_3^{1/2}}{N^{1/2}}\prod_{j=1}^{4}\|u_j\|_{X_\delta^{0,1/2+}}\cdot\|u_5\|_{X_\delta^{1/2,1/2+}}\|u_6\|_{X_\delta^{1/2,1/2+}}.$$

Comparing with (42) it is enough to have

$$\delta^{1/2-}\frac{N_1 N_3^{1/2}(N_5N_6)^{1/2}}{N^{1/2}} \lesssim \delta^{1/4-}N^{-\frac{3}{2}+}N_{max}^{-\epsilon}N_2\cdots N_6,$$

which is true. If $N_3 \gg N_4$ again without loss of generality we assume that $N_4 \leq N$ and we apply Lemma 3. Moreover,

$$\left|\frac{N_1 m_1}{m_2\cdots m_6}\right| \lesssim \frac{N_1 N_3^{1/2}}{m_3 N_3^{1/2}} \lesssim \frac{N_1 N_3^{1/2}}{N^{1/2}}$$

and thus the left-hand side of (42) is

$$\lesssim \frac{N_1 N_3^{1/2}}{N^{1/2}N_1^{1/2}}\|(D^{1/2}u_1)u_4\|_{L_t^2L_x^2}\|u_2u_3\|_{L_t^2L_x^2}\|u_5\|_{L_t^\infty L_x^\infty}\|u_6\|_{L_t^\infty L_x^\infty}$$

$$\lesssim \frac{N_1 N_3^{1/2}}{N^{1/2}N_1^{1/2}}\|u_1\|_{X_\delta^{0,1/2+}}\|u_4\|_{X_\delta^{0,1/2+}}\|u_2\|_{L_t^6L_x^6}\|u_3\|_{L_t^3L_x^3}\|u_5\|_{X_\delta^{1/2,1/2+}}\|u_6\|_{X_\delta^{1/2,1/2+}}$$

$$\lesssim \delta^{1/4-}\frac{N_1 N_3^{1/2}}{N^{1/2}N_1^{1/2}}\prod_{j=1}^{4}\|u_j\|_{X_\delta^{0,1/2+}}\|u_5\|_{X_\delta^{1/2,1/2+}}\|u_6\|_{X_\delta^{1/2,1/2+}},$$

where we used Lemmas 2 and 3 and (7), (9), and (11). Comparing with (42) it is enough to have

$$\delta^{1/4-}\frac{N_1 N_3^{1/2}(N_5N_6)^{1/2}}{N^{1/2}N_1^{1/2}} \lesssim \delta^{1/4-}N^{-\frac{3}{2}+}N_{max}^{-\epsilon}N_2\cdots N_6,$$

which is true.

(b) $N_2 \sim N_3 \gtrsim N$ and $N_2 \gg N_1$. Since N_1 is in the numerator on the left-hand side of (42), this case is easier than the previous and similar analysis gives the same (or ever better) bounds as in (a). The details are omitted.

To conclude the proof of Proposition 5 it remains to show that

$$\left|\int_0^\delta \int I(|u|^4 u)\left(I(|u|^4 u) - Iu|Iu|^4\right) dxdt\right| \lesssim \delta^{\frac{1}{2}-} N^{-2+} \|Iu\|^{10}_{X^{1,1/2+}_\delta}.$$

By Plancherel's theorem,

$$\left|\int_0^\delta \int_{\Gamma_{10}} m_{12345}\{m_{678910} - m_6 \cdots m_{10}\}\hat{u}(\xi_1,t)\cdots\hat{\bar{u}}(\xi_{10},t)d\xi dt\right| \lesssim \delta^{\frac{1}{2}} N^{-2+}\|Iu\|^{10}_{X^{1,1/2+}_\delta}. \tag{43}$$

As we noted before, if $N_{max} \ll N$, the multiplier is zero so we assume that

$$N_{max} \sim N_{med} \gtrsim N.$$

In addition since $m(\xi) \leq 1$, we have that

$$|m_{12345}\{m_{678910} - m_6 \cdots m_{10}\}| \lesssim C.$$

Finally the last pointwise estimate that we use is

$$\frac{1}{m_{max} N_{max}} \lesssim N^{-1},$$

which follows easily since $N_{max} \gtrsim N$. The left-hand side of (43) is

$$\begin{aligned}
&\lesssim \int_0^\delta \int_{\Gamma_{10}} \prod_{j=1}^{10} \hat{u}(\xi_j,t)d\xi dt \lesssim \int_0^\delta \int_{\Gamma_{10}} \frac{m_{max}N_{max}\hat{u}_{max} \cdot m_{med}N_{med}\hat{u}_{med}}{m_{max}N_{max}m_{med}N_{med}} \cdot \\
&\qquad\qquad \prod_{j \neq j_{max}, j_{med}} \hat{u}(\xi_j,t)d\xi dt \\
&\qquad \lesssim N^{-2+}N^{-\epsilon}_{max} \int_0^\delta \int_{\Gamma_{10}} \widehat{DIu_{max}} \cdot \widehat{DIu_{med}} \prod_{j \neq j_{max}, j_{med}} \hat{u}(\xi_j,t)d\xi dt.
\end{aligned}$$

Now reversing Plancherel's theorem and using the estimates

$$\|u\|_{L^6_t L^6_x} \lesssim \|u\|_{X^{0,1/2+}_\delta},$$

$$\|u\|_{L^3_t L^3_x} \lesssim \delta^{1/4-}\|u\|_{X^{0,1/2+}_\delta},$$

$$\|u\|_{X^{0,1/2+}_\delta} \lesssim \|Iu\|_{X^{1,1/2+}_\delta},$$

$$\|u\|_{X^{1/2,1/2+}_\delta} \lesssim \|Iu\|_{X^{1,1/2+}_\delta},$$

we get that the left-hand side of (43) is

$$\lesssim N^{-2+}N^{-\epsilon}_{max}\|JIu_{max}\|_{L^6_t L^6_x} \cdot \|JIu_{med}\|_{L^6_t L^6_x}\|u\|^2_{L^3_t L^3_x}\|u\|^6_{L^\infty_t L^\infty_x}$$

$$\lesssim \delta^{1/2-} N^{-2+} N_{max}^{-\epsilon} \|JIu\|^2_{X_\delta^{0,1/2}} \|u\|^2_{X_\delta^{0,1/2}} \|u\|^6_{X_\delta^{1/2,1/2}} \lesssim \delta^{\frac{1}{2}-} N^{-2+} \|Iu\|^{10}_{X_\delta^{1,1/2+}},$$

where in the process we sum the different Littlewood–Paley pieces, taking advantage of the factor $N_{max}^{-\epsilon}$. ∎

Now we are finally ready to prove Proposition 3.

Proof. When $s = 1$ we can choose $N(T) = +\infty$ and thus $I_{N(T)} = 1$ and the proposition is true with $p(s) = 0$ since the energy is conserved and the kinetic energy blows up as time approaches $T^\star$. Therefore we can fix $10/11 < s < 1$ and take T near $T^\star$. Now let $N = N(T)$ to be chosen later in the argument. Recall that $\delta \sim (\Sigma(T))^{-4-\epsilon}$ gives the time of the local well-posedness. Thus if we divide the interval $[0, T]$ into $\frac{T}{\delta}$-subintervals of size $\sim \delta$, the local well-posedness result uniformly applies. Moreover for any t in this subinterval we have that

$$\|I\langle\nabla\rangle u(t)\|_{L^2} = \sigma(t) \leq \Sigma(T).$$

The next step is to iterate the almost conservation of the energy. It is apparent that after $\frac{T}{\delta}$ steps the growth of the modified energy is

$$\begin{aligned} E(Iu(T)) &\lesssim E(Iu(0)) + \frac{T^\star}{\delta}\{\delta^{\frac{1}{4}-} N^{-\frac{3}{2}+}\Sigma(T)^6 + \delta^{\frac{1}{2}-} N^{-2+}\Sigma(T)^{10}\} \\ &\leq N^{2(1-s)}\lambda(0) + \frac{T^\star}{\delta}\{\delta^{\frac{1}{4}-} N^{-\frac{3}{2}+}\Sigma(T)^6 + \delta^{\frac{1}{2}-} N^{-2+}\Sigma(T)^{10}\} \\ &\lesssim N^{2(1-s)} + \{\delta^{-\frac{3}{4}-} N^{-\frac{3}{2}+}\Sigma(T)^6 + \delta^{-\frac{1}{2}-} N^{-2+}\Sigma(T)^{10}\} \\ &\lesssim N^{2(1-s)} + N^{-\frac{3}{2}+}\Sigma(T)^{9+} + N^{-2+}\Sigma(T)^{12+}, \end{aligned}$$

where in the third inequality we dismiss the irrelevant constants. Now if we switch from $\Sigma(T)$ to $\Lambda(T)$ we have

$$E(Iu(T)) \lesssim N^{2(1-s)} + N^{-\frac{3}{2}+} N^{9(1-s)+}\Lambda(T)^{9+} + N^{-2+} N^{12(1-s)+}\Lambda(T)^{12+}.$$

We know choose $N = N(T)$ so that

$$N^{2(1-s)} \sim N^{-2+} N^{12(1-s)+}\Lambda(T)^{12+},$$
$$N(T) \sim \Lambda(T)^{\frac{12}{10s-8}+}.$$

This establishes Proposition 3 with

$$p(s) = \frac{2\cdot 12(1-s)}{10s-8} < 2$$

which is true for $s > 10/11$. We emphasize that for $1 > s > 10/11$ the second term in the conservation of the modified energy formula produces a smaller correction. ∎

Acknowledgment. I would like to thank Jim Colliander for introducing me to this problem. I am also grateful to him for the beautiful discussions that we had while I was preparing the manuscript.

REFERENCES

[1] J. Bourgain, *Refinements of Strichartz' inequality and applications to* 2*D-NLS with critical nonlinearity,* Internat. Math. Res. Notices, 5 (1998), pp. 253–283.

[2] J. Bourgain and W. Wang, *Construction of blow-up solutions for the nonlinear Schrödinger equation with critical nonlinearity,* Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) 25, (1997), pp. 197–215.

[3] T. Cazenave, *Semilinear Schrödinger Equations,* Courant Lect. Notes Math. 10, New York University, Courant Institute of Mathematical Sciences, New York, AMS, Providence, RI, 2003.

[4] T. Cazenave and F. B. Weissler, *The Cauchy problem for the critical nonlinear Schrödinger equation in* H^s*,* Nonlinear Anal., 14 (1990), pp. 807–836.

[5] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Global well-posedness for Schrödinger equations with derivative,* SIAM J. Math. Anal., 33 (2001), pp. 649–669.

[6] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Sharp multi-linear periodic KdV estimates and applications,* J. Funct. Anal., 211 (2004), pp. 173–218.

[7] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *A refined global well-posedness result for Schrödinger equations with derivative,* SIAM J. Math. Anal., 34 (2002), pp. 64–86.

[8] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Sharp global well-posedness for KdV and modified KdV on* $\mathbb{R}$ *and* $\mathbb{T}$, J. Amer. Math. Soc., 16 (2003), pp. 705–749.

[9] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Almost conservation laws and global rough solutions to a nonlinear Schrödinger equation*, Math. Res. Lett., 9 (2002), pp. 659–682.

[10] J. Colliander, M. Keel, G. Staffilani, H. Takaoka, and T. Tao, *Global well-posedness and scattering for the energy-critical nonlinear Schrödinger equation in* $\mathbf{R^3}$, Ann. of Math., to appear.

[11] J. Colliander, S. Raynor, C. Sulem, and J. D. Wright, *Ground state mass concentration in the* L^2 *critical nonlinear Schrödinger equation below* H^1, Math. Res. Lett., 12 (2005), pp. 357–375.

[12] J. Ginibre, Y. Tsutsumi, and G. Velo, *On the Cauchy problem for the Zakharov system,* J. Funct. Anal., 151 (1997), pp. 384–436.

[13] C. E. Kenig, G. Ponce, and L. Vega, *A bilinear estimate with applications to the Kdv equations,* J. Amer. Math. Soc., 9 (1996), pp. 573–603.

[14] M. J. Landman, G. C. Papanicolaou, C. Sulem, and P-L. Sulem, *Rate of blow up for solutions of the nonlinear Schrödinger equation at critical dimensions,* Phys. Rev A. (3) 38 (1988), pp. 3837–3843.

[15] P.-L Lions, *The concentration-compactness principle in the calculus of variations. The locally compact case.* I, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), pp. 109–145.

[16] P.-L Lions, *The concentration-compactness principle in the calculus of variations. The locally compact case.* II, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), pp. 223–283.

[17] Y. Martel and F. Merle, *Blow up in finite time and dynamics of blow up solutions for the* L^2*-critical generalized KdV equation,* J. Amer. Math. Soc., 15 (2002), pp. 617–664.

[18] F. Merle and P. Raphael, *Sharp upper bound on the blow-up rate for the critical nonlinear Schrödinger equation,* Geom. Funct. Anal., 13 (2003), pp. 591–642.

[19] F. Merle and P. Raphael, *Blow-up dynamic and upper bound on the blow-up rate for critical nonlinear Schrödinger equation,* Ann. of Math., 161 (2005), pp. 157–222.

[20] F. Merle and Y. Tsutsumi, L^2 *concentration of blow-up solutions for nonlinear Schrödinger equation with critical power nonlinearity,* J. Differential Equations, 84 (1990), pp. 205–214.

[21] F. Merle and L. Vega, *Compactness at blow-up time for* L^2 *solutions of the critical nonlinear Schrödinger equation in* 2*D,* Internat. Math. Res. Notices, 8 (1998), pp. 399–425.

[22] H. Nawa, *"Mass concentration" phenomenon for the nonlinear Schrödinger equation with critical power nonlinearity,* Funkcial. Ekvac., 35 (1992), pp. 1–18.

[23] H. Pecher, *Global solutions with infinite energy for the* 1*-dimensional Zakharov system,* Electron. J. Differential Equations, 41 (2005), 18pp (electronic).

[24] G. Perelman, *On the formation of singularities in solutions of the critical nonlinear Schrödinger equation,* Ann. Henri Poincaré, 2 (2001), pp. 605–673.

[25] W. Strauss, *Existence of solitary waves in higher dimensions,* Comm. Math. Phys., 55 (1977), pp. 149–162.

[26] C. Sulem and P-L. Sulem, *The nonlinear Schrödinger equation. Self-focusing and wave collapse,* Appl. Math. Sci. 139 Springer, 1999.

[27] N. Tzirakis, *The Cauchy problem for the semilinear quintic Schrödinger equation in one dimension,* Differential Integral Equations, 18 (2005), pp. 947–960.
[28] M. Weinstein, *On the structure and formation of singularities of solutions to nonlinear dispersive evolution equations,* Comm. Partial Differential Equations, 11 (1986), pp. 545–565.
[29] M. Weinstein, *Nonlinear Schrödinger equations and sharp interpolation estimates,* Comm. Math. Phys., 87 (1983), pp. 567–576.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1947–1977

FREE BOUNDARY PROBLEMS FOR NONLINEAR WAVE SYSTEMS: MACH STEMS FOR INTERACTING SHOCKS*

SUNČICA ČANIĆ†, BARBARA LEE KEYFITZ‡, AND EUN HEUI KIM§

Abstract. We study a family of two-dimensional Riemann problems for compressible flow modeled by the nonlinear wave system. The initial constant states are separated by two jump discontinuities, $x = \pm\kappa_a y$, which develop into two interacting shock waves. We consider shock angles in a range where regular reflection is not possible. The solution is symmetric about the y-axis and on each side of the y-axis consists of an incident shock, a reflected compression wave, and a Mach stem. This has a clear analogy with the problem of shock reflection by a ramp. It is well known that no triple point structure exists in which incident, reflected, and Mach stem shocks meet at a point. In this paper, we model the reflected wave by a continuous function with a singularity in the derivative. This fails to be a weak solution across the sonic line. We show that a solution to the free boundary problem for the Mach stem exists, and we conjecture that the global solution can be completed by the construction of a reflected shock, by a similar free boundary technique.

The point of our paper is the capability to deal analytically with a Mach stem by solving a free boundary problem. The difficulties associated with the analysis of solutions containing Mach stems include (1) loss of obliqueness in the derivative boundary condition corresponding to the jump conditions across the Mach stem, and (2) loss of ellipticity at the formation point of the Mach stem.

We use barrier functions to show that for sufficiently large values of κ_a the subsonic solution is continuous up to the sonic line at the Mach stem.

Key words. two-dimensional conservation laws, degenerate elliptic equations, free boundary problems, self-similar solutions, Riemann problems

AMS subject classifications. Primary: 35L65, 35J70, 35R35; Secondary: 35M10, 35J65, 76J20.

DOI. 10.1137/S003614100342989X

1. Introduction.

This paper marks another step in our program to solve two-dimensional Riemann problems for hyperbolic conservation laws. Our first results involved a method [6] for solving the free boundary problems which arise in the study of small time-independent perturbations of steady transonic shocks in the small disturbance equation. We extended this technique to analyze quasi-steady transonic shocks that are not necessarily small perturbations of known solutions by focusing on weak shock reflection by a wedge, modeled by the unsteady transonic small disturbance (UTSD) equation. We solved this problem in two stages: first, a case corresponding to strong regular reflection in which the free boundary involved a strictly elliptic subsonic state [3] and, second, the case of weak regular reflection, in which proving existence of the free boundary was complicated by the failure of strict ellipticity in the downstream state [4]. The use of a simplified equation was necessary, as

*Received by the editors June 13, 2003; accepted for publication (in revised form) August 4, 2005; published electronically March 3, 2006.
http://www.siam.org/journals/sima/37-6/42989.html
†Department of Mathematics, University of Houston, Houston, TX 77204-3008 (canic@math.uh.edu). This author's research was supported by National Science Foundation grants DMS-0225948, DMS-0244343, and DMS-0245513.
‡Department of Mathematics, University of Houston, Houston, TX 77204-3008 (bkeyfitz@fields.utoronto.ca). This author's research was supported by Department of Energy grant DE-FG-03-94-ER25222.
§Department of Mathematics, California State University, Long Beach, CA 90840-1001 (ekim4@csulb.edu). This author's research was supported by National Science Foundation grant DMS-0103823 and Department of Energy grant DE-FG-02-03-ER25571.

our construction relied in an essential way on reducing the self-similar system to a second-order equation with particular structure which changed type from hyperbolic to elliptic across the sonic line. In [3, 4] we obtained an existence result in only a finite neighborhood of the shock reflection point.

More recently, we have outlined a program for extending our results to a larger class of equations, choosing for a model the nonlinear wave system [5]. This system is a slightly more realistic simplification of the compressible Euler equations of gas dynamics and hence is a better test case for the program. It offers the advantage of being linearly well-posed in space and time (which the UTSD equation is not) and of having a nonlinear acoustic-wave dependence similar to the gas dynamics equations. It also has the convenient feature, just as the UTSD equations have, of reducing to a second-order quasi-linear self-similar equation which, at the sonic line, changes type from hyperbolic to elliptic. It has the additional feature, a more realistic prototype for gas dynamics, of being coupled to a transport equation, so that the change of type takes one from a hyperbolic to a mixed type system. The feature that makes the system more tractable than gas dynamics is that the coupling is very weak: it comes into play only at the point of reconstructing the solution in primitive variables.

As indicated in [5], a number of obstacles must be overcome before a theory for the general Riemann solution, even for this simplified model, can be given. The present result looks at a prototype for a Mach stem. We consider a problem characterized by symmetry and otherwise simplified data. Our eventual goal is to cover all situations which arise with general sectorially constant data. The innovations in this paper are twofold:

1. We are able handle the entire Mach stem without cutoff functions.
2. We overcome the technical difficulty posed by the fact that at the foot of a Mach stem the static boundary condition on the free boundary is no longer a uniformly oblique derivative condition.

We prove existence of a solution in a case where the equation is sonic at the formation point of the Mach stem. However, the correct modeling of the shock interaction is limited to the Mach stem and interaction point itself; we have not attempted to construct the reflected shock. Rather, we have replaced the reflected shock by a weak shock at the sonic boundary, which does not give a weak solution in the neighborhood of the sonic line. Although we have not completely solved the problem, we feel that our result is a significant advance and that this approach will help in solving the full problem. We explain this in section 5.

The analysis applies to the nonlinear wave system (NLWS), a reduction of the inviscid system for compressible isentropic gas dynamics, obtained by neglecting the inertial terms. The system is

$$\begin{aligned} \rho_t + m_x + n_y &= 0, \\ m_t + p(\rho)_x &= 0, \\ n_t + p(\rho)_y &= 0. \end{aligned} \tag{1.1}$$

We consider (1.1) with sectorially constant Riemann data consisting of two states separated by discontinuities at $x = \pm\kappa_a y$ for $y \geq 0$ and with the states chosen so that the one-dimensional Riemann problems at each discontinuity are solved by upward-moving shocks and linear waves only. These determine the solution in the far field. One expects to see a shock interaction consisting either of regular reflection or of Mach reflection, depending on whether the angle between the incident shocks is small or large (see [16]). The scenario we study here for Mach reflection places the formation

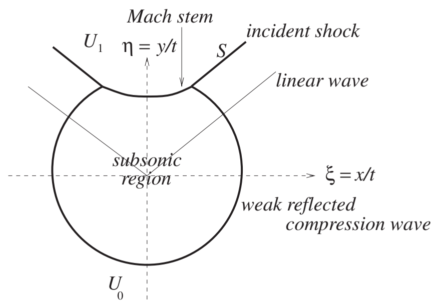


FIG. 1.1. *Sketch of global solution structure.*

point of the Mach stem exactly at the sonic circle, and hence, since this system does not admit triple points, the reflected wave has strength zero at this formation point. This scenario thus requires that the angle between the incident shocks be large enough that the shocks intersect the sonic circle before their extensions intersect each other. Numerical simulations in [17] suggest that such a formation does indeed occur and suggest, further, that the reflected shock is weak.

In this paper, we match a piecewise constant solution outside the sonic circle with a solution of the self-similar equation inside the sonic circle, demanding continuity at the circle. See Figure 1.1 for a sketch of such a solution. Our main result is the existence of a solution to the subsonic problem. The composite function is not a weak solution across the sonic circle. This leaves open the question of what is the actual solution; it differs from the construction here and from the simulations. One possibility is that the reflected wave is a weak, nearly circular shock, which has strength zero at the formation point. Based on the successful construction of the Mach stem in this paper, it may be possible to solve the complete problem by finding this reflected shock as the solution of another free boundary problem. Another possibility is a cascade of supersonic patches, as reported by Tesdall and Hunter for the UTSD equation [25]. We leave this for a future paper.

The techniques we use in this paper to prove global existence of a solution rely on an application of the Schauder fixed point theorem, developed in [6, 3, 4]. A similar approach was used by Chen and Feldman to prove stability of steady transonic shocks for the full potential equation [8, 9]. Chen and Feldman use the potential formulation of the equation to obtain a second-order operator. Both approaches prove existence of a fixed point which solves the underlying free boundary problem. The main difference lies in the compactness arguments used. Owing to the presence of the gradient of the potential in the principal coefficient of the full potential operator, the mapping in [8, 9] is not compact, but it is shown to operate on a compact space. Steady transonic shock perturbation analyses, both in [6] and in [8, 9], examine small perturbations of a uniform solution. A perturbation analysis of steady transonic shocks is also given by Chen, Geng, and Li in [12]. Using partial hodograph transformations which map the free boundary (shock) into a fixed boundary, combined with classical elliptic techniques, [12] obtains stability results for perturbations of conical shocks attached to the tip of a perturbed cone. Chen has used this same partial hodograph technique in a quasisteady problem [11] and has also found an analytical solution for a linearized problem corresponding to quasisteady regular reflection in the gas dynamics equations, [10].

The compressible Euler equations cannot, in general, be written in potential form and self-similar reduction of the compressible Euler equations (see [5, 24, 26]) leads to a system related in structure to the model studied in the present paper. In this connection, we mention also recent work by Zheng on diverging shocks in the pressure gradient system, a type of nonlinear wave system, [27].

In section 2, we derive the second-order operator and derivative boundary condition at the shock for the nonlinear wave system, (1.1); give the technical statement of our result, Theorem 2.3; set up the mapping to find the free boundary; and establish some preliminary estimates. In section 3, using a regularized differential operator, with $\varepsilon\Delta$ added, we prove the existence of a fixed point corresponding to the free boundary for the uniformly elliptic problem. The main point here is to deal with loss of obliqueness in the derivative boundary condition. In section 4, we proceed to the limit $\varepsilon \to 0$. The novelty here is that a uniform upper barrier at the intersection of the Mach stem with the sonic line cannot be found by standard barrier estimates. In section 5, we explain the significance of the result in providing a first step in the construction of Mach stems and other configurations where oblique derivative boundary conditions can become degenerate and where shocks cross the sonic line.

2. Background on the nonlinear wave system. Our point of departure is the compressible Euler system for isentropic flow in two space dimensions,

$$\begin{aligned} \rho_t + (u\rho)_x + (v\rho)_y &= 0, \\ (u\rho)_t + (u^2\rho + p)_x + (uv\rho)_y &= 0, \\ (v\rho)_t + (uv\rho)_x + (v^2\rho + p)_y &= 0, \end{aligned} \tag{2.1}$$

where ρ, u, and v are the density and the components of velocity, respectively, and $p = p(\rho)$ is the pressure. While we have in mind a power-law relation $p(\rho) = A\rho^\gamma$, where $\gamma > 1$ is the ratio of specific heats, all that we require in this paper is $p' > 0$ and $p'' > 0$. We recall that the local speed of sound is c and that $c^2 = dp/d\rho$. The nonlinear wave system is a reduction of (2.1) obtained by neglecting the quadratic terms in u and v. (We do not know if any physical situation is represented by this assumption. However, it underlies the scaling for Stokes flow and was used by Pironneau [23] in a case study of the shallow-water equations, which are modeled by (2.1) with $\gamma = 2$.) In the resulting nonlinear wave system, (1.1), we work with the conserved momentum variables $(m, n) = (\rho u, \rho v)$. The NLWS (1.1) can be written as a second-order nonlinear wave equation for the density and a transport equation for the specific vorticity $\omega = n_x - m_y$:

$$\begin{aligned} \rho_{tt} &= \nabla(c^2(\rho)\nabla\rho), \\ \omega_t &= 0. \end{aligned} \tag{2.2}$$

Since ω is stationary in this simplification of (2.1), then in any regions where the initial data satisfy the irrotationality condition $n_x = m_y$, the solutions, classical or weak, satisfy the same condition.

Introducing self-similar coordinates $\xi = x/t, \eta = y/t$, we can write the system (1.1) as

$$-\xi\rho_\xi - \eta\rho_\eta + m_\xi + n_\eta = 0, \tag{2.3}$$

$$-\xi m_\xi - \eta m_\eta + c^2(\rho)\rho_\xi = 0, \tag{2.4}$$

$$-\xi n_\xi - \eta n_\eta + c^2(\rho)\rho_\eta = 0. \tag{2.5}$$

In self-similar coordinates the nonlinear wave equation in (2.2), with its principal part in divergence form, is

$$Q(\rho) \equiv \big((c^2 - \xi^2)\rho_\xi - \xi\eta\rho_\eta\big)_\xi + \big((c^2 - \eta^2)\rho_\eta - \xi\eta\rho_\xi\big)_\eta + \xi\rho_\xi + \eta\rho_\eta = 0. \tag{2.6}$$

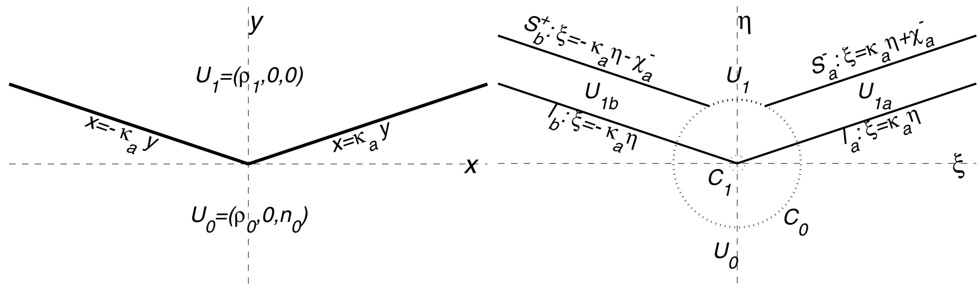


FIG. 2.1. *Riemann data and far-field solution.*

The equation is hyperbolic when $c^2(\rho) < \xi^2 + \eta^2$, elliptic when $c^2(\rho) > \xi^2 + \eta^2$, and degenerate on the sonic circle $c^2(\rho) = \xi^2 + \eta^2$.

It is because we can formulate the problem in terms of ρ that we can apply our fixed point method to this equation.

2.1. Setting up the problem. We consider two-dimensional Riemann data which are constant in sectors. Specifically, in this paper we look at data which correspond to two symmetric converging shocks. This may alternatively be regarded as the reflection of an oblique shock at a vertical wall. The data are constant in two sectors bounded by $\{x = \pm\kappa_a y, y \geq 0\}$ and symmetric with respect to $x = 0$, as shown in Figure 2.1. Let U denote the vector of conserved quantities, $U = (\rho, m, n)$. The Riemann data are

$$U(x, y, 0) = \begin{cases} U_1 \equiv (\rho_1, 0, 0), & -\kappa_a y < x < \kappa_a y, \quad y > 0, \\ U_0 \equiv (\rho_0, 0, n_0) & \text{otherwise.} \end{cases} \tag{2.7}$$

To obtain converging shocks in the far field, we choose $\rho_0 > \rho_1$ and determine n_0, depending on ρ_1, ρ_0, and κ_a, so that the one-dimensional wave between U_0 and U_1 at angle κ_a consists of a backward shock, S_a^-, and a linear wave, l_a, with a state U_{1a} between them:

$$S_a^- : \{\xi = \kappa_a \eta + \chi_a^-\}, \quad l_a : \{\xi = \kappa_a \eta\}, \quad U_{1a} = (\rho_0, m_{1a}, n_{1a}). \tag{2.8}$$

Using the formula (6.1) in [5] these values are

$$\begin{aligned} \chi_a^- &= -\sqrt{1+\kappa_a^2}\sqrt{\frac{p(\rho_0) - p(\rho_1)}{\rho_0 - \rho_1}}; \\ m_{1a} &= -\sqrt{\frac{(p(\rho_0) - p(\rho_1))(\rho_0 - \rho_1)}{1+\kappa_a^2}}; \quad n_{1a} = -\kappa_a m_{1a}; \\ n_0 &= \frac{1}{\kappa_a}\sqrt{(1+\kappa_a^2)(p(\rho_0) - p(\rho_1))(\rho_0 - \rho_1)}. \end{aligned} \tag{2.9}$$

By symmetry, the resolution of the discontinuity at $x = -\kappa_a y$ is

$$S_b^+ : \{\xi = -\kappa_a \eta - \chi_a^-\}, \quad l_b : \{\xi = -\kappa_a \eta\}, \quad U_{1b} = (\rho_0, -m_{1a}, n_{1a}).$$

For the Riemann data (2.7), the sonic circle is important:

$$C_0 \equiv \{(\xi, \eta) : \xi^2 + \eta^2 = c_0^2 \equiv c^2(\rho_0)\}. \tag{2.10}$$

We also define $C_1 \equiv \{(\xi, \eta); \xi^2 + \eta^2 = c_1^2 \equiv c^2(\rho_1)\}$.

Several types of shock interaction seem possible in this model, depending on the relative positions of the incident shock and the sonic circle. They are described in more detail in [17]. For small κ_a, the shocks intersect at a point $\Xi_c \equiv (0, \eta_c) = S_b^+ \cap S_a^- = (0, -\chi_a^-/\kappa_a)$ on the η axis, and two symmetric downward-moving shocks leave Ξ_c. For values of κ_a less than a critical value κ_R which depends on ρ_0 and ρ_1 one expects two solutions of this form, corresponding to "weak" and "strong" regular reflection. For $\kappa_a > \kappa_R$, no solutions of this form exist. On the other hand, for κ_a greater than a value κ_A (with $\kappa_A > \kappa_R$), one finds that $\eta_c < c_0$, so Ξ_c is inside the sonic circle C_0, and the farfield shocks intersect C_0 before reaching the symmetry axis. In this case, it is appealing to believe that a solution like that shown in Figure 1.1 is possible: the subsonic flow interacts with the shocks, which bend to form a single discontinuity; and the flow is continuous at C_0 below the shock. This phenomenon can be thought of as a perturbation of the uniform case $\kappa_a = \infty$.

In this paper, we prove the existence of a solution to the subsonic problem which contains a Mach stem and is continuous up to the sonic line, for sufficiently large values of κ_a; that is, $\kappa_a > \kappa_* > \kappa_A$. In the remainder of the paper, we assume $\kappa_a > \kappa_A$. The paper [17] gives a more detailed discussion of the regions. There, we also give scenarios (without proof) for solutions in the intermediate range of κ where neither regular reflection nor a solution with a weak reflected wave exists.

2.2. The shock evolution equation. At a shock, the Rankine–Hugoniot jump conditions are satisfied across the line of discontinuity. A key element of our solution method has been to rewrite the equations as a problem for a single variable—in this case, ρ. With this goal, we reformulate the Rankine–Hugoniot conditions to obtain two equations: an evolution equation for the shock curve—that is, a relation between the slope of the curve, $\eta' = d\eta/d\xi$, and the variable ρ which appears in (2.6)—and an oblique derivative boundary condition for ρ—that is, an equation linear in the gradient of ρ with coefficients depending on (ξ, η), ρ, and η'. The second equation then becomes a boundary condition for the differential equation (2.6), and we play these two conditions against each other to obtain a mapping on approximate shock positions.

We proceed to derive the jump conditions and formulate the shock evolution equation using the Rankine–Hugoniot conditions.

Writing $U \equiv (\rho, m, n)$ and $\Xi = (\xi, \eta)$, system (2.3)–(2.5) can be put in conservation form:

$$\partial_\xi F(U, \Xi) + \partial_\eta G(U, \xi) = -2U \tag{2.11}$$

with

$$F \equiv \begin{pmatrix} m - \xi\rho \\ p(\rho) - \xi m \\ -\xi n \end{pmatrix} \quad \text{and} \quad G \equiv \begin{pmatrix} n - \eta\rho \\ -\eta m \\ p(\rho) - \eta n \end{pmatrix}.$$

Inside the sonic circle $C_0 = \{\xi^2 + \eta^2 = c^2(\rho_0)\}$, the incident shock need no longer be rectilinear. The state ahead of the shock, U_1, is constant, but the state on the other side, U, is subsonic and is not uniform. The Rankine–Hugoniot conditions along the line of discontinuity $\eta = \eta(\xi)$ are, from (2.11),

$$\frac{d\eta}{d\xi} = \frac{-\eta[\rho] + [n]}{-\xi[\rho] + [m]}, \tag{2.12}$$

$$\frac{d\eta}{d\xi} = \frac{-\eta[m]}{[p] - \xi[m]}, \tag{2.13}$$

$$\frac{d\eta}{d\xi} = \frac{[p] - \eta[n]}{-\xi[n]}, \tag{2.14}$$

where $[f] = f - f_1$ denotes a jump in the state f across the shock $\eta(\xi)$. There are three families of discontinuities; two are genuinely nonlinear, and one is linear (see [5]). For nonlinear waves, $[\rho] \neq 0$. Solving for $[m]$ in (2.13) and for $[n]$ in (2.14) yields

$$[m] = \frac{-[p]\eta'}{-\eta'\xi + \eta}, \quad [n] = \frac{[p]}{-\eta'\xi + \eta}. \tag{2.15}$$

A simple consequence of (2.15) is

$$[m] = -\eta'[n]. \tag{2.16}$$

Using (2.16) in (2.13) we obtain

$$\eta = \frac{[p] - \xi[m]}{[n]}, \tag{2.17}$$

while equating the right sides of (2.12) and (2.13) and using (2.17) gives a relation

$$[p][\rho] = [m]^2 + [n]^2 \tag{2.18}$$

valid for states across a shock.

Using equations (2.15) in (2.12) we get an equation for η' involving only the state variable ρ:

$$([p] - \xi^2[\rho])(\eta')^2 + 2\xi\eta[\rho]\eta' + [p] - \eta^2[\rho] = 0. \tag{2.19}$$

To streamline the discussion, we define a function

$$s(a, b) \equiv \sqrt{\frac{(p(a) - p(b))}{(a - b)}}; \tag{2.20}$$

s is the speed of a one-dimensional shock between states with densities a and b.

PROPOSITION 2.1. *If p is a convex function of ρ, then s^2 is an increasing function of a for fixed b; $s(b, b) \equiv \lim_{a\to b} s(a, b) = c(b)$; and $s(a, b) < c(a)$ for $a > b$.*

Proof. We have

$$\frac{d}{da}s^2 = \frac{p'(a)}{a - b} - \frac{p(a) - p(b)}{(a - b)^2} = \frac{p'(a)(a - b) - (p(a) - p(b))}{(a - b)^2}.$$

Expanding $p(b) = p(a) + p'(a)(b - a) + p''(\beta)(b - a)^2/2$ for some $\beta \in (a, b)$, we obtain $ds^2/da = p''(\beta)/2 > 0$ if p is convex. As $a \to b$, $s^2 \to p'(b) = c^2(b)$ and if $a > b$,

$$c^2(a) - s^2(a, b) = \frac{p'(a)(a - b) - (p(a) - p(b))}{a - b} > 0. \qquad \square$$

For fixed b, we can write

$$a = s_b^{-1}(\eta) \quad \text{when} \quad s(a,b) = \eta. \tag{2.21}$$

Now, solving (2.19) for η' in terms of ρ and writing s^2 for $[p]/[\rho]$ yields

$$\frac{d\eta}{d\xi} = \frac{-\xi\eta \pm \sqrt{s^2(\xi^2+\eta^2-s^2)}}{s^2-\xi^2}. \tag{2.22}$$

Since the subsonic region is symmetric with respect to $\xi = 0$, we solve the problem in the half of the domain in the right half-plane, $\xi \geq 0$, and impose a zero Neumann boundary condition on $\xi = 0$. We may now specify the plus sign in (2.22) for the shock curve Σ in the first quadrant, as we anticipate (and will prove) that the shock slope is nonnegative. This gives the shock evolution equation

$$\frac{d\eta}{d\xi} = \frac{-\xi\eta + \sqrt{s^2(\xi^2+\eta^2-s^2)}}{s^2-\xi^2} = \frac{\eta^2-s^2}{\xi\eta + \sqrt{s^2(\xi^2+\eta^2-s^2)}}. \tag{2.23}$$

The second expression is equivalent to the first, and so both are well defined provided

$$s^2 \leq \xi^2 + \eta^2. \tag{2.24}$$

We will establish this condition in Proposition 2.5. We define $\Xi_s \equiv (0, \eta_s) \equiv (0, \eta(0))$, the point at the foot of the shock, and observe that we want $\eta'(0) = \sqrt{\eta^2 - s^2}/s$ to equal zero, by symmetry, and so $\eta^2 = s^2$ at Ξ_s. Thus we require

$$\eta_s = \eta(0) = s(\rho, \rho_1) = \sqrt{\frac{p(\rho) - p(\rho_1)}{\rho - \rho_1}}. \tag{2.25}$$

This can be interpreted as a condition which determines $\rho(\Xi_s)$ in the subsonic region at the base of the shock (the symmetry boundary).

We also define $\Xi_0 \equiv (\xi_0, \eta_0) = S_a^- \cap C_0$, the point where the incident shock S_a^- and the sonic circle C_0 meet. Using (2.8) for S_a^- and (2.10) for C_0 we determine Ξ_0:

$$\xi_0 = \frac{\kappa_a\sqrt{c_0^2 - s_0^2} - s_0}{\sqrt{1+\kappa_a^2}}, \quad \eta_0 = \frac{\kappa_a s_0 + \sqrt{c_0^2 - s_0^2}}{\sqrt{1+\kappa_a^2}}, \tag{2.26}$$

where $s_0^2 = (p(\rho_0) - p(\rho_1))/(\rho_0 - \rho_1)$. The initial condition for the shock position is $\eta(\xi_0) = \eta_0$.

2.3. The oblique derivative boundary condition. We next use the Rankine–Hugoniot conditions to formulate a boundary condition along the shock $\Sigma = \{(\xi, \eta(\xi))\}$.

Since vorticity is confined to the lines of discontinuity of the Riemann data (see (2.2) and [5]), and these lie below the shock (see Figure 2.1), the vorticity is zero along the shock:

$$m_\eta - n_\xi = 0. \tag{2.27}$$

Using this equation and (2.3)–(2.5), we express all the partial derivatives of m and n in terms of the derivatives of ρ:

$$n_\xi = m_\eta = \frac{1}{\xi^2+\eta^2}\Big(\eta(c^2-\xi^2)\rho_\xi + \xi(c^2-\eta^2)\rho_\eta\Big), \tag{2.28}$$

$$m_\xi = \frac{1}{\xi^2+\eta^2}\Big(\xi(c^2+\eta^2)\rho_\xi - \eta(c^2-\eta^2)\rho_\eta\Big), \tag{2.29}$$

$$n_\eta = \frac{1}{\xi^2+\eta^2}\Big(\xi(-c^2+\xi^2)\rho_\xi + \eta(c^2+\xi^2)\rho_\eta\Big). \tag{2.30}$$

Differentiating (2.18) along Σ ($' = d/d\xi = \partial_\xi + \eta'\partial_\eta$) we get

$$\begin{aligned}(c^2(\rho)[\rho] + [p])(\rho_\xi + \eta'\rho_\eta) &= 2[m]m' + 2[n]n' \\ &= 2[n](-\eta' m' + n') = 2[n](-\eta' m_\xi + (1-(\eta')^2)m_\eta + \eta' n_\eta),\end{aligned}$$

where $[m] = -\eta'[n]$ (equation (2.16)) is used in the second equality and $m_\eta = n_\xi$ (equation (2.27)) in the last equality. We simplify the last expression, replacing derivatives Dm and Dn by $D\rho$ using (2.28), (2.29), (2.30), and

$$[n] = \frac{[p]}{-\eta'\xi + \eta} \tag{2.31}$$

from (2.15), and finally we get

$$\beta \cdot \nabla\rho \equiv \beta_1\rho_\xi + \beta_2\rho_\eta = 0, \tag{2.32}$$

where β is a function of Ξ, ρ, and η' with components

$$\begin{aligned}\beta_1 = {} & (\xi^2+\eta^2)(-\eta'\xi+\eta)(c^2(\rho)+s^2(\rho,\rho_1)) \\ & - 2s^2\left\{-\eta'\xi(c^2+\eta^2) + (1-(\eta')^2)\eta(c^2-\xi^2) + \eta'\xi(-c^2+\xi^2)\right\}\end{aligned} \tag{2.33}$$

and

$$\begin{aligned}\beta_2 = {} & \eta'(\xi^2+\eta^2)(-\eta'\xi+\eta)(c^2(\rho)+s^2(\rho,\rho_1)) \\ & - 2s^2\left\{\eta'\eta(c^2-\eta^2) + (1-(\eta')^2)\xi(c^2-\eta^2) + \eta'\eta(c^2+\xi^2)\right\}.\end{aligned} \tag{2.34}$$

We now examine the obliqueness condition by comparing β with the inward normal to Ω at Σ, $\nu = (\eta', -1)$. It turns out that the operator $\beta\cdot\nabla$ in (2.32) is oblique at all points on the shock except the symmetry point. In fact, obliqueness holds along any monotonic curve which satisfies the shock equation (2.23) at (ξ_0, η_0), that is, $\eta(\xi_0) = \eta_0$ and $\eta'(\xi_0) = 1/\kappa_a$, and for any subsonic function ρ. We prove the following result.

PROPOSITION 2.2. *Let $\Sigma = \{(\xi, \eta(\xi))\}$ be any curve which has positive slope on $(0, \xi_0]$, lies inside the sonic circle C_0, and at $\xi = \xi_0$ satisfies (2.23) and $\eta = \eta_0$; let ν be its inward normal. Then for any function $\rho(\xi,\eta)$ with $c^2(\rho) > \xi^2 + \eta^2$, we have $\beta\cdot\nu > 0$ on Σ for $\xi \in (0, \xi_0]$.*

Proof. We calculate

$$\begin{aligned}\beta\cdot\nu &= \beta_1\eta' - \beta_2 \\ &= -2s^2\left\{-(\eta')^2\xi(c^2+\eta^2) + \eta'(1-(\eta')^2)\eta(c^2-\xi^2) + (\eta')^2\xi(-c^2+\xi^2)\right. \\ &\qquad \left. -\eta'\eta(c^2-\eta^2) - (1-(\eta')^2)\xi(c^2-\eta^2) - \eta'\eta(c^2+\xi^2)\right\} \\ &= 2s^2(\eta'\eta+\xi)\left\{(c^2-\xi^2)(\eta')^2 + 2\xi\eta\eta' + c^2 - \eta^2\right\}.\end{aligned}$$

Now $s^2 = [p]/[\rho] \neq 0$, since $c^2(\rho) > \xi^2+\eta^2 > c^2(\rho_1)$. Also, if $\eta' > 0$ and $\xi > 0$ we have $\eta'\eta + \xi > 0$; so to get obliqueness we need only verify that

$$(c^2-\xi^2)(\eta')^2 + 2\xi\eta\eta' + c^2 - \eta^2 > 0. \tag{2.35}$$

We first note that (2.35) holds at $\xi = \xi_0$, since $c^2(\rho_0) > s^2(\rho_0,\rho_1)$ and $(s^2-\xi^2)(\eta')^2 + 2\xi\eta\eta' + s^2 - \eta^2 = 0$ (equation (2.19)) holds at (ξ_0, η_0).

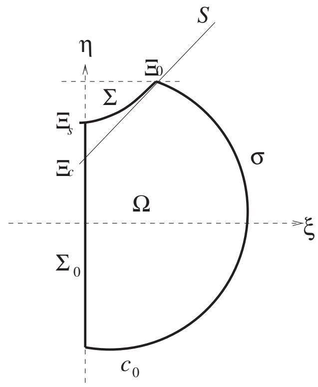


Fig. 2.2. *Sketch of the domain.*

Now, the left-hand side of (2.35) is a quadratic polynomial, $P(\eta')$, where $P(Y) = (c^2-\xi^2)Y^2 + 2\eta\xi Y + (c^2-\eta^2)$, with coefficients depending smoothly on ξ, η, and ρ. For any (ξ,η,ρ) with $\xi^2+\eta^2 < c^2(\rho)$, $P(Y)$ has a fixed sign for all Y since $\mathrm{disc}(P) = c^2\big(\xi^2+\eta^2-c^2(\rho)\big) < 0$. Thus, P has a fixed sign inside C_0. Since $P(\eta') > 0$ at (ξ_0,η_0), then $P > 0$ on $\{(\xi,\eta(\xi)) \mid \xi \in [0,\xi_0]\}$.

Thus obliqueness holds for $\xi > 0$. However, obliqueness fails at $\xi = 0$, where the factor $\eta'\eta + \xi$ vanishes because we impose the condition $\eta' = 0$. ∎

2.4. The free boundary problem. We can now give a technical statement of the main result in this paper. The subsonic domain is bounded by the part of the circle $\xi^2+\eta^2 = c^2(\rho_0)$ which lies below the shock and by the a priori unknown curved transonic shock itself. Taking advantage of the symmetry, we solve the problem in the right half of this domain, which we will call Ω in the remainder of the paper. We define σ to be the closed segment of C_0 bounding Ω and Σ_0 to be the relatively open segment of the η axis which forms the symmetry boundary. See Figure 2.2. The use of a half-domain results in a technical issue at the bottom corner, where Σ_0 meets σ, which is easily dealt with by standard continuity arguments. In addition, the fact that the upper boundary Σ is free means that Σ_0 is also not defined a priori. This matter of nomenclature we shall also ignore in the interest of simplicity.

We define Q to be the governing second-order quasi-linear operator in the domain Ω, given in (2.6) (repeated indices are summed):

$$\begin{aligned} Q\rho &= \big((c^2(\rho)-\xi^2)\rho_\xi - \xi\eta\rho_\eta\big)_\xi + \big((c^2(\rho)-\eta^2)\rho_\eta - \xi\eta\rho_\xi\big)_\eta + \xi\rho_\xi + \eta\rho_\eta \\ &\equiv D_i(a_{ij}(\Xi,\rho)D_j\rho) + b_i(\Xi)D_i\rho = 0. \end{aligned} \tag{2.36}$$

In principle, we should modify Q so that it is elliptic in Ω for any value of ρ. However, in Proposition 2.4, we immediately obtain a priori bounds which enable us to use the original operator. We denote by M the quasi-linear oblique derivative boundary operator on $\Sigma = \{(\xi,\eta(\xi)) \mid \xi \in (0,\xi_0)\}$:

$$M\rho \equiv \beta(\Xi,\rho,\eta')\cdot\nabla\rho = 0. \tag{2.37}$$

Here β is the vectorfield defined by (2.33) and (2.34). The second condition on the free boundary is the shock evolution equation (2.23) for Σ:

$$\frac{d\eta}{d\xi} = f(\Xi,\rho) \equiv \frac{-\xi\eta + \sqrt{s^2(\xi^2+\eta^2-s^2)}}{s^2-\xi^2} \quad \text{with} \quad \eta(\xi_0) = \eta_0. \tag{2.38}$$

Here $s = s(\rho(\xi, \eta(\xi)), \rho_1)$ is the function given by (2.20). On the fixed segments of the boundary, Σ_0 and σ, we impose Neumann and Dirichlet conditions, respectively:

$$\rho_\xi = 0 \quad \text{on} \quad \Sigma_0 \subset \{\xi = 0\}; \quad \rho = \rho_0 \quad \text{on} \quad \sigma \subset \{\xi^2 + \eta^2 = c^2(\rho_0)\}. \tag{2.39}$$

At the Dirichlet boundary, the equation is degenerate elliptic, in a manner described in our previous work, [1, 2, 7]. In particular, we expect that the solution will have an algebraic singularity along this boundary segment.

Now, it is easy to see that the trivial solution $\rho(\xi, \eta) \equiv \rho_0$ solves this problem, with Σ simply the straight-line extension of the incoming shock S_a^-, except at the point where the shock meets the symmetry boundary. Thus, we must in addition impose a one-point condition at this a priori unknown point, which we label Ξ_s. We impose the condition that the curved shock is smooth for the full domain problem, and hence that $\eta'(0) = 0$. As shown in section 2.2, this is equivalent to (2.25). We may alternatively express this as a one-point Dirichlet condition at the corner Ξ_s by solving $\eta_s = s(\rho(0, \eta_s), \rho_1)$, for $\rho(\Xi_s)$, or, using the notation of equation (2.21),

$$\rho(\Xi_s) = s_{\rho_1}^{-1}(\eta_s). \tag{2.40}$$

We establish the following existence theorem.

THEOREM 2.3. *There is a value κ_* such that for any Riemann data* (2.7) *with $\kappa_a > \kappa_*$, the free boundary problem consisting of* (2.36)*,* (2.37)*,* (2.38)*,* (2.39)*, and* (2.40) *has a classical solution $\rho \in C^{2+\alpha}(\Omega) \cap C(\overline{\Omega})$ which is twice continuously differentiable up to Σ and Σ_0 except at Ξ_s and Ξ_0. The free boundary is of Hölder class $H_{2+\alpha}$ for some α which is determined by the Riemann data of the problem.*

We prove this theorem using the fixed point argument we developed in our earlier papers and in work with Lieberman [3, 4, 6] for the slightly simpler small disturbance equations. The main technical difficulty which is new in this case is that the boundary condition on the free boundary is no longer uniformly oblique. To be precise, obliqueness fails at the point Ξ_s. On the other hand, because it is the nature of the Mach stem to strengthen as it approaches the wall, we find that we can control the quantity under the square root sign in (2.38). Thus our result is not restricted to being local, as in [3] and [4], or perturbative, as in [6].

We formulate the fixed point argument in terms of the position of the free boundary. We work with a regularized, uniformly elliptic, operator $Q^\varepsilon = Q + \varepsilon\Delta$ and then, as in [4], send the regularizing parameter, ε, to zero. The mapping on the free boundary is obtained by solving a fixed boundary problem using the oblique derivative condition on the shock boundary and then integrating the shock evolution equation to update the position of the shock. However, unlike our problem in [4], obliqueness fails at the corner Ξ_s representing the foot of the Mach stem. Following ideas outlined by Lieberman [20, 21], we establish local Schauder estimates at Ξ_s which are independent of the obliqueness ratio (Theorem 3.5). In section 3.2 we apply these results to the nonlinear regularized fixed boundary problem. The regularized free boundary problem is solved in section 3.3, and results for the limit $\varepsilon \to 0$ are obtained in section 4.

Before beginning the analysis, we establish that the equations above are well-defined for the approximations we use. The following monotonicity result is used throughout.

PROPOSITION 2.4. *For a given monotonic function $\eta(\xi)$ forming the boundary Σ, suppose that $\rho \in C^1(\Omega \cup \Sigma \cup \Sigma_0) \cap C(\overline{\Omega})$ is a solution of the boundary value problem* (2.36)*,* (2.37)*,* (2.39)*, and* (2.40) *with $\rho \geq \rho_0$. Then $\rho(0, \eta_s) = \rho_{\max}$ is the maximum value of ρ in $\overline{\Omega}$ and ρ is monotonic on Σ.*

Proof. Since the operator Q in (2.36) has no undifferentiated terms, the classical and Hopf maximum principles apply. That is, the local and absolute extrema of ρ occur on the boundary $\partial\Omega$ (classical); and at any point on $\partial\Omega$ where ρ has a local extremum, the normal derivative is nonzero (Hopf [15, p. 34]). On the Neumann and oblique derivative boundaries, Σ_0 and Σ, if ρ has an extremum along the boundary then two linearly independent directional derivatives of ρ are zero, and so $\nabla\rho$ is zero there, which is impossible, by the Hopf maximum principle. Thus there are no local extrema in the interior of Σ_0 or of Σ. There cannot be absolute extrema, either, and hence $\rho_{\max} = \rho(0, \eta_s)$ is the absolute maximum of ρ in Ω, and we obtain the bounds $\rho_0 < \rho < \rho_{\max}$ in Ω from the classical maximum principle. And since in Ω we have $\xi^2 + \eta^2 < c^2(\rho_0) < c^2(\rho)$, it follows that the solution is strictly subsonic in Ω.

To prove monotonicity, we argue by contradiction. Let us first examine the C^1 function ρ restricted to Σ. This is now a function of a single variable, say, the first component of a point $\Xi = (\xi, \eta)$ on Σ. Without confusion, we can label this component by the name of the point, we can order the points along Σ by this component, and we can refer to intervals along Σ by the labels. Then lack of monotonicity means there exist points Z_1 and Z_2 on Σ with $\Xi_s < Z_1 < Z_2 < \Xi_0$ at which $\rho(Z_1) < \rho(Z_2)$. We immediately deduce that

1.
2.
$$\text{in} \quad (\Xi_s, Z_2) \quad \exists \quad \widetilde{C} \quad \text{with} \quad \rho(\widetilde{C}) = \min_{[\Xi_s, Z_2]} \rho;$$
$$\text{in} \quad (\widetilde{C}, \Xi_0) \quad \exists \quad D \quad \text{with} \quad \rho(D) = \max_{[\widetilde{C}, \Xi_0]} \rho.$$

We want to identify points C and D, $C < D$, on Σ such that the following three properties hold:

(i) $\rho(\Xi_s) \geq \rho \geq \rho(C)$ on $[\Xi_s, C]$;
(ii) $\rho(C) \leq \rho \leq \rho(D)$ on $[C, D]$;
(iii) $\rho(D) \geq \rho \geq \rho(\Xi_0)$ on $[D, \Xi_0]$.

Now, property (ii) may not hold with $C = \widetilde{C}$ because $\rho(\widetilde{C})$ is the minimum value of ρ only on the interval $[\Xi_s, Z_2]$, and we may have $D > Z_2$. So, if there is a point in (Z_2, D) at which $\rho < \rho(\widetilde{C})$, then we let C be a point at which ρ has its minimum value in this interval; if there is no such point, then let $C = \widetilde{C}$. Then all three properties hold.

Now we look at the function ρ in the domain Ω. The idea is to partition Ω into subdomains by two curves Γ_C and Γ_D from C and D, respectively, to points A and B, respectively, on Σ_0, in such a way that $\rho(A) < \rho(B)$ and so that we can deduce that there is a point m on Σ_0 at which ρ reaches a minimum on either the domain Ω_A or the domain Ω_B, thus violating the Hopf maximum principle, as stated in the first paragraph of this proof. See Figure 2.3. It is of course sufficient to show that $\rho(m)$ is the minimum value of ρ on the boundary of Ω_A or Ω_B.

It would be simplest to find curves on which ρ is monotonic, but it is not clear that such curves exist, or what properties they would have. Instead, we construct Lipschitz curves on which ρ is monotonic on average. To be precise, we construct curves on which, for a certain number μ,

$$\rho(A) \leq \rho \leq \rho(C) + \mu \quad \text{on } \Gamma_C \quad \text{and } \rho(A) < \rho(C); \tag{2.41}$$
$$\rho(B) \geq \rho \geq \rho(D) - \mu \quad \text{on } \Gamma_D \quad \text{and } \rho(B) > \rho(D). \tag{2.42}$$

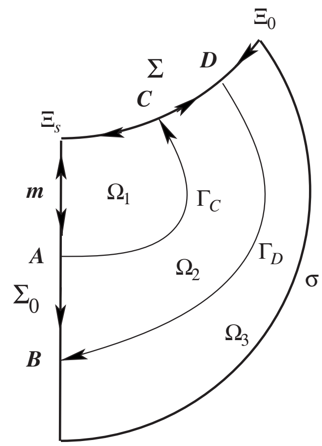


FIG. 2.3. *Illustration of the proof of Proposition* 2.4.

We begin by identifying some useful constants. Let

$$\mu = \frac{1}{4}\min\{\rho(D) - \rho(C), \rho(\Xi_s) - \rho(D), \rho(C) - \rho(\Xi_0)\}.$$

Since $\rho \in C(\overline{\Omega})$, then ρ is uniformly continuous, and there is an $\epsilon > 0$ such that $\rho(\Xi) \leq \rho_0 + \mu$ if $\operatorname{dist}(\Xi, \sigma) < \epsilon$. Let $\Omega_\epsilon = \{\Xi \in \Omega \mid \operatorname{dist}(\Xi, \sigma) > \epsilon\}$, and let $\sigma_\epsilon = \{\Xi \in \Omega \mid \operatorname{dist}(\Xi, \sigma) = \epsilon\}$. As we shall see, we can restrict our attention to $\overline{\Omega_\epsilon}$. The purpose of constructing this domain is to be able to bound $|\nabla\rho|$. Since $\rho \in C^1(\overline{\Omega_\epsilon})$, we have $|\nabla\rho| \leq M$ there, say. (We could estimate M from Schauder theory, but this is not important here.)

Now, on any ball of radius r, the oscillation of ρ is bounded by $2Mr$, and we now choose a radius, $R = \mu/(2M)$, so that

$$\operatorname*{osc}_{B_R \cap \Omega_\epsilon} \rho \leq \mu.$$

Now we construct Γ_D as follows. Consider a ball $B_R(D)$ centered at D. In $B_R(D) \cap \Omega_\epsilon$, $\rho(D)$ cannot be the maximum value of ρ (because D is not a point of local maximum in Ω); hence there are points of $\partial B_R(D) \cap \Omega_\epsilon$ where $\rho > \rho(D)$. Let X_1 be a point at which ρ attains its maximum value in $\overline{B_R(D)}$. The first segment of Γ_D is a straight line from D to X_1. We have $\rho(X_1) > \rho(D)$, and on the segment, $\rho(X) \geq \rho(D) - \mu$ and $\rho(X) < \rho(X_1)$.

Now we continue inductively, forming a sequence of line segments with corners at $\{X_i\}$ (take $D = X_0$), along which $\rho \geq \rho(D) - \mu$ and such that $\rho(X_1) < \rho(X_2) < \cdots$. To show that we can do this, let

$$\Omega_j = \Omega_\epsilon \backslash \overline{\{\cup_0^{j-1} B_R(X_i)\}};$$

we have $X_j \in \partial\Omega_j$, and we consider $B_R(X_j)$. We note that $\rho(X_j)$ is the largest value of ρ on the part of $B_R(X_j)$ inside the complement of Ω_j. However, $\rho(X_j)$ is less than the maximum value of ρ on $B_R(X_j)$, by the mean value property. Hence there is a point $X_{j+1} \in \partial B_R(X_j) \cap \Omega_j$ at which ρ attains its maximum value in $\overline{B_R(X_j)}$. Again, along the straight line from X_j to X_{j+1} we have $\rho \geq \rho(X_j) - \mu > \rho(D) - \mu$.

Now,

$$\operatorname{dist}(X_{i-1}, \Omega_i) = R$$

and

$$\Omega_j \subset \Omega_{j-1} \subset \cdots \subset \Omega_1,$$

so

$$\operatorname{dist}(X_i, \Omega_k) \geq R$$

for $k \geq i + 1$; and since $X_k \in \partial\Omega_k$, the estimate

$$\operatorname{dist}(X_{j+1}, X_i) \geq R \quad \forall \quad i \leq j$$

follows.

Hence $\operatorname{dist}(X_i, X_j) \geq R$ for $i \neq j$ for all points in the sequence. But only a finite number of balls with radius R and separated centers will fit in Ω_ϵ, so this process must terminate after a finite number of steps when we reach a point $X_L = B \in \partial\Omega_\epsilon$. By construction, Γ_D has the properties indicated in (2.42).

Similarly, we construct Γ_C, with termination point $A \in \partial\Omega_\epsilon$.

Next we show that the points A and B lie on Σ_0. First, the curves cannot cross each other, because at every point on Γ_D, $\rho \geq \rho(D) - \mu > \rho(C) + \mu$, while at every point on Γ_C we have $\rho < \rho(C) + \mu$. Also, Γ_D cannot terminate at σ_ϵ where $\rho \leq \rho_0 + \mu < \rho(D)$. For the same reason, B cannot lie on Σ in the segments $[D, \Xi_0]$ or $[C, D]$ where $\rho \leq \rho(D)$. Finally, B cannot lie in the segment $[\Xi_s, C]$ of Σ because this would trap Γ_C in a region where $\rho \geq \rho(C)$ (or, more simply, this would contradict the fact that C is not a local minimum in Ω). Hence $B \in \Sigma_0$.

Similarly, A cannot lie on Σ, where $\rho \geq \rho(C)$ in the interval $[\Xi_s, D]$, and must lie on Σ_0, between B and Ξ_s.

Now we find our final contradiction. Since there is a point, A, in the interval $[\Xi_s, B]$ of Σ_0 where ρ is smaller than its value at either endpoint, then there must be a point m where ρ, restricted to the interval $[\Xi_s, B]$ of Σ_0 attains its minimum. We recall that m cannot be a local minimum in Ω, and so it cannot be a minimum in Ω_1 or in Ω_2. The relevant domain is Ω_1 if $m \in [\Xi_s, A]$; otherwise it is Ω_2. In particular, there would have to be points on the boundary of the relevant domain at which $\rho < \rho(m)$. But the construction we have performed prevents this. To verify this, suppose first that $m \in [\Xi_s, A]$. Then $\rho \geq \rho(m)$ on $[\Xi_s, A]$. In particular, $\rho(m) \leq \rho(A) \leq \rho(X)$ for $X \in \Gamma_C$, and $\rho(m) \leq \rho(A) < \rho(C)$, by (2.41). In addition, $\rho \geq \rho(C)$ on the top boundary, $[\Xi_s, C]$ in Σ, of Ω_1. Thus, we have a contradiction to the maximum principle if $m \in [\Xi_s, A]$.

But if $m \in [A, B]$, then again there are no points on the interval $[A, B]$ of Σ_0 at which $\rho < \rho(m)$, and again $\rho \geq \rho(A) \geq \rho(m)$ along Γ_C. As before, $\rho(C) > \rho(A) \geq \rho(m)$. Now, $\rho \geq \rho(C)$ on the interval $[C, D]$ of the top boundary, Σ, of Ω_2, and by (2.42) we have $\rho \geq \rho(D) - \mu > \rho(C)$ along Γ_D. Thus in this case also, $\rho(m)$ is the smallest value of ρ along the entire boundary of Ω_2. This again contradicts the maximum principle, as stated in the first paragraph of the proof.

We conclude that C and D do not exist, and hence that Z_1 and Z_2 do not exist, and ρ is monotonic on Σ. $\square$

As a second basic result, we prove that the shock evolution equation can always be integrated, defining the mapping whose fixed point is the free boundary. Beginning with a given curve $\eta(\xi)$, assume we have solved the fixed boundary value problem (2.36), (2.37), (2.39), and (2.40). We then produce a new approximate shock position $\tilde{\eta}(\xi)$ by integrating (2.38):

$$\tilde{\eta}(\xi) = \eta_0 + \int_{\xi_0}^{\xi} f(x, \eta(x), \rho(x, \eta(x)))\, dx, \tag{2.43}$$

where f is defined in (2.38). Note that on the right side of (2.43) we evaluate all quantities along the old shock position, $\eta(\xi)$. We have the following proposition.

PROPOSITION 2.5. *Suppose that η is a monotone function and that ρ satisfies the boundary value problem* (2.36), (2.37), (2.39), *and* (2.40). *Then $\eta^2 > s^2$ and $\eta^2 + \xi^2 > s^2$ for all $\xi \in (0, \xi_0)$ so the new curve $\tilde{\eta}$ is defined for all $\xi \in [0, \xi_0]$ and is monotonic. Furthermore, $\tilde{\eta}'(0) = 0$.*

Proof. Because ρ satisfies (2.40), we see that at $\xi = 0$ the quantity under the square root sign in (2.38) is zero. Since η is monotonic, the quantity $\eta^2(\xi)$ is an increasing function of ξ. We use Proposition 2.4 to conclude that s^2 along Σ is a decreasing function of ξ (since ρ decreases and s is a monotonic function of ρ). Hence $\eta^2 - s^2$ is strictly positive when $\xi > 0$. In addition, this implies that $\xi^2 + \eta^2 - s^2$ is positive, and so the right-hand side of (2.43) is well defined (see the equivalent form in (2.23)). In addition, (2.23) also shows that $d\tilde{\eta}/d\xi$ is positive as long as $\eta^2 - s^2 > 0$. Finally, this derivative is zero at $\xi = 0$, where the right side of (2.38) vanishes. ∎

We now define $\mathcal{K} = \mathcal{K}^{\varepsilon}$, a closed, convex subset of a Hölder space $H_{1+\alpha_1}([0, \xi_0])$; the value of $\alpha_1 \in (0, 1)$ depends on the regularizing parameter ε and will be specified later. The functions in $\mathcal{K}$ satisfy

(K1) $\eta(\xi_0) = \eta_0$, and $\eta'(\xi_0) = 1/\kappa_a$, where ξ_0 and η_0 are defined in (2.26);
(K2) $\eta'(0) = 0$;
(K3) $\eta_c \le \eta(\xi) \le \eta_0$; recall that $\eta_c = \sqrt{1 + \kappa_a^2}\, s_0/\kappa_a < \eta_0 < c_0$ if $\kappa_a > \kappa_A$;
(K4) $0 \le \eta' \le \sqrt{c_0^2/s_0^2 - 1}$.

Then (2.43) defines a mapping on $\mathcal{K}$:

$$J : \eta \mapsto \tilde{\eta}. \tag{2.44}$$

The upper bound in **(K4)** is justified by the following proposition.

PROPOSITION 2.6. *If $\eta(\xi)$ is a monotonic function with $\eta(\xi_0) = \eta_0$ and ρ a solution to* (2.36), (2.37), (2.39), *and* (2.40), *then the function f given by* (2.38) *is bounded above by $\sqrt{c_0^2 - s_0^2}/s_0 \equiv 1/\kappa_A$.*

Proof. By Proposition 2.4, $s(\xi, \eta)$ is a decreasing function on $\eta(\xi)$ with $s^2(0, \eta(0)) = \eta^2(0)$, and by Proposition 2.5, $\eta \ge s$ on $\eta(\xi)$. For the function f defined by (2.38), a calculation shows

$$\frac{\partial f}{\partial \xi} = -\frac{(\eta^2 - s^2)\big(s\xi + \eta\sqrt{\xi^2 + \eta^2 - s^2}\big)}{\sqrt{\xi^2 + \eta^2 - s^2}\big(\xi\eta + \sqrt{s^2(\xi^2 + \eta^2 - s^2)}\big)^2} < 0,$$

$$\frac{\partial f}{\partial \eta} = \frac{\xi^2 + \eta^2}{\sqrt{\xi^2 + \eta^2 - s^2}\big(s\eta + \xi\sqrt{\xi^2 + \eta^2 - s^2}\big)} > 0,$$

$$\frac{\partial f}{\partial s^2} = -\frac{\frac{1}{2}\eta^2(\xi^2 + \eta^2 - s^2) + \frac{1}{2}s^2\xi^2 + \xi\eta s\sqrt{\xi^2 + \eta^2 - s^2}}{\big(\xi\eta + \sqrt{s^2(\xi^2 + \eta^2 - s^2)}\big)^2} < 0,$$

Hence, f is largest when η has its maximum value η_0, and ξ and s their minimum values, 0 and s_0, respectively. This gives the stated upper bound, which is the reciprocal of the limiting value κ_A, as calculated in [17]. ∎

We also note the upper bound for the solution ρ of (2.36), (2.37), (2.39), and (2.40) when $\eta \in \mathcal{K}$. Since $\eta_s \leq \eta_0$ and s^2 is monotonic, for given Riemann data $(\rho_0, \rho_1, \kappa_a)$, the value of $\rho_{\max}$ in Proposition 2.4 is bounded above by ρ_M, where, from (2.40),

$$\rho_M = s_{\rho_1}^{-1}(\eta_0). \tag{2.45}$$

We will use this upper bound in the proofs.

We prove Theorem 2.3 in two stages. First, in section 3 we solve the regularized free boundary value problem for $Q^\varepsilon = Q + \varepsilon\Delta$. In section 4, we consider the limit $\varepsilon \to 0$ and show that this limit yields a solution of (2.36)–(2.40).

3. The regularized problem. For a fixed $\varepsilon \in (0,1)$ we solve the free boundary problem defined at the beginning of section 2.4, but with Q replaced by the regularized operator Q^ε. The equation for ρ in the subsonic region is now

$$Q^\varepsilon \rho = Q\rho + \varepsilon\Delta\rho = 0; \tag{3.1}$$

the shock evolution equation remains the same,

$$\eta' = f(\xi, \eta, \rho), \quad \eta(\xi_0) = \eta_0; \tag{3.2}$$

and the boundary conditions are, as before,

$$M\rho = \beta \cdot \nabla\rho = 0 \quad \text{on} \quad \Sigma \equiv \{(\xi, \eta(\xi)); 0 < \xi < \xi_0\}, \tag{3.3}$$

$$\rho = \rho_0 \quad \text{on} \quad \sigma; \quad \rho_\xi = 0 \quad \text{on} \quad \Sigma_0, \tag{3.4}$$

and

$$\rho(\Xi_s) = \rho_s \equiv s_{\rho_1}^{-1}(\eta_s). \tag{3.5}$$

The theorem we prove in this section is as follows. (See (3.7) for the spaces.)

THEOREM 3.1. *For each $\varepsilon \in (0,1)$, there exists a solution $(\rho^\varepsilon, \eta^\varepsilon) \in H_{1+\alpha}^{(-\gamma)}(\Omega^\varepsilon) \times H_{1+\alpha}([0, \xi_0])$ to the regularized free boundary problem* (3.1), (3.2), (3.3), (3.4), *and* (3.5) *such that*

$$\rho_0 < \rho^\varepsilon \leq \rho_s \leq \rho_M \quad \textit{and} \quad c^2(\rho^\varepsilon) > \xi^2 + \eta^2 \quad \textit{in} \quad \overline{\Omega^\varepsilon} \setminus \sigma. \tag{3.6}$$

Here, $\alpha, \gamma \in (0,1)$ both depend on ε and on the Riemann data κ_a, ρ_0, and ρ_1. The curve $\eta^\varepsilon(\xi)$, defining the position of the free boundary Σ^ε, is in $\mathcal{K}^\varepsilon$; Ω^ε is bounded by σ, Σ_0, and Σ^ε.

We prove Theorem 3.1 in the following steps (which take up the three subsections of this section).

Step 1. First we show the existence of a solution to a linear problem with fixed boundary Σ defined by $\eta(\xi) \in \mathcal{K}$ and establish Hölder and Schauder estimates at Σ. For this, it is convenient to define a weighted Hölder space; see [15] for the general definition of weighted Hölder spaces. Let $V = \{\Xi_0\}$ denote the corner point at which Σ meets the degenerate boundary σ. Set $\Omega' = \Omega \cup \sigma \cup \Sigma_0 \setminus V$. We anticipate loss of regularity at V, because of the mixed boundary condition and the degeneracy of the operator Q at σ. At Ξ_s, we also find loss of regularity because of loss of obliqueness of the operator M. The third corner, between Σ_0 and σ, is an artifact of our decision to work in a half-domain. Since it does not contribute to any loss of regularity, we ignore it in the discussion. We define the corner region near Ξ_0:

$$\Omega_V(\delta) = \{x \in \Omega : \mathrm{dist}(V, x) \leq \delta\}.$$

In [3, 4, 6], in which the derivative condition was uniformly oblique, the only loss of regularity came from the corners. In the present problem, we overcome the loss of obliqueness at a single point on Σ, but at a cost: the Schauder estimates are no longer independent of the gradients of the coefficients, and hence we do not get a compact mapping in the same spaces. In this paper, we therefore modify the weighted Hölder spaces, as follows. We define a region which is close to Σ but does not contain the corner Ξ_0 by taking a covering of Σ with balls of radius δ centered at points on Σ which are bounded away from Ξ_0. Define $\Sigma''(\delta) = \{\Xi \in \Sigma \mid \mathrm{dist}(\Xi, \Xi_0) > \delta\}$ and

$$\Sigma(\delta) = \left\{ x \in \Omega \cap \bigcup_{\Xi \in \Sigma''(\delta)} B_\delta(\Xi) \right\},$$

where $B_\delta(\Xi)$ is a ball of radius δ centered at Ξ. We then define

$$H_a^{(b)} \equiv \left\{ \|u\|_a^{(b)} \equiv \sup_{\delta > 0} \delta^{a+b} |u|_{a, \overline{\Omega} \setminus \{\Sigma(\delta) \cup \Omega_V(\delta)\}} < \infty \right\}. \tag{3.7}$$

For the linear problem, we establish a priori Schauder and Hölder bounds at Σ, in particular near the point where the data lose obliqueness; we use Hölder estimates near V, and $C_{2+\alpha}$ estimates locally in the rest of the domain. We prove existence of a solution by regularizing the oblique boundary condition to be uniformly oblique, then passing to the limit using the a priori bounds.

Step 2. Using the Hölder gradient bounds to the linear problem, we establish existence results for the nonlinear fixed boundary problem, via the Schauder fixed point theorem.

Step 3. We apply the Schauder fixed point theorem again to prove existence of a solution to the nonlinear free boundary problem.

3.1. The regularized linear fixed boundary problem. Replace ρ in the coefficients a_{ij} of (2.36) and β_i of (2.33), (2.34) by a function w in a set $\mathcal{W}$ defined with respect to a given boundary component Σ, and depending on given values Ξ_s and ρ_s (see (3.5)), as follows.

DEFINITION 3.2. *The elements of* $\mathcal{W} \subset H_2^{(-\gamma_1)}$ *satisfy*

(W1) $\rho_0 \le w \le \rho_M$, $w = \rho_0$ *on* σ, $w(\Xi_s) = \rho_s$, $w_\xi = 0$ *on* Σ_0;

(W2) $\|w\|_2^{(-\gamma_1)} \le K$;

(W3) $|w|_{\alpha_0, \Omega'_{loc}} \le K_0$.

The weighted Sobolev space is defined by (3.7); the values of $\gamma_1, \alpha_0 \in (0,1)$ will be specified following (3.29), as will the values of K and K_0. The set $\mathcal{W}$ is clearly closed, bounded, and convex.

The quasilinear equations (3.1) and (3.3) are now replaced by linear partial differential and boundary equations (repeated indices are summed)

$$\begin{aligned} L^\varepsilon u &= D_i(a_{ij}(\Xi, w) D_j u) + \varepsilon \Delta u + b_i(\Xi) D_i u = 0 \quad \text{in} \quad \Omega, \\ N u &= \beta_i D_i u = \beta_i(\Xi, w) D_i u = 0 \quad \text{on} \quad \Sigma = \{\eta = \eta(\xi)\}, \end{aligned} \tag{3.8}$$

with a given $\eta \in \mathcal{K} \subset H_{1+\alpha_1}$ and $w \in \mathcal{W}$. Because of the bound (W1), L^ε is uniformly elliptic in Ω with ellipticity ratio depending on the Riemann data and on ε. In this section, we demonstrate the key point that for a given function $w \in \mathcal{W}$, the solution u to the linear equations (3.8) with the remaining boundary conditions

$$u = \rho_0 \text{ on } \sigma, \; u_\xi = 0 \text{ on } \Sigma_0 \text{ and } u(\Xi_s) = \rho_s, \tag{3.9}$$

satisfies Hölder and Schauder estimates in Ω' and a uniform $H_{1+p,\Sigma(d_0)}$ bound near Σ for any $p < \min\{\gamma_1, \alpha_1\}$. This bound gives rise to enough compactness to establish the existence of a solution to the quasilinear problem by applying the Schauder fixed point theorem.

We first note L^∞ a priori bounds for the solution u to the linear problem.

PROPOSITION 3.3. *The solution u to the linear problem* (3.8), (3.9) *satisfies*

$$\rho_0 < u \le \rho_s \le \rho_M \quad in \quad \Omega \cup \Sigma \cup \Sigma_0, \tag{3.10}$$

where $\rho_s = \rho(0, \eta_s)$ is defined in (3.5) *and ρ_M, defined in* (2.45), *is independent of ε. Moreover,*

$$c^2(u) > c^2(\rho_0) > \xi^2 + \eta^2 \quad in \quad \Omega \cup \Sigma \cup \Sigma_0. \tag{3.11}$$

Proof. The linear problem is uniformly elliptic for $\varepsilon > 0$ and $w \in \mathcal{W}$, so the classical maximum principle applies, as well as the boundary considerations used in the proof of Proposition 2.4. ▯

Next, we state the Schauder estimates including the Dirichlet and fixed Neumann boundaries, σ and Σ_0, and the Hölder estimates at the corner, Ξ_0.

THEOREM 3.4. *Assume that Σ is given by $\{(\xi, \eta(\xi))\}$ with $\eta \in \mathcal{K}$ for some $\alpha_1 \in (0,1)$ and that w is in $\mathcal{W}$ for given K, K_0, α_0, and γ_1. Then there exist $\gamma_V, \alpha_\Omega \in (0,1)$ such that any solution $u \in H_{2+\alpha_\Omega,\Omega'} \cap H_{\gamma_V,\Omega_V(d_0)}$ to the linear problem* (3.8), (3.9) *satisfies*

$$|u|_{\gamma,\Omega_V(d_0)} \le C_1 |u|_0 \tag{3.12}$$

for any $\gamma \le \gamma_V$ and

$$|u|_{2+\alpha,\Omega'_{loc}} \le C_2 |u|_0 \tag{3.13}$$

for any $\alpha \le \alpha_\Omega$. The exponent γ_V depends on the Riemann data, and both α_Ω and γ_V depend on ε but are independent of α_1 and γ_1. The constant C_1 is independent of the bounds K and K_0. The constant C_2 is independent of K but depends on K_0.

Proof. The proof is immediate. We refer to Theorem 1 of Lieberman [22] for the corner estimate. Here γ_V depends on the angle between Σ and σ at V, a fixed value that depends only on the Riemann data, and on the obliqueness ratio at V, which is also fixed, as well as on the ellipticity ratio ε, but not on γ_1, α_1, K, or K_0.

Standard interior and boundary Schauder estimates, for example, [15, p. 98], give the local estimate (3.13). The constant C_2 depends on ε, on the H_α norm of the coefficients a_{ij}, and on the domain. ▯

Because interior Schauder estimates can be applied once more, a solution in $H_{2+\alpha,\Omega'}$ is actually in $C^3(\Omega)$.

Finally, we establish Hölder gradient estimates at Σ. It is at this point that we need to derive basic estimates at the point Ξ_s where the boundary operator N is not oblique. To avoid discussing the Neumann boundary separately at each step of this proof, we reflect Ω across the ξ axis, without further comment; Ω includes Σ_0 and we let Σ stand for the full $H_{1+\alpha_1}$ boundary in Theorem 3.5. The remaining assumptions are the same as in Theorem 3.4.

THEOREM 3.5. *Assume that Σ is given by $\{(\xi, \eta(\xi))\}$ with $\eta \in \mathcal{K}$ for some $\alpha_1 \in (0,1)$ and that w is in $\mathcal{W}$ for given K, K_0, α_0, and γ_1. Then, there exists a* *positive constant d_0 such that for every $d \leq d_0$, any solution $u \in C^1(\Omega \cup \Sigma) \cup C^3(\Omega)$ to the linear problem* (3.8), (3.9) *satisfies*

$$|u|_{1+p,\Sigma(d)} \leq C(\varepsilon, \alpha_1, \gamma_1, K, d_0)|u|_0 \tag{3.14}$$

for any $p < \min\{\gamma_1, \alpha_1\}$.

Proof. Away from a neighborhood $B_{d_0}(\Xi_s)$ of Ξ_s the boundary operator N in (3.8) is oblique and thus we can apply known regularity theory, for example, [15, Theorem 6.30], to get (3.14) in $\Sigma(d_0) \setminus B_{d_0}(\Xi_s)$, with a constant C which depends on ε, α_1, Ω, d_0, and K_0. Hence we consider only estimates near Ξ_s in the remainder of the proof.

For a given solution u to (3.8) and (3.9) we define

$$v = \frac{u}{1 + |Du|_0} \quad \text{and} \quad z = Nv = \beta_i(\Xi)D_i v. \tag{3.15}$$

We construct a barrier function f for z on $B \equiv B_{d_0}(\Xi_s) \cap \overline{\Omega}$ to get a Hölder estimate for the gradient of the solution of (3.8), (3.9). Let $\psi = z + f(\zeta)$, where ζ is the regularized distance function (from the boundary component Σ); see [18]. A smooth approximation to $d(\Xi) = \operatorname{dist}(\Sigma, \Xi)$ is necessary since Σ has minimal regularity. The regularized distance function has the properties $1 \leq \zeta/d \leq 2$, $0 < \zeta_0 \leq |D\zeta| \leq \zeta_D$ and $|D^2\zeta| \leq \zeta_D d^{\alpha_1 - 1}$. We let $f(0) = 0$ and we first construct the lower barrier, $-f$, by finding a suitable positive, increasing function f such that $\psi > 0$. Note that, with f positive, we get $\psi \geq z$ on ∂B. Where no confusion is likely, we let subscripts denote partial derivatives and calculate

$$D_i\psi = \beta_j D_{ij} v + D_i\beta_j D_j v + f'\zeta_i, \tag{3.16}$$

whence

$$\beta_j D_{ij} v = D_i\psi - (D_i\beta_j D_j v + f'\zeta_i). \tag{3.17}$$

We also have

$$D_{ij}\psi = \beta_k D_{ijk} v + D_j\beta_k D_{ik} v + D_i\beta_k D_{jk} v + D_{ij}\beta_k D_k v + f'\zeta_{ij} + f''\zeta_i\zeta_j. \tag{3.18}$$

In addition, since w satisfies (W2) with a given constant K, we get estimates on the derivatives of a_{ij}. Using the definition of the weighted norms, we have (noting $|Dw| \leq |w|_1$ and so on)

$$\begin{aligned} |D(a_{ij})| &\leq |a_{ij,x}| + |a_{ij,u}||Dw| \leq |a_{ij,x}| + |a_{ij,u}|\|w\|_1^{(-\gamma_1)} d^{\gamma_1 - 1} \leq m d^{\gamma_1 - 1}, \\ |D^2(a_{ij})| &\leq |a_{ij,x,x}| + 2|a_{ij,x,u}||Dw| + |a_{ij,u,u}||Dw|^2 + |a_{ij,u}||D^2 w| \\ &\leq |a_{ij,x,x}| + 2|a_{ij,x,u}|\|w\|_1^{(-\gamma_1)} d^{\gamma_1 - 1} + |a_{ij,u,u}|(\|w\|_1^{(-\gamma_1)} d^{\gamma_1 - 1})^2 \\ &\quad + |a_{ij,u}|\|w\|_2^{(-\gamma_1)} d^{\gamma_1 - 2} \\ &\leq m(d^{\gamma_1 - 2} + d^{2\gamma_1 - 2}). \end{aligned} \tag{3.19}$$

Here subscripts denote derivatives of a_{ij} with respect to the variables in Ξ $(, x)$ and with respect to w $(, u)$. The symbol $m = m(K)$ denotes a quantity which depends on the structure of the derivatives of a_{ij} and the bound K on w. We absorb terms which are less singular as $d \to 0$. We also get estimates on the derivatives of β_i. Let $\gamma_2 = \min\{\gamma_1, \alpha_1\}$. Then

$$|D\beta_i| \le md^{\gamma_2 - 1}, \qquad |D^2\beta_i| \le m(d^{\gamma_2-2} + d^{2\gamma_2-2}), \tag{3.20}$$

where $m = m(K) > 0$ depends on the structure of the derivatives of β. In deriving this estimate, we use the fact that η', η'', and η''' are bounded by d^{α_1}, d^{α_1-1}, and d^{α_1-2}, respectively, as we can apply Lemma 2.8 of [14] to $\eta(\xi) - \eta$.

Since $\beta_2(\Xi_s, w) = 0$ and $\beta_1(\Xi_s, w) \neq 0$, we can take $0 < d_1 \le 1$ small enough so that for all $0 < d_0 \le d_1$ and all $w \in \mathcal{W}$ we have $\beta_1(\Xi, w) \neq 0$ in B_{d_0}. Now we solve the two equations in (3.17) along with $Lv = 0$, that is,

$$a_{ij}D_{ij}v = -(D_j a_{ij} D_i v + b_i D_i v), \tag{3.21}$$

as a linear system for the three derivatives $D_{ij}v$. The assumption that β_1 is bounded away from zero, coupled with the ellipticity of L, gives a uniform bound $c_1(\Lambda, \lambda, |\beta|_0)$ on the inverse of the coefficient matrix of the linear system. Here we may let Λ and λ be the eigenvalues of (a_{ij}) restricted to B. These are order one constants which depend only on the Riemann data. Furthermore, we can estimate the right-hand sides of (3.17) and (3.21) using (3.19) and (3.20). We get

$$|D^2 v| \le c_1(\Lambda, \lambda, |\beta|_0)\big(|D\psi| + (md^{\gamma_2-1} + |b|_0)|Dv| + f'\zeta_D\big). \tag{3.22}$$

This bounds the second derivatives of v in terms of $|D\psi|$. Now we proceed to obtain bounds for ψ. The idea is to find an elliptic operator for which ψ is a subsolution in B and simultaneously to force $\psi > 0$ on ∂B, by choice of the function f. A second-order operator for ψ involves third derivatives of v, so we estimate these. By using $Lv = 0$, (3.22), (3.19), and (3.20) (recall that $|Dv| \le 1$), we get

$$\begin{aligned}
a_{ij}D_{ijk}v &= -\big(D_k a_{ij} D_{ij}v + D_j a_{ij} D_{ik}v + b_i D_{ik}v + D_{jk}a_{ij}D_i v + D_k b_i D_i v\big)\\
&\le (md^{\gamma_2-1} + |b|_0)|D^2 v| + (md^{\gamma_2-2} + md^{2\gamma_2-2} + |b|_1)|Dv|\\
&\le c_1(md^{\gamma_2-1} + |b|_0)|D\psi| + c_1(md^{\gamma_2-1} + |b|_0)^2\\
&\quad + c_1(md^{\gamma_2-1} + |b|_0)f'\zeta_D + md^{\gamma_2-2} + md^{2\gamma_2-2} + |b|_1\\
&\le c_2\big\{(md^{\gamma_2-1} + 1)|D\psi| + (md^{\gamma_2-1} + 1)^2\\
&\quad + (md^{\gamma_2-1} + 1)f' + md^{\gamma_2-2} + md^{2\gamma_2-2} + 1\big\},
\end{aligned}$$

where $c_2 = c_2(\Lambda, \lambda, \rho_M, |\beta|_0, |b|_0, |b|_1, \zeta_D)$. Thus, using (3.18) and making the estimates indicated, we have

$$\begin{aligned}
a_{ij}D_{ij}\psi &\le c_2|\beta|_0\big\{(md^{\gamma_2-1} + 1)|D\psi| + (md^{\gamma_2-1} + 1)^2 + (md^{\gamma_2-1} + 1)f'\\
&\qquad + md^{\gamma_2-2} + md^{2\gamma_2-2} + 1\big\}\\
&\qquad + 2\Lambda md^{\gamma_2-1}c_1\big\{|D\psi| + md^{\gamma_2-1} + |b|_0 + f'\zeta_D\big\}\\
&\qquad + \Lambda m(d^{\gamma_2-2} + d^{2\gamma_2-2}) + \Lambda f'|\zeta_{ij}| + f'' a_{ij}\zeta_i\zeta_j\\
&\le c_3\big\{(md^{\gamma_2-1} + 1)|D\psi| + md^{\gamma_2-2} + (m^2 + m)d^{2\gamma_2-2}\\
&\qquad + md^{\gamma_2-1}(1 + f')\big\} + \Lambda f'|\zeta_{ij}| + f'' a_{ij}\zeta_i\zeta_j.
\end{aligned}$$

Here c_3 is a constant depending on the same parameters as c_1 and c_2, and terms which are bounded as $d \to 0$ have again been omitted. Now we define

$$L_1\psi \equiv a_{ij}D_{ij}\psi - c_3(md^{\gamma_2-1} + 1)|D\psi|$$

and we calculate

(3.23)
$$L_1\psi \le c_3\big\{md^{\gamma_2-2} + (m^2+m)d^{2\gamma_2-2} + md^{\gamma_2-1}(1+f')\big\} + \Lambda\zeta_D f' d^{\alpha_1-1} + \lambda f''\zeta_0^2.$$

To obtain this estimate, we have assumed $f'' < 0$ and estimated

$$f'' a_{ij}\zeta_i\zeta_j \le f'' \min a_{ij}\zeta_i\zeta_j = f''\lambda|D\zeta|^2 \le f''\lambda\zeta_0^2.$$

We have also used the property of regularized distance: $|\zeta_{ij}| \le \zeta_D d^{\alpha_1-1}$. We now specify $f(\zeta) = f_0\zeta^p$ for any $p < \gamma_2$, so that

$$f'' = f_0 p(p-1)\zeta^{p-2} \le f_0 p(p-1)d^{p-2} < 0,$$

and $f' d^{\alpha_1-1} \le 2^{p-1} f_0 p d^{p+\alpha_1-2}$. Finally, we choose f_0 big enough and $d_2 \in (0,1)$ small enough to get $L_1\psi < 0$ in B_{d_0} for every $d_0 \le d_2$. We now define $d_0 \equiv \min\{d_1, d_2\}$.

Additionally, since (3.14) holds near Σ, away from Ξ_s, and hence is valid on ∂B, we can choose f_0 larger if necessary so that $\psi > 0$ on ∂B. Therefore, by the maximum principle, $\psi > 0$ in B. Thus, $z > -f$ in B.

Similarly, f is an upper barrier for z. We now have an estimate for z. In addition we have, since $\psi = z + f$,

$$|\psi| \le c_4(m^2+1)d^p \quad \text{for} \quad d \le d_0.$$

Since $\psi = 0$ on Σ, we can use Schauder estimates, applying [15, Lemma 6.20] or [14, Lemma 7.1, Theorem 7.2], using the fact that ψ and $-\psi$ are upper and lower solutions of an operator L_1 with a Dirichlet boundary condition and estimating the right side of (3.23), to obtain

$$\|\psi\|^{(-p)}_{2+\gamma_2} \le C_1\big(\sup d^{-p}|\psi| + |\psi|_0 + |\psi|_{p,\partial B}\big) \le c_4(m^2+1) + c(m) = C(m).$$

The constant C_1 depends only on λ and Λ (the ellipticity constants in B) and on γ_2. To obtain the second inequality in this expression, we have used the fact that $|\psi|_{p,\partial B}$ is bounded, with a bound which depends only on $|\psi|_0$ and on Λ/λ. This follows from $\psi = 0$ on Σ and from interior Schauder estimates for v, a solution to a linear problem, on $\partial B \cap \Omega$. Finally, this leads to

$$|D\psi| \le \|D\psi\|^{(1-p)}_{\gamma_2+1} d^{p-1} \le C(m)d^{p-1} \quad \text{for} \quad d < d_0. \tag{3.24}$$

We now use (3.24) in (3.22) and drop lower-order terms to get

$$|D^2 v| \le c_1(|D\psi| + md^{\gamma_2-1} + f') \le c_1(C(m)d^{p-1} + md^{\gamma_2-1} + f_0 p d^{p-1}) \le Cd^{p-1}.$$

Now Hölder estimates on Dv follow by integrating the last inequality. More precisely, $|D^2 v| \le Cd^{p-1}$ implies that $\|Dv\|^{(-p)}_1 \le C$, and by [14, Lemma 2.1] we have

$$|Dv|_p = \|Dv\|^{(-p)}_p \le C(p)\|Dv\|^{(-p)}_1,$$

and therefore we get

$$|v|_{1+p} \le C. \tag{3.25}$$

Finally, using the definition of v in (3.15), we apply the interpolation inequality, [15, Lemma 6.32], with a small $\delta > 0$ to get

$$|u|_{1+p} \leq C(1 + |Du|_0) \leq C(1 + \delta|u|_{1+p} + C_\delta|u|_0) \tag{3.26}$$

and thus (3.14) holds. Therefore we get Hölder gradient estimates at Σ for the solution u of (3.8). ∎

Now we can establish existence of a solution to (3.8) and (3.9).

THEOREM 3.6. *Assume that* Σ *is given by* $\{(\xi, \eta(\xi))\}$ *with* $\eta \in \mathcal{K}$ *for some* $\alpha_1 \in (0,1)$ *and that* w *is in* $\mathcal{W}$ *for given* K, K_0, α_0, *and* γ_1. *Then there exist* $\gamma_V, \alpha_\Omega \in (0,1)$, *and* $d_0 > 0$, *where* γ_V, α_Ω, *and* d_0 *are independent of* γ_1 *and* α_1, *such that a solution* $u \in H_{1+p,\Sigma(d)} \cap H_{2+\alpha,\Omega'} \cap H_{\gamma,\Omega_V(d_0)}$ *for the linear problem* (3.8) *and* (3.9) *exists for any* $\alpha \leq \alpha_\Omega$, $p < \min\{\gamma_1, \alpha_1\}$, $\gamma \leq \gamma_V$, *and* $d \leq d_0$ *and satisfies* (3.12), (3.13), *and* (3.14).

Proof. To show the existence of a solution u to (3.8) and (3.9), we approximate the oblique derivative boundary condition on Σ. To be precise, noting that the unit inward normal to Σ at Ξ_s is $(0,-1)$, for $0 < \delta < 1$ we let $\beta_\delta = \beta + (0, -\delta)$ so that $\beta_\delta \cdot \nu = \beta \cdot \nu + \delta \geq \delta > 0$ at Ξ_s. Then, for sufficiently small δ, β_δ is uniformly oblique. The boundary condition is now discontinuous at the corner Ξ_s, where Σ and Σ_0 meet. Results from [21] and [19] imply that there exists a solution u^δ to $Lu^\delta = 0$ in Ω, $\beta_\delta \cdot \nabla u^\delta = 0$ on Σ, and (3.9). Now we apply Theorems 3.4 and 3.5, which are independent of δ, to see that the sequence u^δ is uniformly bounded in $H_{1+p,\Sigma(d_0)} \cap H_{2+\alpha_\Omega,\Omega'} \cap H_{\gamma_V,\Omega_V(d_0)}$ for any $p < \min\{\gamma_1, \alpha_1\}$. Thus by the Arzela–Ascoli theorem, there exists a subsequence converging uniformly to a function u. Using the uniform bounds (3.12), (3.13), and (3.14), we conclude that the limiting function solves the problem (3.8), (3.9). ∎

3.2. The regularized nonlinear fixed boundary problem. This subsection is devoted to proving the existence of solutions to the nonlinear problem (3.1) with a fixed boundary. We again assume that an approximate shock boundary Σ is given by a function $\eta = \eta(\xi) \in \mathcal{K}$. We also are given the value $\rho_s = s_{\rho_1}^{-1}(\eta(0))$. We prove the following theorem.

THEOREM 3.7. *For each* $\varepsilon \in (0,1)$, *and for given* $\eta \in \mathcal{K} \subset H_{1+\alpha_1}$, *there exists a solution* $\rho^\varepsilon \in H_{2+\alpha}^{(-\gamma)}(\Omega^\varepsilon)$ *to* (3.1), (3.3), (3.4), *and* (3.5) *such that*

$$\rho_0 < \rho^\varepsilon \leq \rho_s \leq \rho_M, \quad \textit{and} \quad c^2(\rho^\varepsilon) > \xi^2 + \eta^2 \quad \textit{in} \quad \overline{\Omega}^\varepsilon \setminus \sigma \tag{3.27}$$

for some $\alpha(\varepsilon), \gamma(\varepsilon) \in (0,1)$. *Moreover, for some* $d_0 > 0$ *the solution* ρ^ε *satisfies*

$$|\rho^\varepsilon|_{\gamma,\Sigma(d_0)\cup\Omega_V(d_0)} \leq K_1, \tag{3.28}$$

where γ *and* K_1 *depend on* ε, γ_V *and* K *but both are independent of* α_1.

Proof. We suppress the dependence on ε to simplify the notation.

Recall that $\mathcal{K} \subset H_{1+\alpha_1}([0,\xi_0])$ is a closed convex set of functions satisfying the additional conditions (K1) to (K4) given in section 2.4. For any function w in $\mathcal{W}$ we define a mapping $T : \mathcal{W} \subset H_2^{(-\gamma_1)} \to H_2^{(-\gamma_1)}$ by letting $\rho = Tw$ be the solution to the linear regularized fixed boundary problem, (3.8), (3.9) solved in Theorem 3.6. Because w satisfies (W1), L^ε is strictly elliptic, with ellipticity ratio depending on ε. By Theorem 3.6, T maps $\mathcal{W} \subset H_2^{(-\gamma_1)}$ to a bounded set in $H_{2+\alpha}^{(-\gamma_V)}$, where γ_V is the value given by Theorem 3.6. Since γ_V is independent of γ_1, we may take $\gamma_1 = \gamma_V/2$ and then $T(\mathcal{W})$ is precompact in $H_2^{(-\gamma_1)}$.

To show T maps $\mathcal{W}$ into itself, we need to show that Tw satisfies (W1), (W2), and (W3). Now, (W1) is immediate by Proposition 3.3 and the boundary conditions.

By applying interior and boundary Hölder estimates (see [15, Theorems 8.22 and 8.27]), we get the local estimate

$$|\rho|_{\alpha*,\Omega_1'} \leq C_0, \tag{3.29}$$

where $0 < \alpha* < 1$ and C_0 depend only on ε (the ellipticity ratio), the Riemann data, and on $d' = \operatorname{dist}(\Omega_1', \partial\Omega')$ with $\Omega_1' \subset \Omega'$. Notice that, as in the remark following Theorem 8.24 in [15, p. 202], the constant C_0 is nondecreasing and the constant $\alpha*$ nonincreasing with respect to d'. Since $\Omega' \subset \Omega$ is bounded, we can find an upper bound for C_0 and a lower bound for $\alpha*$ depending only on the size of Ω and the ellipticity ratio. Thus, if we define $\mathcal{W}$ with $K_0 = C_0$ and $\alpha_0 = \alpha*$, with C_0 the upper bound and $\alpha*$ the lower bound, then $\rho = Tw$ satisfies (W3). Note that K_0 and $\alpha*$ are independent of α_1 and γ_V.

To verify (W2), we need to find a value K such that

$$\sup_{\delta>0} \delta^{2-\gamma_1} |\rho|_{2,\overline{\Omega}\setminus\{\Sigma(\delta)\cup\Omega_V(\delta)\}} < K, \tag{3.30}$$

assuming $\|w\|_2^{-\gamma_1} \leq K$. We start by noting that Theorem 3.5 implies the existence of a positive constant $d_0 > 0$ such that for every $d \leq d_0$, any solution $u \in C^1(\Omega\cup\Sigma)\cup C^3(\Omega)$ to the linear problem (3.8), (3.9) satisfies the Hölder gradient estimate (3.14), where the constant C depends on K but is uniform in $d \leq d_0$. Based on this estimate, we get a local bound for the weighted norm of ρ on $\Sigma(d_0)$ of the form

$$d^{2-\gamma_1}|\rho|_2 \leq d^{1-\gamma_1+p} C \tag{3.31}$$

which holds for all $d < d_0$. Here C depends on K, α_1, and γ_1. To show (3.30) we estimate the supremum by considering separately domains $\overline{\Omega} \setminus \{\Sigma(\delta) \cup \Omega_V(\delta)\}$ for which $\delta > \tilde{d}$, where $\tilde{d} \leq d_0$ will be specified later, and domains for which $\delta \leq \tilde{d}$.

In domains of the first kind, $\overline{\Omega}\setminus\{\Sigma(\delta)\cup\Omega_V(\delta)\}$ with $\delta > \tilde{d}$, the solution is smooth and its C^2-norm bound is independent of K. More precisely, we can use the uniform Hölder estimate (3.29) and bootstrap iteratively (see [15, Theorem 6.6]) to get the local Schauder estimate

$$|\rho|_{2+\alpha_\Omega,\Omega'} \leq C(K_0). \tag{3.32}$$

Notice that since the Hölder estimate (3.29) is independent of the distance between Ω_1' and the boundary Σ, so is the Schauder estimate (3.32). The interpolation inequality [15, Lemma 6.32] gives

$$|\rho|_{2,\Omega'} \leq c|\rho|_0 + \mu|\rho|_{2+\alpha,\Omega'} \leq c\rho_M + \mu C(K_0) \tag{3.33}$$

for any $\mu > 0$ and $c = c(\mu)$. We fix $\mu = 1$ and get

$$\sup_{\delta>\tilde{d}} \delta^{2-\gamma_1} |\rho|_{2,\overline{\Omega}\setminus\{\Sigma(\delta)\cup\Omega_V(\delta)\}} \leq K', \tag{3.34}$$

where K' depends on the size of the domain Ω, on $C(K_0)$, and on ρ_M but is independent of the distance to Σ.

Next we study $\delta^{2-\gamma_1}|\rho|_{2,\overline{\Omega}\setminus\{\Sigma(\delta)\cup\Omega_V(\delta)\}}$ when $\delta \leq \tilde{d}$. We divide the subdomain $\overline{\Omega} \setminus \{\Sigma(\delta) \cup \Omega_V(\delta)\}$ into two: the part for which $\delta > \tilde{d}$ and the complement. Then the upper bound over the subdomain $\overline{\Omega} \setminus \{\Sigma(\delta) \cup \Omega_V(\delta)\}$ is equal to the larger of the suprema over the two subdomains. The supremum over the subdomain for which $\delta > \tilde{d}$ has been calculated above. The supremum over the complement is calculated using the estimates for the behavior of the solution near Σ, namely, estimate (3.31) and the corner estimate (3.12). In (3.12), the constants C_1 and γ_V are independent of K, K_0, and α_1, while $|\rho|_0$ is bounded by ρ_M from Proposition 3.10. By the interpolation inequality [14, Lemma 2.1], since $\gamma_1 = \gamma_V/2$ we have

$$|\rho|_{\gamma_1,\Omega_V(d_V)} \le C_V|\rho|_{\gamma_V,\Omega_V(d_V)} \le C_V C_1 \rho_M, \tag{3.35}$$

where $C_V = C_V(\gamma_1, \gamma_V, \Omega_V(d_V))$, for some $d_V > 0$. From here we get that

$$d^{2-\gamma_1}|\rho|_2 \le K_V \quad \forall d < d_V,$$

where K_V is independent of K. Hence we can take $K \equiv \max\{K_V, K'\}$, using the bound (3.34), and now K is independent of α_1 and of $\tilde{d}$. Since K_V and K' are independent of $\tilde{d}$ we can change $\tilde{d}$ without affecting K. Therefore, we can choose $\tilde{d} \le \min\{d_0, d_V\}/2$ in (3.31) small enough that $\tilde{d}^{1-\gamma_1+p}C < K$. Therefore, (3.30) is satisfied and we have chosen parameters K, K_0, and α_0 defining $\mathcal{W}$ so that T maps $\mathcal{W}$ into itself.

Now, by the Schauder fixed point theorem, there exists a fixed point ρ such that $T\rho = \rho \in H_2^{(-\gamma_1)}$. Thus, ρ solves (3.1), (3.3), (3.4), and (3.5). By a bootstrap argument we get $\rho \in H_{2+\alpha}^{(-\gamma_1)}$ for any $\alpha \le \alpha_\Omega$, the value given in Theorem 3.6. For reference, we note that we have chosen $\gamma_1 = \gamma_V/2$; the exponent $\gamma_V \in (0,1)$ depends on the corner angle at Ξ_0 and α_Ω and γ_V depend on ε. The bounds on ρ in Proposition 3.3 give the first estimate in (3.27), and the second follows.

Finally, since $T(\mathcal{W}) \subset \mathcal{W}$ is a bounded set in $H_2^{(-\gamma_1)}$, then by (W2) and by the interpolation inequality [14, Lemma 2.1], any fixed point ρ satisfies (3.28) for any $\gamma \le \gamma_1 = \gamma_V/2$. Note that K_1 and γ_1 are independent of α_1. ◻

3.3. The regularized nonlinear free boundary problem. We now prove existence of a solution to the regularized free boundary problem.

Proof of Theorem 3.1. Again, we suppress the ε dependence.

For each $\eta \in \mathcal{K} \subset H_{1+\alpha_1}$, using the solution ρ of the nonlinear fixed boundary problem (3.1), (3.3), (3.4), and (3.5) given by Theorem 3.7, we define the map J on $\mathcal{K}$, $\tilde{\eta} = J\eta$ as in (2.44), by integrating (2.43):

$$\tilde{\eta}(\xi) = \eta_0 + \int_{\xi_0}^{\xi} f(x, \eta(x), \rho(x, \eta(x)))\, dx. \tag{3.36}$$

First, we check that J maps $\mathcal{K}$ into itself. Property (K1) follows from (3.36). By Proposition 2.5, property (K2) holds, while the upper and lower bounds in (K4) hold by Proposition 2.6 and in turn imply (K3).

The Hölder class of ρ at Σ is given by the estimate (3.28), along with a bound on the Hölder γ-norm, and from estimate (3.28) in the proof of Theorem 3.7 we saw that we could choose $\gamma = \gamma_V/2$. Evaluating $f(\Xi, \rho(\Xi))$, we get a bound $|f|_{\gamma_V/2} \le C(K_1)$, and thus $|\tilde{\eta}|_{1+\gamma_V/2} \le C(K_1)$. The constants here are simple functions of the Riemann data and the structure of the pressure function. The important feature of the mapping is that γ_V is independent of α_1, the Hölder exponent of the space $\mathcal{K}$. Thus, we have $J(\mathcal{K}) \subset H_{1+\gamma_V/2}$; since properties (K1)–(K4) hold, we then have $J(\mathcal{K}) \subset \mathcal{K}$ if

$\alpha_1 \le \gamma_V/2$. Furthermore, J is compact if $\alpha_1 < \gamma_V/2$. We now take $\alpha_1 = \gamma_V/3$. By standard arguments, the map J is continuous.

Therefore, J has a fixed point $\eta^\varepsilon \in H_{1+\gamma_V/3}([0,\xi_0])$ by the Schauder fixed point theorem. This gives a curve Σ^ε on which (3.2) holds. Together with the corresponding solution ρ^ε from Theorem 3.7, this establishes the existence of a solution $(\rho^\varepsilon, \eta^\varepsilon) \in H_{2+\alpha}^{(-\gamma)} \times H_{1+\alpha}$ of the regularized free boundary problem (3.1), (3.2), (3.3), (3.4), and (3.5) for sufficiently small $\gamma(\varepsilon)$ and $\alpha(\varepsilon)$.

This completes the proof of Theorem 3.1. ∎

4. The limiting solution. In this section we study the limiting solution, as the elliptic regularization parameter ε tends to zero. We start with the regularized solutions of (3.1), (3.2), (3.3), (3.4), and (3.5), whose existence is guaranteed by Theorem 3.1. Denote by ρ^ε a sequence of regularized solutions of the partial differential equation.

PROPOSITION 4.1. *For each ε the constant function ρ_0 is a lower barrier for ρ^ε and $c^2(\rho_0) > \xi^2 + \eta^2$ in $\overline{\Omega^\varepsilon} \setminus \sigma$.*

Proof. For each ε we have $\rho^\varepsilon > \rho_0$, and by the monotonicity of c^2 we get $c^2(\rho^\varepsilon) > c^2(\rho_0) > \xi^2 + \eta^2$ in $\Omega^\varepsilon \cup \Sigma_0$. The same inequality holds on Σ^ε since $(\xi, \eta^\varepsilon(\xi))$ lies inside C_0. Thus $c^2(\rho_0) > \xi^2 + \eta(\xi)^2$ for $\xi \in [0, \xi_0)$ and ρ_0 is a uniform lower barrier. ∎

The existence of a uniform lower bound ρ_0 in ε allows us to apply standard local compactness arguments (see, for example, [3, Lemma 4.2]) to get a limit ρ, locally, in the interior of the domain. Here, the issue is ensuring ellipticity uniformly in ε in compact subsets of Ω. We first show that the sequence of domains Ω^ε converges to a domain Ω, as $\varepsilon \to 0$.

LEMMA 4.2. *The sequence η^ε has a convergent subsequence, whose limit η belongs to $C^\gamma([0,\xi_0])$ for all $\gamma \in (0,1)$. The limiting curve η is convex.*

Proof. Theorem 3.1 gives the existence of a sequence $(\rho^\varepsilon, \eta^\varepsilon)$ of solutions of the regularized free boundary problems for which η^ε belongs to the set $\mathcal{K}^\varepsilon$ for each ε. Now, $\rho_0 < \rho^\varepsilon \le \rho_s^\varepsilon \le \rho_M$, where ρ_M is independent of ε, and the property (K4) of $\mathcal{K}^\varepsilon$, specified in section 2.4, immediately gives a C^1 bound on η^ε, uniformly in ε. Thus by the Arzela–Ascoli theorem, η^ε has a convergent subsequence, and the limit $\eta \in C^\gamma([0,\xi_0])$ for all $\gamma \in (0,1)$.

To see that η is convex we first show that η^ε is convex for each $\varepsilon > 0$. Recall that $\eta' = f(\xi, \eta(\xi), \rho(\xi, \eta(\xi)))$ and calculate $\eta'' = f_\xi + f_\eta \eta' + f_\rho \rho'$. By observing that if ρ were constant the shock would be a straight line, we get $f_\xi + f_\eta \eta' = 0$. Therefore, the sign of η'' is determined entirely by the sign of f_ρ and ρ'. Since ρ is decreasing by Proposition 2.4, this implies $\rho' \le 0$. Furthermore, by Lemma 2.1 we have $d(s^2)/d\rho \ge 0$ and by the proof of Proposition 2.6 we have $f_{s^2} < 0$, so $f_\rho = f_{s^2}(s^2)' \le 0$. This shows that η^ε is convex for each $\varepsilon > 0$, and so the limiting function is convex. ∎

The limit value $\eta(0) = \lim \eta^\varepsilon(0)$ is also established, and the corresponding subsequence of domains Ω^ε also has a limit, Ω.

In the remaining lemmas, without further comment, we carry out the limiting argument using the convergent subsequence of η^ε, which we again call η^ε.

LEMMA 4.3. *The sequence ρ^ε has a limit $\rho \in C^{2+\alpha'}(\Omega)$ for some $\alpha' > 0$. The limit ρ satisfies the quasi-linear degenerate elliptic equation* (2.36). *In addition, $\rho_0 < \rho < \rho_M$ in Ω.*

Proof. The proof is based on local compactness arguments and on uniform L^∞ bounds for ρ^ε: $\rho_0 < \rho^\varepsilon < \rho_s^\varepsilon \le \rho_M$, where ρ_M is independent of ε. The main ideas follow those used in [13, Theorem 1] and the proof is almost identical to the proof of [4, Lemma 4.2]. We omit the details. ∎

In the next lemma, we show that the limiting functions ρ and η satisfy both the shock evolution equation (2.38) and the oblique derivative boundary condition (2.37), $M\rho = 0$, on Σ.

LEMMA 4.4. *The limits η and ρ satisfy*

$$\eta' = f(\xi, \eta, \rho) \quad \textit{and} \quad M\rho = \beta(\eta(\xi), \rho) \cdot \nabla\rho = 0 \quad \textit{on} \quad \Sigma. \tag{4.1}$$

Furthermore, $\eta \in C^{2+\alpha'}(0, \xi_0) \cap C^1([0, \xi_0))$ and $\rho \in C^{2+\alpha'}(\Omega \cup \Sigma \cup \Sigma_0 \setminus \Xi_s) \cap C(\Omega \cup \Sigma \cup \Sigma_0)$ for some $\alpha' > 0$. In addition, ρ satisfies $\rho = \rho_s$ at $\Xi_s = (0, \eta(0))$, where $\rho_s = s_{\rho_1}^{-1}(\eta(0))$.

Proof. The proof is similar to that of [4, Lemma 4.3] except for the loss of uniform obliqueness at Ξ_s. We omit the local arguments away from Ξ_s and concentrate on dealing with the behavior of the solution near Ξ_s.

The arguments presented in the proof of [4, Lemma 4.3] imply $\eta^\varepsilon(\xi) \to \eta(\xi)$ in $C_{loc}^{2+\alpha'}$ for $\xi \neq 0$, and since the subsequence ρ^ε converges to ρ in $C_{loc}^{1+\alpha'}$, we get

$$(\eta^\varepsilon)' = f(\xi, \eta^\varepsilon, \rho^\varepsilon) \to f(\xi, \eta, \rho) \quad \forall \xi \neq 0,$$

thus $\eta' = f(\xi, \eta, \rho)$ for $\xi \neq 0$. Furthermore, by continuity of β and ρ we have

$$0 = \beta(\eta^\varepsilon, \rho^\varepsilon) \cdot \nabla\rho^\varepsilon(\xi, \eta^\varepsilon(\xi)) \to \beta(\eta, \rho) \cdot \nabla\rho(\xi, \eta(\xi)) \quad \forall \xi \neq 0,$$

and thus $\beta(\eta, \rho) \cdot \nabla\rho = 0$ on $\Sigma \setminus \{(0, \eta(0))\}$.

We now focus on the behavior of the solution at Ξ_s. By Lemma 4.2 we have $\eta^\varepsilon \to \eta$ in $C^\gamma([0, \xi_0])$ for any $0 < \gamma < 1$. Furthermore, by construction, for each $\varepsilon > 0$,

$$s^2(\rho_s^\varepsilon, \rho_1) = (\eta^\varepsilon(0))^2.$$

Therefore, as $\varepsilon \to 0$, the right-hand side converges to $\eta^2(0)$; hence $s^2(\rho_s^\varepsilon, \rho_1) \to \eta^2(0)$. By continuity and monotonicity of s^2 this implies that the sequence of numbers ρ_s^ε also has a limit, R. Moreover, $s^2(R, \rho_1) = \eta^2(0)$. But, this equation defines ρ_s; therefore $R = \rho_s$ and we have shown that the sequence of traces of the functions ρ^ε evaluated at $(0, \eta^\varepsilon(0))$ converges to ρ_s. We still have to show that ρ is continuous at Ξ_s, that is, that $\lim_{\xi\to 0} \rho(\xi, \eta(\xi)) = \rho_s$.

Since η_ε' has a limit $\eta' = f(\xi, \eta(\xi), \rho(\xi, \eta(\xi)))$ in $C^{1+\alpha}$ for $\xi \neq 0$, and since for each $\varepsilon > 0$ we have $\eta_\varepsilon'(0) = 0$, then for any $\delta > 0$ there exists an $h_0 \neq 0$ such that

$$|\eta'(h)| \leq |\eta'(h) - \eta_\varepsilon'(h)| + |\eta_\varepsilon'(h)| \leq \delta$$

for $0 < h < h_0$, which implies continuity of η' at $\xi = 0$ and $\eta'(0) = 0$. Thus

$$f(h, \eta(h), \rho(h, \eta(h))) = \eta'(h) \to \eta'(0) = 0 = f(0, \eta(0), \rho_s) \text{ as } h \to 0.$$

This implies, among other things, that $\rho(h, \eta(h)) \to \rho_s$ and so ρ is continuous at Ξ_s, $\rho(\Xi_s) = \rho_s$ and the boundary condition (2.40) is satisfied. ∎

The final task is to prove continuity of ρ up to the degenerate boundary σ. It is here that we need an additional condition on the Riemann data.

LEMMA 4.5. *For Riemann data satisfying a bound $\kappa_a > \kappa_*(\rho_1, \rho_0)$, the limit ρ satisfies $\rho = \rho_0$ on σ and $\rho \in C(\overline{\Omega})$.*

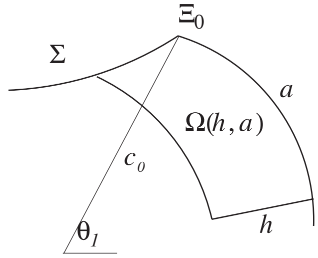


FIG. 4.1. *A sketch of the corner barrier domain.*

Proof. Continuity of solutions of $Q\rho = 0$ up to a degenerate boundary was proved as Corollary 3.3 in [7], at points where the degenerate boundary σ is convex, when the problem satisfies a Dirichlet condition on the entire boundary, and the entire boundary is degenerate. In [7], a pointwise upper barrier function ψ was constructed, uniformly in ε, with $\psi > \rho^\varepsilon$ in Ω and $\psi = \rho^\varepsilon$ at $\Xi \in \sigma$. This proof can easily be adapted to give a local barrier at every interior point of σ in our problem. Thus, to show continuity everywhere on σ we need only to show continuity at Ξ_0. We construct an upper barrier ψ with $\psi(\Xi_0) = \rho_0$ so that $\psi \geq \rho^\varepsilon$ in a fixed set $\Omega(h, a)$ (see Figure 4.1) for all $\varepsilon > 0$. Since ρ_0 is a lower barrier, we then have continuity at Ξ_0.

It is convenient to work in polar coordinates $(\xi, \eta) = (r\cos\theta, r\sin\theta)$. In this coordinate system, the operator Q^ε becomes

$$Q^\varepsilon\rho = \left(c^2(\rho) - r^2 + \varepsilon\right)\rho_{rr} + \frac{c^2}{r^2}\rho_{\theta\theta} + p''(\rho)\left(\rho_r^2 + \frac{1}{r^2}\rho_\theta^2\right) + \left(\frac{c^2}{r} - 2r\right)\rho_r.$$

To compare ψ and ρ^ε we introduce an operator $Q_1^\varepsilon(\rho^\varepsilon)$ which is partially linearized:

$$Q_1^\varepsilon(\rho^\varepsilon)u = \left(c^2(u) - r^2 + \varepsilon\right)u_{rr} + \frac{c^2(\rho^\varepsilon)}{r^2}u_{\theta\theta} + p''(\rho^\varepsilon)\left(u_r^2 + \frac{1}{r^2}u_\theta^2\right) + \left(\frac{c^2(\rho^\varepsilon)}{r} - 2r\right)u_r.$$

The barrier function has the form

$$\psi(r,\theta) = \rho_0 + A(c_0 - r)^b + B(\theta_1 - \theta)^2. \tag{4.2}$$

Here θ_1 is the angle subtended by Ξ_0; A and B are constants to be determined and the exponent b is a value, also to be determined, in $(0, 1)$. The barrier is constructed on a curvilinear quadrilateral, $c_0 \geq r \geq c_0 - h$, $\theta_1 - a \leq \theta \leq \theta^\varepsilon(r)$, where $\theta^\varepsilon(r)$ is the boundary Σ^ε in polar coordinates and h and a are small numbers to be determined. The use of a barrier function with a singular derivative is motivated by [7], following [13]. In fact, we conjecture that the solution to the equation, in this case, does have a square root singularity at C_0 and that our value of b, which can be refined a posteriori to be any number less than $1/2$, is optimal.

Before evaluating $Q_1^\varepsilon\psi$, we write $c^2 = p'$ and expand $c^2 - r^2$ as $c^2(\psi) - r^2 = c^2(\psi) - c_0^2 + c_0^2 - r^2 = (\psi - \rho_0)p''(\overline{\rho}) + c_0^2 - r^2$, where $\overline{\rho}$ is a value in the range of ρ^ε. By assumption, p'' is bounded above and below by positive numbers for $\rho \in [\rho_0, \rho_M]$.

We have

$$
\begin{aligned}
&Q_1^\varepsilon(\rho^\varepsilon)\psi \\
&= \left(p''(\overline{\rho})\left\{A(c_0-r)^b + B(\theta_1-\theta)^2\right\} + (c_0+r)(c_0-r)+\varepsilon\right) b(b-1)A(c_0-r)^{b-2} \\
&\qquad + \frac{c^2(\rho^\varepsilon)}{r^2}(2B) + p''(\rho^\varepsilon)\left\{\left(Ab(c_0-r)^{b-1}\right)^2 + \frac{1}{r^2}\left(2B(\theta_1-\theta)\right)^2\right\} \\
&\qquad + \left(\frac{c^2(\rho^\varepsilon)}{r} - 2r\right) Ab(c_0-r)^{b-1}.
\end{aligned}
$$

The coefficient of $(c_0-r)^{b-2}$, the most singular term as $r \to c_0$, is

$$p''(\overline{\rho})(B(\theta_1-\theta)^2+\varepsilon)b(b-1)A < 0.$$

The next most singular power is $(c_0-r)^{2b-2}$, and its coefficient is

$$A^2 b\left(p''(\overline{\rho})(b-1) + p''(\rho)b\right) \le A^2 k_0 < 0 \qquad \text{if} \qquad b < \frac{\min p''}{2\max p''}, \tag{4.3}$$

which we now assume. The next power is $(c_0-r)^{b-1}$, whose coefficient is a bounded multiple of A; the remaining terms are bounded and involve only powers of B. Once we have fixed the lower limit, c_0-h, for r, and have chosen B, we can then choose A, which appears quadratically in (4.3) with a negative coefficient, large enough to make the entire expression negative. This is sufficient to guarantee that $\psi-\rho^\varepsilon$ does not have a negative minimum in the interior of $\Omega(h,a)$ provided that $\psi-\rho^\varepsilon$ is nonnegative on the boundary of $\Omega(h,a)$. For at a negative interior minimum, $\nabla\psi = \nabla\rho^\varepsilon$, and

$$
\begin{aligned}
0 &\ge Q_1^\varepsilon(\rho^\varepsilon)\psi - Q^\varepsilon(\rho^\varepsilon) \\
&> \left(c^2(\psi)-c^2(\rho^\varepsilon)\right)\psi_{rr} + \left(c^2(\rho^\varepsilon)-r^2+\varepsilon\right)(\psi-\rho)_{rr} + \frac{c^2(\rho^\varepsilon)}{r^2}(\psi-\rho)_{\theta\theta}.
\end{aligned} \tag{4.4}
$$

However, $\psi < \rho^\varepsilon$ implies $c^2(\psi)-c^2(\rho) < 0$, while $\psi_{rr} < 0$ by the concavity of ψ in r; in addition $(\psi-\rho)_{rr}$ and $(\psi-\rho)_{\theta\theta}$ are nonnegative at the minimum, so the sum of the three terms is positive. This contradiction establishes the conclusion that if $\psi \ge \rho^\varepsilon$ on the boundary of $\Omega(h,a)$, then ψ is an upper barrier for each ρ^ε.

We now turn to establishing bounds for ψ on the sides of the quadrilateral. First, on σ: $\rho^\varepsilon = \rho_0 < \psi$. We fix an angular interval by choosing some $a > 0$; then we can choose B large enough that $Ba^2 > \rho_M$. This gives $\psi > \rho^\varepsilon$ on the boundary $\theta = \theta_1 - a$ of $\Omega(h,a)$.

The appropriate condition on the oblique derivative boundary is more delicate. We linearize the boundary condition, obtain an estimate of the form $N_1(\rho^\varepsilon)\psi \le 0$, and use the Hopf maximum principle to show that $\psi-\rho^\varepsilon$ is positive on Σ^ε. Getting the estimate $N_1(\rho^\varepsilon)\psi \le 0$ is rendered difficult by the fact that the part of $\nabla\psi$ which becomes singular near Ξ_0 is not the normal derivative (over which we have some control because the problem is oblique near Ξ_0) but the derivative in the direction r.

We can obtain the bound we need, at least as long as κ_a is large enough. To see this, we compute the derivative of ψ along Σ^ε, using the linearized operator $N_1(\rho^\varepsilon) = \beta(\rho^\varepsilon)\cdot\nabla$. To focus on the singular part, we write $\beta(\rho^\varepsilon)\cdot\nabla$ in terms of its radial and angular components,

$$N_1\psi = \beta^r\psi_r + \beta^\theta\psi_\theta,$$

where

$$\beta^r = \beta \cdot (\cos\theta, \sin\theta) = \frac{1}{r}(\beta_1\xi + \beta_2\eta),$$

and we have an analogous expression for β^θ. Now a calculation gives

$$\beta_1\xi + \beta_2\eta = (\eta - \eta'\xi)(\xi + \eta'\eta)\big(r^2(c^2 + 3s^2) - 4c^2s^2\big). \tag{4.5}$$

The first two factors are uniformly positive near Ξ_0, and if ρ_M is sufficiently close to ρ_0, then we claim there exists an interval $[c_0 - h, c_0]$ in which $\beta_1\xi + \beta_2\eta$ has a positive lower bound, for there will be a value $h > 0$ such that the expression in (4.5) is positive for $r > c_0 - h$ as long as $4c^2s^2/(c^2 + 3s^2) < c_0^2$ for all values in the range of ρ. Since the left side of this expression is monotone increasing in ρ, it is sufficient to impose the restriction on $c^2(\rho_M)$ and $s^2(\rho_M)$. The condition obviously holds for $\rho = \rho_0$, and so it certainly holds for ρ_M sufficiently close to ρ_0. Furthermore, for large κ_a, $\rho_M = \rho_0 + \mathcal{O}(1/\kappa_a)$, by a calculation given in [17]. That is, for κ_a large enough we have $\beta_1\xi + \beta_2\eta \geq C > 0$ in (4.5). Estimates on κ_* are given in [17].

We now complete the calculation of

$$N_1(\rho^\varepsilon)\psi = -\beta^r A(c_0 - r)^{b-1} - 2\beta^\theta B(\theta_1 - \theta)$$

by choosing A large enough that $N_1\psi \leq 0$ on Σ^ε. We also ensure $\psi - \rho^\varepsilon > 0$ at $r = c_0 - h$, by increasing A again if necessary, so that $Ah^b > \rho_M$.

Finally, we confirm that the inequality $N_1(\rho^\varepsilon)\psi < 0$ precludes negative values of $\psi - \rho^\varepsilon$ on Σ^ε. If there are negative values, then there is a negative minimum, at which the tangential derivative of $\psi - \rho^\varepsilon$ vanishes, so we have

$$0 \geq N_1(\psi - \rho^\varepsilon) = \beta^t(\psi - \rho^\varepsilon)_t + \beta^n(\psi - \rho^\varepsilon)_n = \beta^n(\psi - \rho^\varepsilon)_n,$$

where the superscripts mark the tangential and (inward) normal components of β, and the subscripts the derivatives of $\psi - \rho^\varepsilon$. Since $\beta^n > 0$, this implies that $(\psi - \rho^\varepsilon)_n \leq 0$. However, we can write $\overline{L}(\psi - \rho^\varepsilon) \leq 0$ at such a point for a suitable linear operator $\overline{L}$, and thus the Hopf maximum principle requires that $(\psi - \rho^\varepsilon)_n > 0$, a contradiction. Thus we conclude that $\psi - \rho^\varepsilon \geq 0$ on the entire boundary of $\Omega(h, a)$. By the argument following the inequality (4.4), this establishes ψ as an upper barrier. We note that this construction depends on ε only through the location of the curve $\Sigma = \Sigma^\varepsilon$ and that A, B, b, h, and a are independent of ε. Thus, since the domains Ω^ε converge, it follows that ψ is a barrier for all ρ^ε for sufficiently small ε.

Thus the solution ρ is continuous up to the degenerate boundary. ∎

Continuity of ρ at Ξ_0 allows a strengthening of Lemma 4.4, as follows.

COROLLARY 4.6. *The free boundary η is smooth up to the degenerate boundary, namely, $\eta \in C^1[0, \xi_0]$.*

Proof of Theorem 2.3. Lemmas 4.2, 4.3, 4.4, and 4.5 show that there exists a solution pair $(\rho, \eta) \in C^{2+\alpha'}(\overline{\Omega} \setminus \{\sigma \cup \Xi_s\}) \cap C(\overline{\Omega}) \times C^{2+\alpha'}(0, \xi_0)$ satisfying (2.36), (2.37), (2.38), (2.39), and (2.40). This completes the proof of Theorem 2.3. ∎

5. Conclusions. Theorem 2.3 has constructed a solution ρ of the differential equation (2.36) in Ω; combining this function with the piecewise constant solution far from the origin, we obtain a function which is piecewise constant in the supersonic region, continuous across the degenerate boundary σ, and consistent with the derived form of the Rankine–Hugoniot conditions across the Mach stem. To recover the momentum components, m and n, we could in principle integrate equations (2.4) and (2.5), which can be written as transport equations in the radial variable r,

$$\frac{\partial m}{\partial r} = \frac{1}{r} c^2(\rho)\rho_\xi, \qquad \frac{\partial n}{\partial r} = \frac{1}{r} c^2(\rho)\rho_\eta, \tag{5.1}$$

and integrated from the boundary of the subsonic region toward the origin. We note that the sonic boundary can be written $r = r_0(\theta)$ and the boundary conditions for m and n are of the form $m(r_0(\theta), \theta) = m_0(\theta)$, where m_0 is piecewise continuous on σ and is determined from the Rankine–Hugoniot relation (2.15) on Σ; the component n is treated exactly the same way.

At σ and at Ξ_s, where we have proved only that ρ is continuous, equations (5.1), may not be meaningful. Elsewhere, m and n have the same regularity as ρ, except that discontinuities in m and n on the line $\xi = \kappa_a \eta$ may persist all the way in to the origin. In addition, the behavior of $c^2\rho_\eta/r$ in (5.1) at the origin causes a logarithmic singularity in n (but not in m: $c^2\rho_\xi/r$ remains bounded since $\rho_\xi(0, 0) = 0$).

Remark. There is some evidence of the unbounded behavior near the origin in the numerical simulations in [17]. This may presage difficulties in extending these results to the gas dynamics equations.

We argue heuristically that there is a difficulty at σ. Because of the construction, (ρ, m, n) is a weak solution of the system (1.1), or equivalently of the self-similar form (2.3)–(2.5), except possibly at the sonic boundary. It can be checked that the system (1.1) and the second-order equation (2.6), $Q(\rho) = 0$, are equivalent for weak solutions (that is, they conserve the same quantities). We can write (2.6) in the form $\mathrm{div}A = S$, with

$$A = (p_\xi - \xi^2\rho_\xi - \xi\eta\rho_\eta + \xi\rho, p_\eta - \eta^2\rho_\eta - \xi\eta\rho_\xi + \eta\rho),$$

and $S = -2\rho$. The usual multiplication by a smooth test function ϕ supported on a compact set D containing a segment of the degenerate boundary σ, followed by integration by parts, gives the weak form of the equation which must be satisfied for any weak solution in which $\nabla\rho$ is integrable (as is the case for our constructed function). Integrating by parts in the opposite sense on each side of $\Gamma \equiv \sigma \cap D$ yields the condition

$$\int_\Gamma \phi[A \cdot \nu]\, ds = 0,$$

where $[\ \ ]$ denotes the jump in the quantity and ν is the normal to σ. Since this must hold for all choices of D and ϕ, it holds pointwise. Furthermore, since the normal direction is the radial direction at σ, this means we need the function ρ inside Ω to satisfy

$$\lim_{r \to r_0} r\big(c^2(\rho) - r_0^2\big)\rho_r = 0. \tag{5.2}$$

We observe that for a linear wave equation, c^2 is constant and $r_0 = c$, and so this equation holds. However, for the function we constructed in Theorem 2.3 we have only the estimate $\rho - \rho_0 < A(r_0 - r)^\beta$ with $\beta < 1/2$ (see Lemma 4.5 and [7]) and this is not strong enough to give the limit (5.2). In fact, we have calculated, in [2] and [5], that the the behavior of solutions near a degenerate boundary like σ is exactly a square root singularity ($\beta = 1/2$), and so the function we have constructed fails to give a weak solution in the neighborhood of σ.

Acknowledgment. We would like to recognize the careful reading of an earlier version of the paper by the referees, whose efforts have improved the results.

REFERENCES

[1] S. Čanić and B. L. Keyfitz, *An elliptic problem arising from the unsteady transonic small disturbance equation*, J. Differential Equations, 125 (1996), pp. 548–574.

[2] S. Čanić and B. L. Keyfitz, *Quasi-one-dimensional Riemann problems and their role in self-similar two-dimensional problems*, Arch. Ration. Mech. Anal., 144 (1998), pp. 233–258.

[3] S. Čanić, B. L. Keyfitz, and E. H. Kim, *Free boundary problems for the unsteady transonic small disturbance equation: Transonic regular reflection*, Methods Appl. Anal., 7 (2000), pp. 313–336.

[4] S. Čanić, B. L. Keyfitz, and E. H. Kim, *A free boundary problem for a quasilinear degenerate elliptic equation: Regular reflection of weak shocks*, Comm. Pure Appl. Math., 55 (2002), pp. 71–92.

[5] S. Čanić, B. L. Keyfitz, and E. H. Kim, *Mixed hyperbolic-elliptic systems in self-similar flows*, Bol. Soc. Brasil. Mat., 32 (2002), pp. 1–23.

[6] S. Čanić, B. L. Keyfitz, and G. M. Lieberman, *A proof of existence of perturbed steady transonic shocks via a free boundary problem*, Comm. Pure Appl. Math., 53 (2000), pp. 1–28.

[7] S. Čanić and E. H. Kim, *A class of quasilinear degenerate elliptic equations*, J. Differential Equations, 189 (2003), pp. 71–98.

[8] G.-Q. Chen and M. Feldman, *Multidimensional transonic shocks and free boundary problems for nonlinear equations of mixed type*, J. Amer. Math. Soc., 16 (2003), pp. 461–494.

[9] G.-Q. Chen and M. Feldman, *Steady transonic shock and free boundary problems in infinite cylinders for the Euler equations*, Comm. Pure Appl. Math., 57 (2004), pp. 310–356.

[10] S. Chen, *Linear approximation of shock reflection at a wedge with large angle*, Comm. Partial Differential Equations, 21 (1996), pp. 1103–1118.

[11] S. Chen, *A singular multi-dimensional piston problem in compressible flow*, J. Differential Equations, 189 (2003), pp. 292–317.

[12] S. Chen, Z. Geng, and D. Li, *Existence and stability of conic shock waves*, J. Math. Anal. Appl., 277 (2003), pp. 512–532.

[13] Y. S. Choi and P. J. McKenna, *A singular quasilinear elliptic boundary value problem*, II, Trans. Amer. Math. Soc., 350 (1998), pp. 2925–2937.

[14] D. Gilbarg and L. Hörmander, *Intermediate Schauder estimates*, Arch. Ration. Mech. Anal., 74 (1980), pp. 297–318.

[15] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, New York, 1983.

[16] B. L. Keyfitz, *Self-similar solutions of two-dimensional conservation laws*, J. Hyperbolic Differential Equations, 1 (2004), pp. 445–492.

[17] B. L. Keyfitz, A. Kurganov, S. Čanić, and E. H. Kim, *Pattern and paradox: Shock interactions in the nonlinear wave system*, in preparation, 2006.

[18] G. M. Lieberman, *Regularized distance and its applications*, Pacific J. Math., 117 (1985), pp. 329–352.

[19] G. M. Lieberman, *Mixed boundary value problems for elliptic and parabolic differential equation of second order*, J. Math. Anal. Appl., 113 (1986), pp. 422–440.

[20] G. M. Lieberman, *Oblique derivative problems in Lipschitz domains.* I. *Continuous boundary data*, Boll. Unione Mat. Ital. Ser. B, 7 (1987), pp. 1185–1210.

[21] G. M. Lieberman, *Oblique derivative problems in Lipschitz domains.* II. *Discontinuous boundary data*, J. Reine Angew. Math., 389 (1988), pp. 1–21.

[22] G. M. Lieberman, *Optimal Hölder regularity for mixed boundary value problems*, J. Math. Anal. Appl., 143 (1989), pp. 572–586.

[23] O. Pironneau, *Shallow Water Computations*, 1991, private commuication.

[24] D. Serre, *Écoulements de fluides parfaits en deux variables indépendentes de type espace. Réflexion d'un choc plan par un dièdre compressif*, Arch. Ration. Mech. Anal., 132 (1995), pp. 15–36.

[25] A. M. Tesdall and J. K. Hunter, *Self-similar solutions for weak shock reflection*, SIAM J. Appl. Math., 63 (2002), pp. 42–61.

[26] T. Zhang and Y.-X. Zheng, *Conjecture on the structure of solutions of the Riemann problem for two-dimensional gas dynamics systems*, SIAM J. Math. Anal., 21 (1990), pp. 593–630.

[27] Y. Zheng, *A global solution to a two-dimensional Riemann problem involving shocks as free boundaries*, Acta Math. Appl. Sinica (English Ser.), 19 (2003), pp. 559–572.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1978–1987

CONVERGENCE OF AN ALGORITHM FOR THE ANISOTROPIC AND CRYSTALLINE MEAN CURVATURE FLOW*

ANTONIN CHAMBOLLE† AND MATTEO NOVAGA‡

Abstract. We give a simple proof of convergence of the anisotropic variant of a well-known algorithm for mean curvature motion, introduced in 1992 by Merriman, Bence, and Osher. The algorithm consists in alternating the resolution of the (anisotropic) heat equation, with initial datum the characteristic function of the evolving set, and a thresholding at level 1/2.

Key words. anisotropic mean curvature flow, heat equation, signed distance function

AMS subject classifications. 35K65, 35K55, 53C44

DOI. 10.1137/050629641

1. Introduction: The algorithm.

More than ten years ago, Merriman, Bence, and Osher [26] proposed the following algorithm for the computation of the motion by mean curvature of a surface. Given a closed set $E \subset \mathbb{R}^N$, they let $T_h E = \{u(\cdot, h) \geq 1/2\}$, where u solves the following heat equation:

$$\left\{ \begin{array}{rcll} \dfrac{\partial u}{\partial t}(x,t) & = & \Delta u(x,t)\,, & t > 0,\ x \in \mathbb{R}^N\,, \\ u(\cdot, 0) & = & \chi_E & (t = 0)\,. \end{array} \right. \tag{1}$$

Then, they let $E_h(t) = T_h^{[t/h]} E$ (with $[t/h]$ the integer part of t/h), and conjectured that $\partial E_h(t)$ converges to $\partial E(t)$, as $h \to 0$, where $\partial E(t)$ is the (generalized) evolution by mean curvature starting from ∂E.

The proof of convergence of this scheme was established by Evans [17] and Barles and Georgelin [2]. Other proofs were given by Ishii [22] and Cao [12], where the evolution in (1) was replaced by the convolution of χ_E with a more general symmetric kernel. This was generalized by Ishii, Pires, and Souganidis [23] to the case of the convolution with an arbitrary kernel (with some growth assumptions). This approach was also studied by Ruuth and Merriman [29] (see also [28]). Vivier [34] and Leoni [25] have considered other generalizations with (1) replaced with a time and space dependent anisotropic heat equation with a lower order term. The space dependence is an additional difficulty and it is not clear how what we will present could be adapted to such situations; on the other hand, in the two latter papers, "only" the case of Riemannian anisotropies is considered, in contrast to what we will study here.

We propose here to study the generalization of this algorithm to the so-called *anisotropic* and *crystalline curvature motion*, as defined in [21, 33, 32, 31]. We follow the definition in [10]: we consider (ϕ, ϕ°) a pair of mutually polar convex 1-homogeneous functions in $\mathbb{R}^N$ (i.e., $\phi^\circ(\xi) = \sup_{\phi(\eta)\leq 1} \xi \cdot \eta$, $\phi(\eta) = \sup_{\phi^\circ(\xi)\leq 1} \xi \cdot \eta$; see [27]). These are assumed to be locally finite, and, to simplify, even. The pair

*Received by the editors April 20, 2005; accepted for publication (in revised form) December 12, 2005; published electronically March 3, 2006.
http://www.siam.org/journals/sima/37-6/62964.html

†CMAP, CNRS UMR 7641, École Polytechnique, 91128, Palaiseau Cedex, France (antonin.chambolle@polytechnique.fr).

‡Dipartimento di Matematica, Università di Pisa, via Buonarroti 2, 56127 Pisa, Italy (novaga@dm.unipi.it).

(ϕ,ϕ°) will be referred as *the anisotropy.* The local finiteness implies that there is a constant $c>1$ such that

$$c^{-1}|\eta| \le \phi(\eta) \le c|\eta| \quad\text{and}\quad c^{-1}|\xi| \le \phi^\circ(\xi) \le c|\xi|$$

for any η and ξ in $\mathbb{R}^N$. We refer to [9, 10] for the main properties of ϕ and ϕ°.

Being convex and 1-homogeneous, ϕ° (and ϕ) is also subadditive, so that the function $(x,y)\mapsto\phi(x-y)$ defines a distance—the "ϕ-distance." For $E\subset\mathbb{R}^N$ and $x\in\mathbb{R}^N$, we denote by $\mathrm{dist}^\phi(x,E) := \inf_{y\in E}\phi(x-y)$ the ϕ-distance of x to the set E, and by

$$d_E^\phi(x) \;:=\; \mathrm{dist}^\phi(x,E) \,-\, \mathrm{dist}^\phi(x,\mathbb{R}^N\setminus E)$$

the signed ϕ-distance to ∂E, negative in the interior of E and positive outside its closure. One easily checks that

$$|d_E^\phi(x)-d_E^\phi(y)| \;\le\; \phi(x-y) \;\le\; c|x-y|$$

for any $x,y\in\mathbb{R}^N$, so that d_E^ϕ is differentiable a.e. in $\mathbb{R}^N$. The former inequality shows moreover that $\nabla d_E^\phi(x)\cdot h \le \phi(h)$ for any $h\in\mathbb{R}^N$ if x is a point of differentiability; hence $\phi^\circ(\nabla d_E^\phi(x))\le 1$. If ϕ and ϕ° are smooth, one shows quite easily that d_E^ϕ is differentiable at each point x which has a unique ϕ-projection $y\in\partial E$ (solving $\min_{y\in\partial E}\phi(x-y)$). In this case, $\nabla d_E^\phi(x)$ is given by $\nabla\phi((x-y)/d_E^\phi(x))$, so that $\phi^\circ(\nabla d_E^\phi(x))=1$. If ϕ, ϕ° are just Lipschitz-continuous, one still shows that $\phi^\circ(\nabla d_E^\phi(x))=1$ a.e. in $\mathbb{R}^N$; see [9, 10] for details.

A *Cahn–Hoffman* vector field n_ϕ is a vector field on ∂E such that $n_\phi(x)\in\partial\phi^\circ(\nu_E(x))=\partial\phi^\circ(\nabla d_E^\phi(x))$ a.e. on ∂E, where $\partial\phi^\circ$ is the (0-homogeneous) subgradient of ϕ° (see [27, 16]) and ν_E is the (Euclidean) exterior normal to ∂E. If such a field is given in a neighborhood of ∂E, then it is characterized by

$$\phi^\circ(n_\phi(x)) \;=\; 1 \quad\text{and}\quad n_\phi(x)\cdot\nabla d_E^\phi(x)=1 \quad\text{a.e.}$$

This follows from Euler's identity, since ϕ° is 1-homogeneous. In this case, $\kappa_\phi=\operatorname{div} n_\phi$ is a *ϕ-curvature* of ∂E. The ϕ-curvature flow is then an evolution $E(t)$ such that at each time, the velocity of $\partial E(t)$ is given by

$$V \;=\; -\kappa_\phi\, n_\phi\,, \tag{2}$$

where n_ϕ is a Cahn–Hoffman vector field and κ_ϕ is the associated curvature. If ϕ,ϕ° are smooth (e.g., in $C^2(\Omega\setminus\{0\})$), then n_ϕ,κ_ϕ are uniquely defined, whereas if ϕ, ϕ° are merely Lipschitz (when, for instance, the *Wulff shape* $\{\phi\le 1\}$ is a convex polytope), then n_ϕ can be nonunique and the anisotropy is called *crystalline* [33, 9].

As easily shown by formal asymptotic expansion, the natural anisotropic generalization of the Merriman–Bence–Osher algorithm is as follows. Given E a closed set with compact boundary in $\mathbb{R}^N$, we let $T_h(E)=\{x : u(x,h)\ge 1/2\}$, where $u(x,t)$ is the solution of

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) \;\in\; \operatorname{div}\Big(\phi^\circ(\nabla u)\partial\phi^\circ(\nabla u)\Big)(x,t)\,, & t>0,\ x\in\mathbb{R}^N\,,\\ u(\cdot,0) \;=\; \chi_E & (t=0)\,. \end{cases} \tag{3}$$

The function $u(x,t)$ is well defined and unique by classical results on contraction semigroups [11]: if E is compact, it corresponds to the flow in $L^2(\mathbb{R}^N)$ of the subdifferential of the functional $u \mapsto \int_{\mathbb{R}^N} \phi^\circ(\nabla u)^2/2\,dx$ if $u \in H^1(\mathbb{R}^N)$, and $+\infty$ otherwise. On the other hand, if $\mathbb{R}^N \setminus E$ is compact, one defines u by simply letting $u = 1 + v$, where v solves the same equation with initial data $\chi_E - 1$.

We are interested in the limit of the discrete evolutions $t \mapsto E_h(t) = T_h^{[t/h]}E$, as $h \to 0$. Our main result is a result of consistency with suitable "regular" evolutions: it states that if there exists a regular evolution starting from E in the sense of our Definition 2.1 (which is a variant of a definition first introduced in [9] and includes smooth evolutions when the anisotropy is smooth), then $E_h(t)$ converges to this evolution. This consistency result, together with the monotonicity of the scheme ($E \subseteq F \Rightarrow T_hE \subseteq T_hF$, as follows from the comparison principle for (3)), yields convergence also to all generalized solutions defined (in the smooth case) using barriers [7, 8] or, equivalently, viscosity solutions [14, 15, 4, 5, 3], as long as these are unique. Also, it yields the convergence of the scheme to crystalline evolutions, when the initial set is convex. Existence and uniqueness of such (regular and generalized) evolutions are established, in the convex case, in [6].

Another important consequence of our consistency result is a comparision principle for the regular evolutions of Definition 2.1, which follows from the monotonicity of the scheme. It gives an alternative proof of uniqueness for the convex crystalline evolutions studied in [6] (the original proof is based on [9]).

We observe that evolutions similar to (2) might also be obtained by convolution with appropriate kernels as studied by Ishii, Pires, and Souganidis [23]. However, a complete characterization of these motions in dimension higher than 2 is still not known (see [29, 30] in two dimensions).

Our evolution is also different from the evolutions considered by Leoni [25] (or Vivier [34]); in her paper, the heat equation (1) is replaced with an equation of the form $u_t = A(x,t) : D^2u + H(x,t,Du)$. The resulting surface motion is a variant of the mean curvature motion, with an (x,t) dependent velocity which is a function of a Riemannian curvature (depending on A) plus a lower order forcing term.

It would be interesting to prove a similar consistency result for the variational variant of (3), which is somehow simpler to solve numerically (in the truly nonlinear anisotropic cases): for $E \subset \mathbb{R}^N$ bounded, one would define $T_hE = \{u_h \geq 1/2\}$, where u_h is the solution of (with $\Omega \Supset E$ bounded or $\Omega = \mathbb{R}^N$)

$$\min_{u \in H^1(\Omega)} \int_\Omega \phi^\circ(\nabla u(x))^2 \,+\, \frac{1}{h}(u(x) - \chi_E(x))^2\,dx\,. \tag{4}$$

Although it is likely that this variant produces the same evolution as the original scheme (it is true in the isotropic case, since u_h is given by the convolution of χ_E with a radially symmetric kernel), we could not extend our proof in all cases to this new scheme.

Our proof follows the same idea as our recent proof of consistency [13] for (a generalization of) the variational algorithm of Almgren, Taylor, and Wang [1]. However, we have just learned that Goto, Ishii, and Ogawa [20, 24] have recently given a new proof of the convergence of the Merriman–Bence–Osher algorithm, in the isotropic case, which is very similar to the proof we give here.

2. The consistency result and some consequences. If $E \subset \mathbb{R}^N$ we say that E satisfies the interior rW_ϕ-condition if and only if for any $x \in \partial E$ there exist $y \in E$

with $\phi(x-y) = r$ and $\phi(x'-y) \geq r$ for any $x' \in \mathbb{R}^N \setminus E$. We say that E satisfies the exterior rW_ϕ-condition if $\mathbb{R}^N \setminus E$ satisfies the interior rW_ϕ-condition.

We will show a consistency result with regular evolutions of (2), in the sense of the following definition.

DEFINITION 2.1. *We say that $t \mapsto E(t)$ is an rW_ϕ-regular ϕ-curvature flow on $[t_0, t_1]$, $t_0 < t_1$, if and only if*

(i) *for any $t \in [t_0, t_1]$, $E(t)$ satisfies the interior and exterior rW_ϕ-conditions;*

(ii) *there exists a bounded and relatively open neighborhood A of $\bigcup_{t_0 \leq t \leq t_1} \partial E(t) \times \{t\}$ in $\mathbb{R}^N \times [t_0, t_1]$ such that $d(x,t) := d^\phi_{E(t)}(x)$ is Lipschitz in A;*

(iii) *there exists a vector field $n : A \to \mathbb{R}^N$ with $n \in \partial\phi^\circ(\nabla d)$ a.e. in A, and $\operatorname{div} n \in L^\infty(A)$; and*

(iv) *there exists $\overline{c} > 0$ such that $|\partial d/\partial t - \operatorname{div} n| \leq \overline{c}|d|$ a.e. in A.*

This definition, up to the additional requirement that $E(t)$ satisfies an interior and exterior rW_ϕ-condition, is due to Bellettini and Novaga [9, Def. 2.2].

Such evolutions are known to exist if ϕ, ϕ°, and ∂E are smooth enough (for instance, in $C^{3,\alpha}(\mathbb{R}^N \setminus \{0\})$ [1]), or for any ϕ, ϕ°, when the initial set E is convex and satisfies an interior rW_ϕ-condition (exterior is always true in the case of convex sets) [6]. They also exist in the purely crystalline case, i.e., when both ϕ and ϕ° are piecewise linear in dimension $N = 2$ [18, 19, 31] (see section 4 for an example).

Our main theorem states that the anisotropic Merriman–Bence–Osher scheme is consistent with such evolutions.

THEOREM 2.2. *Let E be a regular flow in the sense of Definition* 2.1*, on a time interval* $[t_0, t_1]$*. Then, for any t and τ with $t_0 \leq t < t + \tau \leq t_1$, $\partial T_h^{[\tau/h]} E(t)$ converges to $\partial E(t+\tau)$ in the Hausdorff sense, as $h \to 0$.*

The following corollary, also shown in [9], is obvious.

COROLLARY 2.3. *Let E, F be two flows in the sense of Definition* 2.1*, on* $[t_0, t_1]$*, and assume $E(t_0) \subseteq F(t_0)$. Then $E(t) \subseteq F(t)$ for all $t \in [t_0, t_1]$. In particular, if $E(t_0) = F(t_0)$, then $E(t) = F(t)$ for all $t \in [t_0, t_1]$.*

The next corollary follows, with a standard proof (see [4, 5]), from the monotonicity and consistency of the scheme.

COROLLARY 2.4. *Assume $E \subset \mathbb{R}^N$ is a closed set with compact boundary such that the generalized ϕ-curvature flow $E(t)$, starting from E, is uniquely defined on a time interval $[0, T)$ (e.g., ϕ, $\phi^\circ \in C^2(\mathbb{R}^N \setminus \{0\})$, and no fattening occurs* [14]*). Then $\partial T_h^{[t/h]} E(t) \to \partial E(t)$ in the Hausdorff sense for any $t < T$, as $h \to 0$. The same conclusion holds for any ϕ, ϕ° if E is convex, by the uniqueness result in* [6].

Let us observe that this result follows easily from Theorem 2.2 when evolutions according to Definition 2.1 are known to exist. If not (e.g., if ϕ, ϕ° are merely C^2), this is still true; however, the proof relies on a comparison with appropriate strict super- and subsolutions, defined according to obvious modifications of Definition 2.1 (as in [13]).

Remark 1. In case ϕ, ϕ° are not even, Theorem 2.2 still holds, but (i) the signed distance to the interface $d^\phi_{E(t)}(x)$ must be defined, in Definition 2.1, in a nonsymmetric way, and (ii) the term $\partial\phi^\circ(\nabla u)$ in (3) must be replaced with $-\partial\phi^\circ(-\nabla u)$ (since ∇u has a reverse orientation with respect to the outer normal to the set E).

3. Proof of Theorem 2.2. The proof of Theorem 2.2 is divided into several steps. The idea is to build appropriate sub- and supersolutions to (3), by means of the function $d(x,t)$, and to compare $T_h E(t)$ with $E(t+h)$.

These barriers will be built by means of the function $\gamma : \mathbb{R} \times [0,+\infty) \to [0,1]$, which solves the heat equation

$$\begin{cases} \dfrac{\partial\gamma}{\partial\tau}(\xi,\tau) \;=\; \dfrac{\partial^2\gamma}{\partial\xi^2}(\xi,\tau)\,, & \xi\in\mathbb{R}\,,\ \tau>0\,,\\ \gamma(\xi,0)=Y(\xi)\,, & \xi\in\mathbb{R}\ \ (\tau=0)\,, \end{cases} \tag{5}$$

where $Y = \chi_{[0,+\infty)}$ is the Heaviside function. It is well known that γ is given by

$$\gamma(\xi,\tau) \;=\; \frac{1}{2\sqrt{\pi\tau}}\int_{-\infty}^{\xi} e^{-\frac{s^2}{4\tau}}\,ds\,.$$

In particular, one readily sees that it is self-similar: indeed, the change of variables $s' = s/\sqrt{\tau}$ yields

$$\gamma(\xi,\tau) \;=\; \frac{1}{2\sqrt{\pi}}\int_{-\infty}^{\frac{\xi}{\sqrt{\tau}}} e^{-\frac{s'^2}{4}}\,ds' \;=\; \gamma\left(\frac{\xi}{\sqrt{\tau}},1\right) \;=:\; \gamma_1\left(\frac{\xi}{\sqrt{\tau}}\right)\,.$$

We first show the following (obvious) result.

LEMMA 3.1. *For any $\varepsilon>0$, there exists $\tau_0>0$ such that if $0\le\tau\le\tau_0$, then $\gamma(\varepsilon,\tau)\ge 1-\tau$.*

Proof. We just need to observe that $\tau\mapsto\gamma(\varepsilon,\tau)$ is C^1 with derivative 0 at 0. This derivative is indeed given by $(-\varepsilon/\tau^{3/2})\gamma_1'(\varepsilon/\sqrt{\tau}) = (-\varepsilon/\tau^{3/2})\exp(-\varepsilon^2/(4\tau))$. There exists τ_0 such that it is in $[-1,0]$ for $\tau\le\tau_0$; hence $\gamma(\varepsilon,\tau)\ge\gamma(\varepsilon,0)-\tau$ if $\tau\in[0,\tau_0]$, which shows the lemma. $\square$

Let us now consider E, $r>0$, $t_0\le t_1$, A, and the functions $d(x,t)$, $n(x,t)$, as in Definition 2.1. Possibly reducing r, we can assume that $\{|d|\le r\}\subset A$. Let us fix $t\in[t_0,t_1)$, $\delta\in[0,r/2]$ and let $F=\{d(\cdot,t)\le\delta\}$. Let u be the solution of

$$\begin{cases} \dfrac{\partial u}{\partial\tau}(x,\tau) \;\in\; \operatorname{div}\Big(\phi^\circ(\nabla u)\partial\phi^\circ(\nabla u)\Big)(x,\tau)\,, & \tau>0,\ x\in\mathbb{R}^N\,,\\ u(\cdot,0) \;=\; \chi_F \;=\; Y(-d(\cdot,t)+\delta) & (\tau=0)\,. \end{cases} \tag{6}$$

We first show the following result.

LEMMA 3.2. *For any $\varepsilon\in(0,r/2)$, there exists $\tau_0>0$ (independent of δ) such that $\tau\le\tau_0$ yields $u(x,\tau)\le\tau$ for any x such that $d(x,t)-\delta=\varepsilon$.*

Proof. Let us fix $x_0\in\mathbb{R}^N\setminus F$ with $d(x_0,t)-\delta=\varepsilon$. Since $E(t)$ satisfies the exterior rW_ϕ-condition, the function $(d(\cdot,t)-\delta)$ is, outside F, equal to $\operatorname{dist}^\phi(\cdot,F)$. Hence, letting $W=\{x : \phi(x-x_0)<\varepsilon\}$, one sees that $W\cap F=\emptyset$. We deduce that $\chi_F\le 1-\chi_W$ in $\mathbb{R}^N$, so that $u(\cdot,\tau)\le 1-w(\cdot,\tau)$, where w is the solution of

$$\begin{cases} \dfrac{\partial w}{\partial\tau}(x,\tau) \;\in\; \operatorname{div}\Big(\phi^\circ(\nabla w)\partial\phi^\circ(\nabla w)\Big)(x,\tau)\,, & \tau>0,\ x\in\mathbb{R}^N\,,\\ w(\cdot,0) \;=\; \chi_W & (\tau=0)\,. \end{cases}$$

This solution is explicitly given by $w(x,\tau)=U(\phi(x-x_0)/\varepsilon,\tau/\varepsilon^2)$, where $U(|x|,\tau)=\tilde U(x,\tau)$ and $\tilde U$ is the (radial) solution of the heat equation $\partial\tilde U/\partial t=\Delta\tilde U$ with initial datum χ_{B_1}, the characteristic function of the unit ball in $\mathbb{R}^N$. It is well known that

$$\tilde U(x,\tau)=\frac{1}{\sqrt{4\pi\tau}^N}\int_{\{|y|\le 1\}}\exp\left(-\frac{|x-y|^2}{4\tau}\right)dy$$

so that

$$U(|x|,\tau) \;=\; \frac{1}{\sqrt{4\pi\tau}^N}\int_{\{|y|\le 1\}} \exp\left(-\frac{(|x|-y_1)^2+\sum_{i=2}^N y_i^2}{4\tau}\right)\,dy\,.$$

Using arguments similar to the proof of the previous lemma (based on the fact that U is smooth near $(\xi,\tau)=0,0$ and $\partial U/\partial t(0,0)=0$), one sees that there exists $\tau_0>0$ such that if $\tau\le\tau_0$, $U(0,\tau)\ge 1-\varepsilon^2\tau$; hence $w(0,\tau)\ge 1-\tau$ if $\tau\le\tau_0'=\varepsilon^2\tau_0$. We deduce that $u(x_0,\tau)\le\tau$ if $\tau\le\tau_0'$, depending only on ε. This shows the lemma. ◻

Let us fix $\varepsilon<r/4$ and let us look for a supersolution of (6) on a time interval $[0,h]$, h small, of the form

$$v(x,\tau) \;=\; \gamma\big(-d(x,t+\tau)+\delta+\overline{c}\,\overline{\varepsilon}\tau,\tau\big)+h\,,$$

in $B=\bigcup_{0\le\tau\le h}\{x\,:\,d(x,t)-\delta\le\varepsilon\,,d(x,t+\tau)-\delta\ge-\varepsilon\}\times\{\tau\}$, where the constant $\overline{\varepsilon}$ will be made precise later on. We observe that since the speed of the motion is bounded at any time for τ small enough, if h is small enough (depending only on r,ε), B remains inside $\{(x,\tau)\in\mathbb{R}^N\times[0,h]\,:\,\delta-\varepsilon\le d(x,t+\tau)\le\delta+2\varepsilon\}$, and $(0,t)+B\subset A$.

At $\tau=0$, $v(x,0)=Y(-d(x,t)+\delta)+h$ is strictly larger than $\chi_F(x)=u(x,0)$. If $0\le\tau\le h$ and $d(x,t)-\delta=\varepsilon$, by Lemma 3.2 we have $u(x,\tau)\le\tau\le h\le v(x,\tau)$, provided h is small enough. If, on the other hand, $d(x,t+\tau)-\delta=-\varepsilon$, then by Lemma 3.1, still for h small enough, $v(x,\tau)=\gamma(-d(x,t+\tau)+\delta+\overline{c}\,\overline{\varepsilon}\tau,\tau)+h\ge\gamma(\varepsilon,\tau)+h\ge 1-\tau+h$; hence $v(x,\tau)\ge 1\ge u(x,\tau)$. We find that $v\ge u$ on $\{(x,\tau)\in\partial B\,:\,\tau<h\}$, which is the parabolic boundary of B (and, in fact, our proof even shows that $v\ge u$ in a neighborhood of this boundary).

Hence, to get that v is a supersolution of (6) in B, one has to show that $\partial v/\partial\tau\ge\operatorname{div}Z$ for some field $Z\in\phi^\circ(\nabla v)\partial\phi^\circ(\nabla v)$ inside B.

One has, a.e. in B,

$$\text{(7)}\quad \frac{\partial v}{\partial\tau}(x,\tau) \;=\; \left(-\frac{\partial d}{\partial t}(x,t+\tau)+\overline{c}\,\overline{\varepsilon}\right)\frac{\partial\gamma}{\partial\xi}(-d(x,t+\tau)+\delta+\overline{c}\,\overline{\varepsilon}\tau,\tau) + \frac{\partial\gamma}{\partial\tau}(-d(x,t+\tau)+\delta+\overline{c}\,\overline{\varepsilon}\tau,\tau)\,,$$

whereas

$$\nabla v(x,\tau) \;=\; -\frac{\partial\gamma}{\partial\xi}(-d(x,t+\tau)+\delta+\overline{c}\,\overline{\varepsilon}\tau,\tau)\nabla d(x,t+\tau)\,.$$

Using the assumption that ϕ° is even, we see that $\phi^\circ(\nabla v)=\partial\gamma/\partial\xi$ (since $\phi^\circ(\nabla d)=1$ a.e. in $\mathbb{R}^N$) and that $\partial\phi^\circ(\nabla v)=-\partial\phi^\circ(\nabla d)$ (since $\partial\gamma/\partial\xi>0$ and $\partial\phi^\circ$ is 0-homogeneous and odd). Let now

$$Z(x,\tau) \;=\; -\frac{\partial\gamma}{\partial\xi}(-d(x,t+\tau)+\delta+\overline{c}\,\overline{\varepsilon}\tau,\tau)\,n(x,t+\tau) \;=\; \phi^\circ(\nabla v(x,\tau))(-n(x,t+\tau)).$$

Since (by assumption) $n(x,t+\tau)\in\partial\phi^\circ(\nabla d(x,t+\tau))=-\partial\phi^\circ(\nabla v(x,\tau))$ for a.e. x in $\mathbb{R}^N$ and any $\tau\in(0,h)$, one has $Z(x,\tau)\in\phi^\circ(\nabla v)\partial\phi^\circ(\nabla v)(x,\tau)$. Since ϕ° is 1-homogeneous, Euler's identity yields $\nabla d\cdot n=\phi^\circ(\nabla d)=1$ as soon as $n\in\partial\phi^\circ(\nabla d)$. We deduce

$$\begin{aligned}(8)\quad \operatorname{div} Z(x,\tau) \;&=\; -\operatorname{div}\left[\frac{\partial\gamma}{\partial\xi}(-d(x,t+\tau)+\delta+\overline{c}\,\overline{\varepsilon}\tau,\tau)n(x,t+\tau)\right]\\ &=\; \frac{\partial^2\gamma}{\partial\xi^2}(-d(x,t+\tau)+\delta+\overline{c}\,\overline{\varepsilon}\tau,\tau)\\ &\qquad -\;\frac{\partial\gamma}{\partial\xi}(-d(x,t+\tau)+\delta+\overline{c}\,\overline{\varepsilon}\tau,\tau)(\operatorname{div} n(x,t+\tau))\,.\end{aligned}$$

Since we have $\partial d/\partial t \le \operatorname{div} n + \overline{c}|d|$ in B, we deduce using (7) and (8) that

$$\begin{aligned}\frac{\partial v}{\partial\tau}(x,\tau) \;&\ge\; \operatorname{div} Z(x,\tau) \;-\; \frac{\partial^2\gamma}{\partial\xi^2}(-d(x,t+\tau)+\delta+\overline{c}\,\overline{\varepsilon}\tau,\tau)\\ &+\overline{c}(\overline{\varepsilon}-|d(x,t+\tau)|)\frac{\partial\gamma}{\partial\xi}(-d(x,t+\tau)+\delta+\overline{c}\,\overline{\varepsilon}\tau,\tau) \;+\; \frac{\partial\gamma}{\partial\tau}(-d(x,t+\tau)+\delta+\overline{c}\,\overline{\varepsilon}\tau,\tau)\,.\end{aligned}$$

Now, γ satisfies the heat equation, so that if $|d(x,t+\tau)| \le \overline{\varepsilon}$ a.e. in B, we get

$$(9)\qquad \frac{\partial v}{\partial\tau}(x,\tau) \;\ge\; \operatorname{div} Z(x,\tau)\,.$$

We choose $\overline{\varepsilon} = \delta + 2\varepsilon$ so that $|d(x,t+\tau)| \le \overline{\varepsilon}$ a.e. in B and (9) holds. By standard comparison results on parabolic equations (see [11]), we deduce that $v(x,h) \ge u(x,h)$. In particular, we have shown that there exists $h_0 > 0$ (depending only on r,ε) such that if $h < h_0$,

$$\begin{aligned}T_hF \;&=\; \left\{u(\cdot,h) \ge \frac{1}{2}\right\} \;\subset\; \left\{v(\cdot,h) \ge \frac{1}{2}\right\}\\ &=\; \left\{x\in\mathbb{R}^N \,:\, d(x,t+h) \le \delta + \overline{c}\,\overline{\varepsilon}h - [\gamma(\cdot,h)]^{-1}\left(\frac{1}{2}-h\right)\right\}.\end{aligned}$$

Since $\gamma_1(0) = 1/2$, $\gamma_1'(0) = 1/(2\sqrt{\pi})$, we have $\gamma_1^{-1}(1/2-h) = -2\sqrt{\pi}h + o(h)$. Now, $\gamma(\xi,h) = \gamma_1(\xi/\sqrt{h})$, so that $[\gamma(\cdot,h)]^{-1} = \sqrt{h}\gamma_1^{-1}$. We find that $[\gamma(\cdot,h)]^{-1}(1/2-h) = (-2\sqrt{\pi}h + o(h))\sqrt{h}$. Hence, possibly reducing h_0, one gets that if $h < h_0$, then $[\gamma(\cdot,h)]^{-1}(1/2-h) \ge -4h^{3/2}$. Recalling that $\overline{\varepsilon} = \delta+2\varepsilon$, we find that if $h<h_0$,

$$T_hF \;\subset\; \left\{x\in\mathbb{R}^N \,:\, d(x,t+h) \le [\delta + (\overline{c}(\delta+2\varepsilon)+4\sqrt{h})h]\right\}.$$

Now, we deduce that ($\varepsilon \in (0,r/4)$ being fixed) if $t \in [t_0,t_1)$, $h \le h_0$, and $k \ge 1$ with $t+kh \le t_1$, one has

$$T_h^k(E(t)) \;\subset\; \left\{x\in\mathbb{R}^N \,:\, d(x,t+kh) \le \delta_k\right\}$$

with $\delta_0 = 0$ and $\delta_k = \delta_{k-1} + (\overline{c}(\delta_{k-1}+2\varepsilon)+4\sqrt{h})h$, as long as $\delta_{k-1} \le r/2$. By induction, we find

$$\delta_k \;=\; \left((1+\overline{c}h)^k - 1\right)\left(2\varepsilon + \frac{4\sqrt{h}}{\overline{c}}\right).$$

In particular, if $\tau > 0$ is fixed, with $t+\tau \le t_1$, and $k = [\tau/h]$, we see that $\lim_{h\to 0}\delta_k = 2\varepsilon(e^{\overline{c}\tau}-1)$. If $\varepsilon < r/4$ is chosen small enough (less than $(r/4)/(e^{\overline{c}\tau}-1)$), we see that for $h>0$ small enough, $\delta_{[\tau/h]} \le r/2$.

We now recall that any sequence of sets in $\mathbb{R}^N$ with equibounded boundaries has a subsequence that converges in the Hausdorff sense to a closed set. If E' is the Hausdorff limit of a converging subsequence of $T_h^{[\tau/h]}E(t)$, as $h\to 0$, we deduce that $E' \subseteq \{d(\cdot,t+\tau) \le 2\varepsilon(e^{\overline{c}\tau}-1)\}$. Since this must be true for all $\varepsilon > 0$ small enough, one sees that $E' \subseteq E(t+\tau)$. On the other hand, a symmetric argument (based on subsolutions of the equation) will yield that if $\mathbb{R}^N \setminus E''$ is the Hausdorff limit of a converging subsequence of $(\mathbb{R}^N \setminus T_h^{[\tau/h]}E(t))_{h>0}$, then $\mathbb{R}^N \setminus E'' \subseteq \overline{\mathbb{R}^N \setminus E(t+\tau)}$; that is, $\mathrm{int}(E(t+\tau)) \subseteq E''$. Without loss of generality, one can choose the same subsequence in both limits above: in this case, one can show that $E'' \subset E'$, and $E' \setminus E''$ is the Hausdorff limit of $\partial T_h^{[\tau/h]}E(t)$ (which might differ from $\partial E'$ or $\partial E''$). Since $\mathrm{int}(E(t+\tau)) \subseteq E'' \subset E' \subseteq E(t+\tau)$, we see that $E'' = \mathrm{int}(E(t+\tau))$, $E' = E(t+\tau)$, $E' \setminus E'' = \partial E(t+\tau)$, and by uniqueness of this Hausdorff limit we deduce Theorem 2.2. $\square$

4. A numerical example. The algorithm is quite easy to implement numerically. Of course, there is some difficulty in computing precisely the solution of (3) in strongly anisotropic or crystalline cases, especially when the subgradient $\partial\phi^\circ$ is multivalued. We experimented with an implicit method, based on iterative resolutions of the variational problem (4). More precisely, we approximate $u(\cdot,h)$ with $w_n(x)$, where $h = nh'$, n is a fixed (small) integer, $w_0 = \chi_E$, and for $i = 0,\dots,n-1$, w_{i+1} solves (in a domain Ω "large" with respect to E)

$$\min_{w\in H^1(\Omega)} \int_\Omega \phi^\circ(\nabla w(x))^2 + \frac{1}{h'}(w(x)-w_i(x))^2\,dx\,.$$

To solve this minimization problem in the crystalline case, we discretize (here, on a bidimensional finite differences grid) and solve the dual problem (see, for instance, [16])

$$\min_{\xi\in L^2(\Omega;\mathbb{R}^N)} \int_\Omega \phi(\xi(x))^2 + h'((w_i(x)/h') - \mathrm{div}\,\xi(x))^2\,dx\,,$$

using a conjugate-gradient method. Then, $w_{i+1} = w_i - h'\mathrm{div}\,\xi$. The thresholding at level $1/2$ is replaced by a "soft thresholding" $w_n(x) \mapsto \min\{1, \max\{1/2 + \sigma(w_n(x) - 1/2), 0\}\}$, where σ is adapted to the spatial discretization step, in order to keep a precision slightly higher than the grid's. In the example shown in Figure 1, the Wulff shape $\{\phi \le 1\}$ is a hexagon.

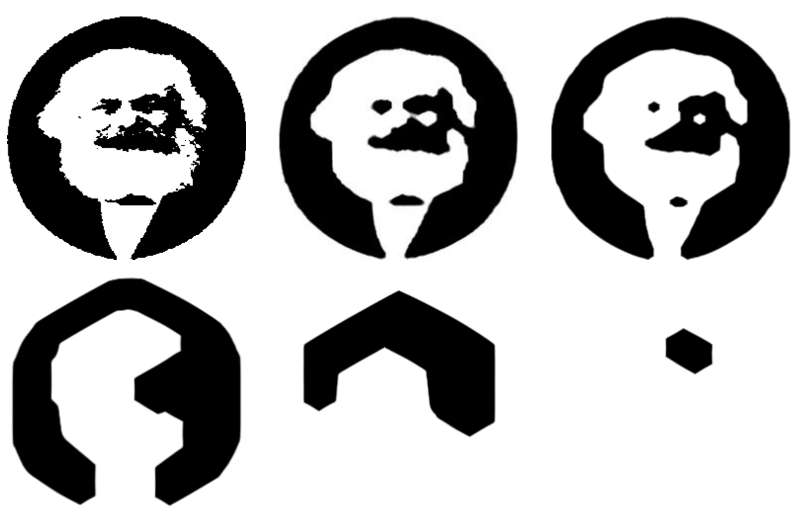


FIG. 1. *Evolutions at times* $t = 0, 5, 25, 60, 400, 800$.

REFERENCES

[1] F. Almgren, J. E. Taylor, and L.-H. Wang, *Curvature-driven flows: A variational approach*, SIAM J. Control Optim., 31 (1993), pp. 387–438.

[2] G. Barles and C. Georgelin, *A simple proof of convergence for an approximation scheme for computing motions by mean curvature*, SIAM J. Numer. Anal., 32 (1995), pp. 484–500.

[3] G. Barles, H. M. Soner, and P. E. Souganidis, *Front propagation and phase field theory*, SIAM J. Control Optim., 31 (1993), pp. 439–469.

[4] G. Barles and P. E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptot. Anal., 4 (1991), pp. 271–283.

[5] G. Barles and P. E. Souganidis, *A new approach to front propagation problems: Theory and applications*, Arch. Ration. Mech. Anal., 141 (1998), pp. 237–296.

[6] G. Bellettini, V. Caselles, A. Chambolle, and M. Novaga, *Crystalline mean curvature flow of convex sets*, Arch. Ration. Mech. Anal., 179 (2006), pp. 109–152.

[7] G. Bellettini and M. Novaga, *Minimal barriers for geometric evolutions*, J. Differential Equations, 139 (1997), pp. 76–103.

[8] G. Bellettini and M. Novaga, *Comparison results between minimal barriers and viscosity solutions for geometric evolutions*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 26 (1998), pp. 97–131.

[9] G. Bellettini and M. Novaga, *Approximation and comparison for nonsmooth anisotropic motion by mean curvature in* $\mathbf{R}^N$, Math. Models Methods Appl. Sci., 10 (2000), pp. 1–10.

[10] G. Bellettini and M. Paolini, *Anisotropic motion by mean curvature in the context of Finsler geometry*, Hokkaido Math. J., 25 (1996), pp. 537–566.

[11] H. Brézis, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland Math. Stud. 5, Notas de Matemática (50), North–Holland, Amsterdam, 1973.

[12] F. Cao, *Partial differential equations and mathematical morphology*, J. Math. Pures Appl. (9), 77 (1998), pp. 909–941.

[13] A. Chambolle and M. Novaga, *Implicit Time Discretization of the Mean Curvature Flow with a Discontinuous Forcing Term*, preprint CMAP 571, École Polytechnique, France, 2005.

[14] Y. G. Chen, Y. Giga, and S. Goto, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, Proc. Japan Acad. Ser. A Math. Sci., 65 (1989), pp. 207–210.

[15] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[16] I. Ekeland and R. Témam, *Convex Analysis and Variational Problems*, Classics Appl. Math. 28, SIAM, Philadelphia, 1999.

[17] L. C. Evans, *Convergence of an algorithm for mean curvature motion*, Indiana Univ. Math. J., 42 (1993), pp. 533–557.

[18] M.-H. Giga and Y. Giga, *Generalized motion by nonlocal curvature in the plane*, Arch. Ration. Mech. Anal., 159 (2001), pp. 295–333.

[19] Y. Giga and M. E. Gurtin, *A comparison theorem for crystalline evolution in the plane*, Quart. Appl. Math., 54 (1996), pp. 727–737.

[20] Y. Goto, K. Ishii, and T. Ogawa, *Method of the distance function to the Bence-Merriman-Osher algorithm for motion by mean curvature*, Commun. Pure Appl. Anal., 4 (2005), pp. 311–339.

[21] M. E. Gurtin, *Thermomechanics of Evolving Phase Boundaries in the Plane*, Oxford Mathematical Monographs, The Clarendon Press, Oxford University Press, New York, 1993.

[22] H. Ishii, *A generalization of the Bence, Merriman and Osher algorithm for motion by mean curvature*, in Curvature Flows and Related Topics (Levico, 1994), GAKUTO Internat. Ser. Math. Sci. Appl. 5, Gakkōtosho, Tokyo, 1995, pp. 111–127.

[23] H. Ishii, G. E. Pires, and P. E. Souganidis, *Threshold dynamics type approximation schemes for propagating fronts*, J. Math. Soc. Japan, 51 (1999), pp. 267–308.

[24] K. Ishii, *Optimal Rate of Convergence of the Bence-Merriman-Osher Algorithm for Motion by Mean Curvature*, preprint 16, available online at http://cpde.iac.rm.cnr.it (2004).

[25] F. Leoni, *Convergence of an approximation scheme for curvature-dependent motions of sets*, SIAM J. Numer. Anal., 39 (2001), pp. 1115–1131.

[26] B. Merriman, J. K. Bence, and S. J. Osher, *Diffusion generated motion by mean curvature*, in Computational Crystal Growers Workshop, J. E. Taylor, ed., Sel. Lectures Math., AMS, Providence, RI, 1992, pp. 73–83.

[27] R. T. Rockafellar, *Convex Analysis. Reprint of the* 1970 *Original*, Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, 1997.

[28] S. J. Ruuth, *Efficient algorithms for diffusion-generated motion by mean curvature*, J. Comput. Phys., 144 (1998), pp. 603–625.

[29] S. J. Ruuth and B. Merriman, *Convolution-generated motion and generalized Huygens' principles for interface motion*, SIAM J. Appl. Math., 60 (2000), pp. 868–890.

[30] S. J. Ruuth and B. Merriman, *Convolution-thresholding methods for interface motion*, J. Comput. Phys., 169 (2001), pp. 678–707.

[31] J. E. Taylor, *Motion of curves by crystalline curvature, including triple junctions and boundary points*, in Differential Geometry: Partial Differential Equations on Manifolds (Los Angeles, CA, 1990), Proc. Sympos. Pure Math. 54, AMS, Providence, RI, 1993, pp. 417–438.

[32] J. E. Taylor, *Surface motion due to crystalline surface energy gradient flows*, in Elliptic and Parabolic Methods in Geometry (Minneapolis, MN, 1994), A. K. Peters, Wellesley, MA, 1996, pp. 145–162.

[33] J. E. Taylor, J. W. Cahn, and C. A. Handwerker, *Geometric models of crystal growth*, Acta Metall., 40 (1992), pp. 1443–1474.

[34] L. Vivier, *Convergence of an approximation scheme for computing motions with curvature dependent velocities*, Differential Integral Equations, 13 (2000), pp. 1263–1288.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 1988–2024

DISSIPATIVE SYMMETRIZERS OF HYPERBOLIC PROBLEMS AND THEIR APPLICATIONS TO SHOCK WAVES AND CHARACTERISTIC DISCONTINUITIES*

YURI TRAKHININ†

Abstract. By introducing the notations of dissipative and strictly dissipative p-symmetrizers of initial-boundary-value problems for linear hyperbolic systems we formalize the dissipative integrals technique [A. Blokhin, Yu. Trakhinin, in *Handbook of Mathematical Fluid Dynamics*, Vol. 1, North-Holland, Amsterdam, 2002, pp. 545–652] applied earlier to shock waves and characteristic discontinuities for various concrete systems of conservation laws. This enables us to prove the local in time existence of shock-front solutions of an abstract symmetric system of hyperbolic conservation laws, provided that the corresponding constant coefficients linearized problem has a strictly dissipative p-symmetrizer. Our result does not, in particular, require the fulfillment of Majda's block structure condition. A p-symmetrizer is, in some sense, a "secondary" Friedrichs symmetrizer for the symmetric system for the vector of p-derivatives of unknown functions, and the structure of p-symmetrizer takes into account (if applicable) the set of divergent constraints for the original system. After applying a p-symmetrizer, which is in general a set of matrices and vectors, the boundary conditions for a resulting symmetric system are dissipative (or strictly dissipative). We give concrete examples of p-symmetrizers. Our main examples are strictly dissipative 2-symmetrizers for shock waves in gas dynamics and magnetohydrodynamics. A general procedure for constructing a p-symmetrizer does not however exist. But, if it was somehow constructed, then we do not need to test the Lopatinski condition that is often connected with insuperable technical difficulties. As an illustration, we refer to the author's recent result [Yu. Trakhinin, *Arch. Ration. Mech. Anal.*, 177 (2005), pp. 331–366] for compressible current-vortex sheets for which the construction of a dissipative 0-symmetrizer has first enabled the finding of sufficient conditions for their weak linearized stability.

Key words. symmetric hyperbolic systems, dissipative boundary conditions, multidimensional conservation laws, shock waves

AMS subject classifications. 35L50, 35L65, 35L67

DOI. 10.1137/050628003

1. Introduction: Initial-boundary-value problems for quasi-linear hyperbolic systems.

Consider a system of N *conservation laws*

$$\partial_t \boldsymbol{\mathcal{P}}^0(\mathbf{U}) + \sum_{j=1}^{n} \partial_j \boldsymbol{\mathcal{P}}^j(\mathbf{U}) = 0, \tag{1.1}$$

where $\boldsymbol{\mathcal{P}}^\alpha = \boldsymbol{\mathcal{P}}^\alpha(\mathbf{U}) = (\mathcal{P}_1^\alpha, \dots, \mathcal{P}_N^\alpha)$, $\mathbf{U} = \mathbf{U}(t, \mathbf{x}) = (u_1, \dots, u_N)$, $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$, $\partial_t := \partial/\partial t$, $\partial_j := \partial/\partial x_j$. With the notation

$$\operatorname{div} \mathbf{a} := \sum_{j=1}^{n} \partial_j a^j \qquad (\mathbf{a} = \mathbf{a}(t, \mathbf{x}) = (a^1, \dots, a^n) \quad \text{is a vector})$$

system (1.1) in componentwise form reads

$$\partial_t \mathcal{P}_i^0(\mathbf{U}) + \operatorname{div} \boldsymbol{\mathcal{P}}_i(\mathbf{U}) = 0, \quad i = \overline{1, N},$$

*Received by the editors March 29, 2005; accepted for publication (in revised form) October 25, 2005; published electronically March 24, 2006. This work was supported by EPSRC research grant GR/S96609/01.

http://www.siam.org/journals/sima/37-6/62800.html

†Sobolev Institute of Mathematics, Koptyuga pr.4, 630090 Novosibirsk, Russia (trakhin@math.nsc.ru).

where $\boldsymbol{\mathcal{P}}_i = \boldsymbol{\mathcal{P}}_i(\mathbf{U}) = (\mathcal{P}_i^1, \dots, \mathcal{P}_i^n)$. Assuming that the flux functions $\mathcal{P}_i^\alpha$ are smooth enough (in practice, they are usually C^∞), one can rewrite (1.1) as the quasi-linear system

$$B_0(\mathbf{U})\mathbf{U}_t + \sum_{j=1}^{n} B_j(\mathbf{U})\mathbf{U}_{x_j} = 0, \tag{1.2}$$

with $B_\alpha = (\partial \boldsymbol{\mathcal{P}}^\alpha / \partial \mathbf{U})$.

We will assume that system (1.1) may be supplemented (but not necessarily) by a set of K *divergent constraints*

$$\operatorname{div} \boldsymbol{\Psi}_j(\mathbf{U}) = 0, \quad j = \overline{1, K}, \tag{1.3}$$

where $\boldsymbol{\Psi}_j = \boldsymbol{\Psi}_j(\mathbf{U}) = (\Psi_j^1, \dots, \Psi_j^n)$. For example, for the system of gas dynamics one has no divergent constraints at all, whereas the system of magnetohydrodynamics (MHD) is supplemented by the sole ($K = 1$) divergent constraint $\operatorname{div} \mathbf{H} = 0$ (see, e.g., [28]). We also refer, for instance, to Landau's equations of superfluid [29] (see also [11]) which are supplemented by the tree divergent constrains $\nabla \times \mathbf{v}_{\mathrm{s}} = 0$ ($\mathbf{v}_{\mathrm{s}}$ is the superfluid velocity [29, 11]).

ASSUMPTION 1.1. *The divergent constraints* (1.3) *are the restrictions on the initial data for system* (1.1)*. That is, if* (1.3) *are satisfied initially, they hold for all* $t > 0$.

Remark 1.1. Assumption 1.1 is quite natural because it holds for all the physically relevant models: MHD, Landau's equations of superfluid, etc. (see, e.g., [11] for further examples). In practice, $(1.3)|_{t>0}$ is proved by applying the operator div to appropriate subsystems of (1.1) and taking into account $(1.3)|_{t=0}$. Of course, applying div requires $n \leq N$. The last assumption will be also made for other reasons (see section 3).

Sometimes, by an appropriate choice of the vector of unknowns $\mathbf{U}$ a concrete system of conservation laws can be written in the nonconservative form (1.2) with symmetric matrices B_α. For example, the systems of gas dynamics and MHD are written as symmetric systems for the vectors of unknowns $\mathbf{U} = (p, \mathbf{v}, S)$ and $\mathbf{U} = (p, \mathbf{v}, \mathbf{H}, S)$, respectively. At the same time, it is not always possible to guess an appropriate vector $\mathbf{U}$ for which a system of conservation laws, (1.1), can be rewritten as a symmetric quasi-linear system. But, as was first shown by Godunov [23, 24], system (1.1) can be always symmetrized if we know, a priori, an *additional* conservation law ("entropy" conservation)

$$\partial_t \Phi^0(\mathbf{U}) + \operatorname{div} \boldsymbol{\Phi}(\mathbf{U}) = 0,$$

with $\boldsymbol{\Phi} = \boldsymbol{\Phi}(\mathbf{U}) = (\Phi^1, \dots, \Phi^n)$, which holds on smooth solutions of (1.1). That is, one can find an invertible change of unknowns $\mathbf{U} \to \mathbf{Q}$ such that the system

$$A^0(\mathbf{Q})\mathbf{Q}_t + \sum_{j=1}^{n} A^j(\mathbf{Q})\mathbf{Q}_{x_j} = 0 \tag{1.4}$$

is symmetric: $A^\alpha = (A^\alpha)^*$, where

$$A^\alpha = \left(\frac{\partial \boldsymbol{\mathcal{P}}^\alpha}{\partial \mathbf{Q}} \right) = B_\alpha(\mathbf{U}(\mathbf{Q})) \left(\frac{\partial \mathbf{U}}{\partial \mathbf{Q}} \right),$$

i.e., $A^\alpha = B_\alpha J^{-1}$, with $J = J(\mathbf{Q}) = (\partial\mathbf{Q}/\partial\mathbf{U})$. It should be noted that if system (1.1) is accompanied by a set of divergent constraints (1.3), then these constrains should be generically taken into account under *Godunov's symmetrization*. For the process of symmetrization itself we refer to [13] and references therein (note that, in particular, $\mathbf{Q} = (\partial\Phi^0/\partial\boldsymbol{\mathcal{P}}^0)$).

It is worth noting that the symmetric system (1.4) can be rewritten as a quasi-linear system for the original vector of unknowns $\mathbf{U}$ that is again symmetric. Indeed, (1.4) clearly implies the system

$$A_0(\mathbf{U})\mathbf{U}_t + \sum_{j=1}^{n} A_j(\mathbf{U})\mathbf{U}_{x_j} = 0, \tag{1.5}$$

where the matrices $A_\alpha = A_\alpha(\mathbf{U}) := J^*A^\alpha J = J^*B_\alpha$ are symmetric. Thus, the matrix

$$S = S(\mathbf{U}) = J^* = \left(\frac{\partial\mathbf{Q}}{\partial\mathbf{U}}\right)^*$$

is the one (called *Friedrichs symmetrizer*) that symmetrizes system (1.2):

$$A_\alpha = SB_\alpha = A_\alpha^*.$$

Recall that the quasi-linear symmetric system (1.5) is *symmetric hyperbolic* in the sense of Friedrichs [22] if $A_0(\mathbf{U}) > 0$ (or $A_0(\mathbf{U}) < 0$ if we multiply (1.5) by -1).

The main requirement for the local in time well-posedness of the Cauchy problem for a quasi-linear system of conservation laws is the hyperbolicity condition that is easily verified for symmetric systems. The local existence theorem for the Cauchy problem for symmetric hyperbolic systems was independently proved by Vol'pert and Khudyaev [47], Lax [30], and Kato [26] (see also [34]). In contrast with the Cauchy problem, the conditions for well-posedness of initial-boundary-value problems for hyperbolic systems, in the generic case, cannot be easily found even for linearized problems with constant coefficients.

1.1. Standard boundary conditions. Let us first consider quasi-linear initial-boundary-value problems with standard boundary conditions [40]:

$$L(\mathbf{U})\mathbf{U} = 0 \quad \text{in } [0,T]\times\mathbb{R}^n_+, \tag{1.6a}$$

$$M(t,\mathbf{x}',\mathbf{U})\mathbf{U} = 0 \quad \text{on } [0,T]\times\{x_1 = 0\}\times\mathbb{R}^{n-1}, \tag{1.6b}$$

$$\mathbf{U}|_{t=0} = \mathbf{U}_0 \quad \text{in } \mathbb{R}^n_+, \tag{1.6c}$$

where $L = L(\mathbf{U}) = A_0(\mathbf{U})\partial_t + \sum_{j=1}^n A_j(\mathbf{U})\partial_j$, and system (1.6a) is supposed to be symmetric hyperbolic, and the matrix M is a $d\times N$ matrix. Here and below $\mathbb{R}^n_\pm = \{x_1 \gtrless 0,\ \mathbf{x}' \in \mathbb{R}^{n-1}\}$, $\mathbf{x}' = (x_2,\dots,x_n)$. Without loss of generality we consider the problem in a half-space because the case of a smooth bounded domain Ω is reduced, in some sense, to problem (1.6) by a finite partition of unity subordinated to an open covering of $\overline{\Omega}$.

To prove a local (in time) existence theorem for problem (1.6) we should consider the following linear problem associated to (1.6):

$$L(\widehat{\mathbf{U}})\mathbf{U} = \mathbf{f} \quad \text{in } [0,T]\times\mathbb{R}^n_+, \tag{1.7a}$$

$$M(t,\mathbf{x}',\widehat{\mathbf{U}})\mathbf{U} = \mathbf{g} \quad \text{on } [0,T]\times\{x_1 = 0\}\times\mathbb{R}^{n-1}, \tag{1.7b}$$

$$\mathbf{U}|_{t=0} = \mathbf{U}_0 \quad \text{in } \mathbb{R}^n_+, \tag{1.7c}$$

where $\widehat{\mathbf{U}}$ is a given vector-function. Here we introduce the source terms $\mathbf{f}(t,\mathbf{x})$ and

$\mathbf{g}(t,\mathbf{x}')$ to make the interior equations and the boundary conditions inhomogeneous (this is needed to attack the nonlinear problem).

Recall that the boundary conditions (1.7b) are called *dissipative* if

$$-(A_1\mathbf{U},\mathbf{U})|_{x_1=0} \geq 0 \qquad \forall\ \mathbf{U} \in \ker M, \tag{1.8}$$

where $-A_1 = -A_1(\widehat{\mathbf{U}})$ is the boundary matrix, $M = M(t,\mathbf{x}',\widehat{\mathbf{U}})$. They are *strictly dissipative* if there exist a fixed constant $\gamma > 0$ such that

$$-(A_1\mathbf{U},\mathbf{U})|_{x_1=0} \geq \gamma|\mathbf{V}|^2 \qquad \forall\ \mathbf{U} \in \ker M, \tag{1.9}$$

where $\mathbf{V}$ is the "noncharacteristic part" of the trace $\mathbf{U}|_{x_1=0}$, i.e., the projection of $\mathbf{U}|_{x_1=0}$ orthogonal to $\ker A_1|_{x_1=0}$ ($\mathbf{V} = \mathbf{U}$ for the case of noncharacteristic boundary, i.e., when $\det A_1|_{x_1=0} \neq 0$). Recall also that the boundary conditions (1.7b) are called *maximally dissipative* if they are dissipative and

$$\dim\ker M = \#\ \text{nonpositive eigenvalues of } A_1|_{x_1=0} \text{ counting multiplicity} \tag{1.10}$$

(we use the definition from [39]). Property (1.10) means that the hyperbolic system (1.7a) has the correct number of boundary conditions in (1.7b), i.e.,

$$d = \#\ \text{positive eigenvalues of } A_1|_{x_1=0} \text{ counting multiplicity.}$$

In the following we always assume that the number of boundary conditions is correct and we therefore drop the word "maximally" when speaking about dissipative or strictly dissipative boundary conditions.

For the inhomogenous boundary conditions (1.7b) inequality (1.9) implies that there exists a fixed constant $\delta > 0$ such that

$$-(A_1\mathbf{U},\mathbf{U})|_{x_1=0} \geq \delta|\mathbf{V}|^2 - \delta^{-1}|\mathbf{g}|^2 \tag{1.11}$$

for all $\mathbf{U}$ satisfying (1.7b). To deduce inequality (1.11) from (1.9) it needs to reduce system (1.7a) to the form [35], $\mathcal{A}_0\mathbf{W}_t + \sum_{j=1}^{n}\mathcal{A}_j\mathbf{W}_{x_j} = \dots$, where $\mathbf{U} = \mathcal{T}\mathbf{W}$, $\mathcal{A}_\alpha = \mathcal{T}^*A_\alpha\mathcal{T}$, $\mathcal{A}_1 = \operatorname{diag}(D_1, -D_2, 0)$, $D_i > 0$, $(\mathcal{A}_1\mathbf{W},\mathbf{W}) = (D_1\mathbf{W}_1,\mathbf{W}_1) - (D_2\mathbf{W}_2,\mathbf{W}_2)$, and the boundary conditions (1.7b) are supposed to be rewritten in the form $\mathbf{W}_1 = S\mathbf{W}_2 + \tilde{\mathbf{g}}$. Analogous simple arguments show that the dissipativity hypothesis (1.8) implies that there is a matrix $B = B(t,\mathbf{x}')$ such that

$$-(A_1\mathbf{U},\mathbf{U})|_{x_1=0} \geq (B\mathbf{g},\mathbf{V}) - \gamma|\mathbf{g}|^2 \tag{1.12}$$

for all $\mathbf{U}$ satisfying (1.7b), where γ is a constant, the matrix B can be, in principle, explicitly written out (it depends on $\mathcal{T}$, etc.).

If the boundary conditions (1.6b) are *linear*, i.e., $M = M(t,\mathbf{x}')$, then the boundary conditions for the associated linear problem (1.7) can be considered to be homogenous ($\mathbf{g}$=0) and the dissipativity hypothesis (1.8) is quite enough to prove a local existence theorem for problem (1.6) by standard fixed-point argument. In this case, for the linear problem (1.7) the basic estimate following from assumption (1.8) reads [40]

$$\|\mathbf{U}(t)\|_{L_2(\mathbb{R}^n_+)} \leq C\left\{\|\mathbf{U}_0\|_{L_2(\mathbb{R}^n_+)} + \|\mathbf{f}\|_{L_2([0,T]\times\mathbb{R}^n_+)}\right\},$$

where $C = C(T)$ is a positive constant independent of the initial data and the source terms. For linear boundary conditions on a noncharacteristic boundary the local W_2^s-existence theorem for problem (1.6) was proved by Schochet [41] (see Appendix A of [41]), where $s \geq [n/2]+2$ as for the Cauchy problem [47, 30, 26]. The case of characteristic boundary was considered by Secchi [42] (he has proved a local existence theorem in anisotropic weighted Sobolev spaces [42]).

However, if the boundary conditions in (1.6b) are *nonlinear*, i.e., M depends on $\mathbf{U}$, then one has to consider inhomogenous boundary conditions in the linear problem (1.7) (even if the original conditions (1.6b) were homogenous). The usual way (see, e.g., [40]) to deal with inhomogenous boundary conditions suggests to subtract from the solution a more regular function satisfying the boundary conditions, and reduce problem (1.7) to one with homogenous boundary conditions. But, such a way leads to the loss of "1/2 derivative" from $\mathbf{g}$ (see [40]). That is, the dissipativity hypothesis is already not enough to achieve a nonlinear local existence result by standard iterations.

At the same time, when the boundary conditions in (1.6b) are nonlinear, but conditions (1.7b) are strictly dissipative, using inequality (1.11), we can easily deduce the basic a priori estimate (with no loss of derivatives)

$$\|\mathbf{U}(t)\|_{L_2(\mathbb{R}^n_+)} \leq C\Big\{\|\mathbf{U}_0\|_{L_2(\mathbb{R}^n_+)} + \|\mathbf{f}\|_{L_2([0,T]\times\mathbb{R}^n_+)} + \|\mathbf{g}\|_{L_2([0,T]\times\mathbb{R}^{n-1})}\Big\} \tag{1.13}$$

and its higher order counterparts (see Appendix A in [43]). This enables one to prove a local existence theorem for problem (1.6) by standard fixed-point argument. The proof for the general problem (1.6) has not been written out somewhere, but we can refer to [43] for a concrete example of problem (1.6). Note that the well-posedness of the linear problem (1.7) with $\mathbf{g} \neq 0$ was proved in Appendix A of [43].

That is, when conditions (1.7b) are inhomogenous but strictly dissipative, an unpleasant loss of derivatives is avoided by a direct approach to the original problem (1.7) with inhomogenous boundary conditions (for problem (1.7) this was done in Appendix A of [43]; see also section 4 for shock waves). If, however, conditions (1.7b) are just dissipative (but not strictly dissipative), such a direct approach to the linear problem (1.7) only enables us to obtain the a priori estimate with the loss of one derivative from $\mathbf{g}$:

$$\|\mathbf{U}(t)\|_{W_2^1(\mathbb{R}^n_+)} \leq C\Big\{\|\mathbf{U}_0\|_{W_2^1(\mathbb{R}^n_+)} + \|\mathbf{f}\|_{W_2^1([0,T]\times\mathbb{R}^n_+)} + \|\mathbf{g}\|_{W_2^2([0,T]\times\mathbb{R}^{n-1})}\Big\}. \tag{1.14}$$

Here we suppose that the boundary is noncharacteristic. For the case of characteristic boundary the counterpart of (1.14) indicates a loss of control on derivatives in the normal direction [39, 42, 16]. To deduce estimate (1.14) we should differentiate problem (1.7) with respect to t and $\mathbf{x}'$. Then, we take into account inequality (1.12) and integrate by parts the boundary integral (see section 3 for the case of Rankine–Hugoniot boundary conditions).

Remark 1.2. For the case of linearized Rankine–Hugoniot boundary conditions a priori estimates with loss of derivatives were proved for shock waves and characteristic discontinuities by Coulombel [17], Coulombel and Secchi [19], and the author [45] (in [45] the a priori estimates are formally with no loss of derivatives because the boundary conditions were supposed to be homogenous). For the case of standard boundary conditions, cf. (1.7b), the a priori estimate (1.14) is a basic estimate, and it is not difficult to obtain an estimate for higher order derivatives. But, of course, it will be also with the loss of one derivative from $\mathbf{g}$. This precludes one from using fixed-point argument to prove a local existence theorem. It seems that the only way to overcome the difficulty connected with the loss of derivatives phenomenon is the employment of the Nash–Moser technique (for hyperbolic problems see [2, 31, 21]). Recently, the Nash–Moser method was successfully used by Coulombel and Secchi

[20] for two-dimensional (2D) compressible vortex sheets as well as for weakly stable shock waves (see also preparatory results in this direction in [17, 18, 19]). There is also a great hope to achieve a nonlinear local existence result for current-vortex sheets [45] by using Nash–Moser iterations.

If the boundary conditions (1.7b) are not dissipative, this does not, of course, mean that problem (1.7) is ill-posed. An alternative to the energy method was first suggested by Kreiss [27] for the strictly hyperbolic case. Kreiss has proved that problem (1.7) with constant ("frozen") coefficients obeys an a priori $L_{2,\eta}$-estimate with no loss of derivatives if and only if the boundary conditions satisfy the *uniform Lopatinski condition* [27] ($\|\cdot\|_{L_{2,\eta}} = \|e^{-\eta t}(\cdot)\|_{L_2}$, and $\eta > 0$ is sufficiently large). This estimate follows from a symmetrizer construction (*Kreiss symmetrizer*) and is carried over variable coefficients by using pseudodifferential calculus. That is, the uniform Lopatinski condition is the sharp algebraic criterion of *strong* well-posedness (well-posedness "with no loss of derivatives").

Later Kreiss' symmetrizer analysis was extended by Agranovich [1], Majda and Osher [35], and Majda [32] to hyperbolic systems satisfying a so-called block structure condition, which holds in particular for hyperbolic symmetrizable systems with constant multiplicities [1, 36]. In [32] Majda extends the Kreiss theory to the case of Rankine–Hugoniot boundary conditions. Majda's approach has been then improved by Métivier [37] by using paradifferential calculus of Bony (see also discussion in section 4). And recently Métivier and Zumbrun [38] have extended the Kreiss–Majda theory to a class of hyperbolic symmetrizable systems with characteristics of variable multiplicities. These systems at points of variable multiplicity should satisfy some conditions [38] which hold in particular for the MHD system. Thus, if the symmetric hyperbolic system (1.6a) meets either the Agronovich–Majda–Osher block structure condition [1, 35] or the conditions of Métivier and Zumbrun [38], then the linear problem (1.7) is strongly well-posed, provided that the boundary conditions (1.7b) satisfy the uniform Lopatinski condition. In this case a local existence theorem for the nonlinear problem (1.6) can be proved by analogy with the proofs in [33, 37].

At the same time, it should be noted that in practice the algebraic criterion given by the Lopatinski condition often cannot be tested analytically. Sometimes one succeeds to check it numerically (see [44]), but, frequently, numerical calculations can give only a very rough description of the condition for weak/strong well-posedness because either the domain of parameters for the constant coefficients linearized problem is unbounded or the number of these parameters is too big. Usually this happens for the case of Rankine–Hugoniot boundary conditions (see below), i.e., for shock waves or characteristic discontinuities. For example, such a situation takes place for compressible current-vortex sheets [45]. In this connection, there is no sense to reject at once the energy method as soon as the boundary conditions are not dissipative. The main purpose of the present paper is to formalize the so-called *dissipative integrals technique* (see [13]), which is a kind of "higher order" extension of the usual energy method. Especially, this technique turned out to be effective for shock waves in various hyperbolic models: gas dynamics, Landau's equations of superfluid, MHD, radiation hydrodynamics, etc. (see [13] and references therein). Let us now go on to the case of Rankine–Hugoniot boundary conditions.

1.2. Rankine–Hugoniot boundary conditions. Consider system (1.1) in the whole space $\mathbb{R}^n$. Let

$$\Gamma(t) = \{x_1 - f(t, \mathbf{x}') = 0\}$$

be a smooth hypersurface in $[0,T]\times\mathbb{R}^n$. We assume that $\Gamma(t)$ is a surface of strong discontinuity for solutions of (1.1). Let $\mathbf{U}(t,\mathbf{x})$ be a classical solution of (1.1) on either side of Γ. As is known, $\mathbf{U}$ is a weak solution of (1.1) if and only if the Rankine–Hugoniot jump conditions hold at each point of Γ:

$$f_t[\boldsymbol{\mathcal{P}}^0(\mathbf{U})]+\sum_{k=2}^{n} f_{x_k}[\boldsymbol{\mathcal{P}}^k(\mathbf{U})]-[\boldsymbol{\mathcal{P}}^1(\mathbf{U})]=0, \tag{1.15}$$

where $[a]=a^+-a^-=a|_{x_1-f(t,\mathbf{x}')=+0}-a|_{x_1-f(t,\mathbf{x}')=-0}$.

It should be noted that the initial-boundary-value problem for system (1.5) in the domains $\Omega^\pm(t):=\{x_1\gtrless f(t,\mathbf{x}')\}$ with the boundary conditions (1.15) on the hypersurface $\Gamma(t)$ is a free-boundary-value problem. Indeed, the function $f(t,\mathbf{x}')$ defining Γ is one of the unknowns of problem (1.5), (1.15) with the corresponding initial data

$$f|_{t=0}=f_0 \quad \text{in } \mathbb{R}^{n-1}, \qquad \mathbf{U}|_{t=0}=\mathbf{U}_0 \quad \text{in } \Omega^+(0)\cup\Omega^-(0). \tag{1.16}$$

To work in fixed domains instead of the domains $\Omega^\pm(t)$ we make the following change of variables:

$$\widetilde{t}=t,\quad \widetilde{x}_1=x_1-f(t,\mathbf{x}'),\quad \widetilde{\mathbf{x}}'=\mathbf{x}'.$$

Then, $\widetilde{\mathbf{U}}(\widetilde{t},\widetilde{\mathbf{x}}):=\mathbf{U}(t,\mathbf{x})$ is a smooth vector-function for $\widetilde{\mathbf{x}}\in\mathbb{R}^n_\pm$, and the initial-boundary-value problem (1.5), (1.15), (1.16) is reduced to the following problem (we omit tildes to simplify the notation):

$$L(\mathbf{U},\mathbf{F})\mathbf{U}=0 \quad \text{in } [0,T]\times(\mathbb{R}^n_+\cup\mathbb{R}^n_-), \tag{1.17a}$$

$$B(\mathbf{U}^+,\mathbf{U}^-)\mathbf{F}-[\boldsymbol{\mathcal{P}}^1(\mathbf{U})]=0 \quad \text{on } [0,T]\times\{x_1=0\}\times\mathbb{R}^{n-1}, \tag{1.17b}$$

$$\mathbf{U}|_{t=0}=\mathbf{U}_0 \quad \text{in } \mathbb{R}^n_+\cup\mathbb{R}^n_-, \qquad f|_{t=0}=f_0 \quad \text{in } \mathbb{R}^{n-1}. \tag{1.17c}$$

Here

$$L=L(\mathbf{U},\mathbf{F})=A_0(\mathbf{U})\partial_t+A_\nu(\mathbf{U},\mathbf{F})\partial_1+\sum_{k=2}^{n}A_k(\mathbf{U})\partial_k,$$
$$\mathbf{F}=\mathbf{F}(t,\mathbf{x}')=(f_t,\mathbf{F}'),\quad \mathbf{F}'=\mathbf{F}'(t,\mathbf{x}')=\nabla_{\mathbf{x}'}f,\quad \nabla_{\mathbf{x}'}=(\partial_2,\dots,\partial_n),$$
$$A_\nu=A_\nu(\mathbf{U},\mathbf{F})=\sum_{\alpha=0}^{n}\nu_\alpha A_\alpha=A_1(\mathbf{U})-f_tA_0(\mathbf{U})-\sum_{k=2}^{n}f_{x_k}A_k(\mathbf{U}),$$

$\boldsymbol{\nu}=(\nu_0,\dots,\nu_n)=(-f_t,\mathbf{N})$ and $\mathbf{N}=(1,-\mathbf{F}')$ are, respectively, the space-time and space normal vectors to $\Gamma(t)$. The matrix $B=B(\mathbf{U}^+,\mathbf{U}^-)$ is of order $N\times n$ and determined from the relation

$$B(\mathbf{U}^+,\mathbf{U}^-)\mathbf{F}=f_t[\boldsymbol{\mathcal{P}}^0(\mathbf{U})]+\sum_{k=2}^{n}f_{x_k}[\boldsymbol{\mathcal{P}}^k(\mathbf{U})],\quad \mathbf{U}^\pm:=\mathbf{U}|_{x_1=\pm0}.$$

After straightening of variables above the divergent constraints (1.3) take the form

$$\operatorname{div}\boldsymbol{\psi}_j(\mathbf{U},\mathbf{F}')=0 \quad \text{in } [0,T]\times(\mathbb{R}^n_+\cup\mathbb{R}^n_-), \qquad j=\overline{1,K}, \tag{1.18}$$

where $\boldsymbol{\psi}_j = \boldsymbol{\psi}_j(\mathbf{U},\mathbf{F}') = ((\boldsymbol{\Psi}_j,\mathbf{N}),\Psi_j^2,\ldots,\Psi_j^n)$. Moreover, solutions of problem (1.17) should satisfy the jump conditions

$$(1.19)\qquad [(\boldsymbol{\Psi}_j(\mathbf{U}),\mathbf{N})] = 0 \quad \text{on } [0,T]\times\{x_1=0\}\times\mathbb{R}^{n-1}, \quad j=\overline{1,K},$$

coming from (1.3). At the same time, taking into account Assumption 1.1 and Remark 1.1, it is natural to make the following assumption that should be true for all the physically relevant models.

Assumption 1.2. *The divergent constraints* (1.18) *are the restrictions on the initial data* (1.17c)*, i.e., if* (1.18) *are satisfied initially, they hold for all* $t>0$*. Equations* (1.17b) *and* (1.19) *form a system of* N *independent boundary conditions.*

To clarify Assumption 1.2 we can refer, for example, to MHD. Namely, in MHD the jump condition $f_t[(\mathbf{H},\mathbf{N})]=0$ contained in the corresponding main system, (1.17b), is implied by the equation $[(\mathbf{H},\mathbf{N})]=0$ coming from the divergent constraint $\operatorname{div}\mathbf{H}=0$. That is, in MHD the number of independent Rankine–Hugoniot boundary conditions is equal to the number of conservation laws.

To prove the existence of solutions with a surface of strong discontinuity $\Gamma(t)$ for the system of hyperbolic conservation laws (1.1) one needs to reply to the following question: does there exist a solution $(\mathbf{U},f)$ to problem (1.17) at least locally in time? The necessary (but not sufficient) condition for this is that the hyperbolic problem (1.17) has the correct number of boundary conditions in (1.17b). In this connection, it should be noted that, in contrast with the standard boundary conditions (1.6b), one of the conditions in (1.17b) is needed for determining the function $f(t,\mathbf{x}')$. For noncharacteristic discontinuities, i.e., *shock waves,* the plane $x_1=0$ is not a characteristic boundary for system (1.17). That is, the boundary matrix A_ν is nonsingular at $x_1=0$: $\det A_\nu^\pm \neq 0$, where $A_\nu^\pm = A_\nu|_{x_1=\pm 0}$. As is known, for shock waves the correct number of Rankine–Hugoniot boundary conditions is guaranteed by the *Lax shock conditions.* They can be conveniently written in terms of the eigenvalues of the boundary matrix A_ν:

$$(1.20)\qquad \lambda_k(A_\nu^+) < 0 < \lambda_k(A_\nu^-), \quad \lambda_{k-1}(A_\nu^-) < 0 < \lambda_{k+1}(A_\nu^+),$$

where $\lambda_i(A_\nu^\pm)$ ($i=\overline{1,N}$, $\lambda_1\le\ldots\le\lambda_N$) are the eigenvalues of the matrices $A_\nu^\pm$ and k is a fixed integer number, $1\le k\le N$ (an associated shock wave is called k-shock), $\lambda_0 := -\infty$, $\lambda_{N+1} := +\infty$.

Let $(\widehat{\mathbf{U}}(t,\mathbf{x}),\hat{f}(t,\mathbf{x}'))$ be a given vector-function, where $\widehat{\mathbf{U}}$ is supposed to be smooth for $\mathbf{x}\in\mathbb{R}^n_\pm$. Then the linearization of (1.17) results in the following variable coefficients problem for determining small perturbations $(\delta\mathbf{U},\delta f)$ (below we drop δ):

$$(1.21\text{a})\qquad L(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\mathbf{U} + \widehat{C}\mathbf{U} = \big\{L(\widehat{\mathbf{U}},\widehat{\mathbf{F}})f\big\}\widehat{\mathbf{U}}_{x_1} \quad \text{in } [0,T]\times(\mathbb{R}^n_+\cup\mathbb{R}^n_-),$$

$$(1.21\text{b})\qquad B(\widehat{\mathbf{U}}^+,\widehat{\mathbf{U}}^-)\mathbf{F} - \big[S^{-1}(\widehat{\mathbf{U}})A_\nu(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\mathbf{U}\big] = 0 \quad \text{on } [0,T]\times\{x_1=0\}\times\mathbb{R}^{n-1},$$

and the initial data for the perturbation $(\mathbf{U},f)$ coincide with (1.17c). Here, $\widehat{\mathbf{F}} = (\hat{f}_t,\widehat{\mathbf{F}}')$, $\widehat{\mathbf{F}}' = \nabla_{\mathbf{x}'}\hat{f}$, the matrix $\widehat{C} = \widehat{C}(\widehat{\mathbf{U}},\widehat{\mathbf{U}}_t,\nabla\widehat{\mathbf{U}},\widehat{\mathbf{F}})$ is determined as follows:

$$\widehat{C}\mathbf{U} = (\mathbf{U},\nabla_u A_0(\widehat{\mathbf{U}}))\widehat{\mathbf{U}}_t + (\mathbf{U},\nabla_u A_\nu(\widehat{\mathbf{U}},\widehat{\mathbf{F}}))\widehat{\mathbf{U}}_{x_1} + \sum_{k=2}^n (\mathbf{U},\nabla_u A_k(\widehat{\mathbf{U}}))\widehat{\mathbf{U}}_{x_k},$$

$(\mathbf{U},\nabla_u) := \sum_{i=1}^N u_i\partial/\partial u_i$. Recall that $S(\mathbf{U})$ is the Friedrichs symmetrizer mentioned above, i.e., $S^{-1}A_\nu = B_\nu$, where $B_\nu = \sum_{\alpha=0}^n \nu_\alpha B_\alpha$. Problem (1.21a), (1.17c) is the *genuine linearization* of (1.17) in the sense that we keep all the lower order terms in (1.21a).

It should be noted that the differential operator in system (1.21a) is a first order operator in f. This fact can give some trouble in the application of the energy method to (1.21a). To avoid this difficulty we make the change of unknowns (see [2])

$$\bar{\mathbf{U}} = \mathbf{U} - f\widehat{\mathbf{U}}_{x_1}. \tag{1.22}$$

In terms of the "good unknown" (1.22) problem (1.21a) takes the form

$$L(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\bar{\mathbf{U}} + \widehat{C}\bar{\mathbf{U}} + f\partial_1\{L(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\widehat{\mathbf{U}}\} = 0 \quad \text{in } [0,T]\times(\mathbb{R}^n_+\cup\mathbb{R}^n_-), \tag{1.23a}$$

$$\begin{aligned} B(\widehat{\mathbf{U}}^+,\widehat{\mathbf{U}}^-)\mathbf{F} &- \left[S^{-1}(\widehat{\mathbf{U}})A_\nu(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\bar{\mathbf{U}}\right] \\ &= -f\left[S^{-1}(\widehat{\mathbf{U}})A_\nu(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\widehat{\mathbf{U}}_{x_1}\right] \quad \text{on } [0,T]\times\{x_1=0\}\times\mathbb{R}^{n-1}. \end{aligned} \tag{1.23b}$$

Actually, to prove a local existence theorem for the nonlinear problem (1.17) with strictly dissipative boundary conditions or having a strictly dissipative p-symmetrizer (see section 2), it is not necessary to consider the genuine linearization (1.21a). It is enough to keep only the principal part of the linearized equations, i.e., one can drop the lower order terms in (1.23). At the same time, for the case when the loss of derivatives takes place it needs to perform genuine linearization ("to find a differential") for the purpose of a possible use of the Nash–Moser method (see discussion in Remark 1.2). The linearized equations associated to (1.17a), (1.17b) and obtained by dropping the lower order terms in (1.23) read:

$$L(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\mathbf{U} = \mathbf{f} \quad \text{in } [0,T]\times(\mathbb{R}^n_+\cup\mathbb{R}^n_-), \tag{1.24a}$$

$$B(\widehat{\mathbf{U}}^+,\widehat{\mathbf{U}}^-)\mathbf{F} - \left[S^{-1}(\widehat{\mathbf{U}})A_\nu(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\mathbf{U}\right] = \mathbf{g} \quad \text{on } [0,T]\times\{x_1=0\}\times\mathbb{R}^{n-1}. \tag{1.24b}$$

Here we introduce the source terms $\mathbf{f}(t,\mathbf{x})$ and $\mathbf{g}(t,\mathbf{x}')$, where $\mathbf{f}(t,\mathbf{x}) = \mathbf{f}^\pm(t,\mathbf{x})$ for $\mathbf{x}\in\mathbb{R}^n_\pm$.

For problem (1.24) the definitions of dissipative and strictly dissipative boundary conditions are analogous to those in (1.12) and (1.11). In particular, the homogenous boundary conditions (1.24b) ($\mathbf{g}=0$) are dissipative if $-[(A_\nu\mathbf{U},\mathbf{U})]|_{x_1=0}\geq 0$ for all $\mathbf{U}$ satisfying (1.24b). Assuming that the front can be eliminated, i.e., the vector-function $\mathbf{F}$ can be expressed through $\mathbf{U}^+$, $\mathbf{U}^-$, and $\mathbf{g}$ (see section 3 for more details), for shock waves the counterparts of estimates (1.14) and (1.13) are the following:

$$\begin{aligned} \|f\|_{W_2^1([0,T]\times\mathbb{R}^{n-1})} + \sum_\pm \|\mathbf{U}(t)\|_{W_2^1(\mathbb{R}^n_\pm)} \leq C\Bigg\{ &\|f_0\|_{W_2^1(\mathbb{R}^{n-1})} \\ + \sum_\pm \Big\{\|\mathbf{U}_0\|_{W_2^1(\mathbb{R}^n_\pm)} + \|\mathbf{f}^\pm\|_{W_2^1([0,T]\times\mathbb{R}^n_\pm)}\Big\} &+ \|\mathbf{g}\|_{W_2^2([0,T]\times\mathbb{R}^{n-1})}\Bigg\} \end{aligned} \tag{1.25}$$

(for dissipative boundary conditions),

$$\begin{aligned} \|f\|_{W_2^1([0,T]\times\mathbb{R}^{n-1})} + \sum_\pm \|\mathbf{U}(t)\|_{L_2(\mathbb{R}^n_\pm)} \leq C\Bigg\{ &\|f_0\|_{W_2^1(\mathbb{R}^{n-1})} \\ + \sum_\pm \Big\{\|\mathbf{U}_0\|_{L_2(\mathbb{R}^n_\pm)} + \|\mathbf{f}^\pm\|_{L_2([0,T]\times\mathbb{R}^n_\pm)}\Big\} &+ \|\mathbf{g}\|_{L_2([0,T]\times\mathbb{R}^{n-1})}\Bigg\} \end{aligned} \tag{1.26}$$

(for strictly dissipative boundary conditions).

Since the original nonlinear problem (1.17) is a reduced free-boundary-value problem, we should gain the "additional derivative" for the front perturbation f (cf. (1.26)) to use then fixed-point argument. In this sense the a priori estimate (1.25) indicates the loss of one derivative not only from the source term $\mathbf{g}$ but also from the front f. At the same time, Rankine–Hugoniot boundary conditions are usually not dissipative (not to mention strict dissipativity). Nevertheless, the energy method can be still applied to problem (1.17) under certain circumstances. The main idea is to obtain from the linearized problem a problem for higher order derivatives of $\mathbf{U}$ so that this problem has dissipative (or strictly dissipative) boundary conditions. Such an idea was first realized by Blokhin [7, 9] for shock waves in gas dynamics. In the next section we formalize this idea by introducing the notations of dissipative and strictly dissipative p-symmetrizers and give concrete examples of p-symmetrizers. In section 3 we consider the case of constant coefficients and deduce a priori estimates for Lax shock waves whose linearized problems have a p-symmetrizer. In section 4 we carry these estimates over variable coefficients and outline the proof of the local existence theorem for the original nonlinear problem (1.17), provided that the corresponding constant coefficients linearized problem has a strictly dissipative p-symmetrizer. For all the linear results obtained earlier by the dissipative integrals technique for shock waves in various concrete models (see [13] and references therein) this enables one to conclude the local existence of shock-front solutions of the corresponding nonlinear systems. Eventually, in section 5 we make concluding remarks and discuss open problems.

2. Dissipative *p*-symmetrizers: Definition and examples. We introduce the notations of dissipative and strictly dissipative p-symmetrizers for linear hyperbolic initial-boundary-value problems with constant coefficients. We give the definition of these notations for the case of Rankine–Hugoniot boundary conditions. The corresponding definition for the earthier case of standard boundary conditions (cf. (1.6b)) is given analogously and does not need a separate treatment.

The constant coefficients linearized problem for planar discontinuities is of decisive importance for the subsequent variable coefficients and nonlinear analysis. For planar discontinuities $\hat{f}(t,\mathbf{x}')$ is a linear function:

$$\hat{f}(t,\mathbf{x}') = \sigma t + (\boldsymbol{\sigma}',\mathbf{x}'), \quad \boldsymbol{\sigma} = (\sigma,\boldsymbol{\sigma}') \in \mathbb{R}^n. \tag{2.1}$$

For the case of a piecewise constant solution,

$$\widehat{\mathbf{U}} = \begin{cases} \widehat{\mathbf{U}}^+, & x_1 > \sigma t + (\boldsymbol{\sigma}',\mathbf{x}'), \\ \widehat{\mathbf{U}}^-, & x_1 < \sigma t + (\boldsymbol{\sigma}',\mathbf{x}'), \end{cases}$$

equations (1.24) have constant ("frozen") coefficients:

$$L(\widehat{\mathbf{U}}^{\pm},\boldsymbol{\sigma})\mathbf{U} = \mathbf{f}^{\pm} \quad \text{for } \mathbf{x} \in \mathbb{R}^n_{\pm}, \tag{2.2a}$$

$$\widehat{B}\mathbf{F} - \big[\widehat{S}^{-1}\widehat{A}_\nu\mathbf{U}\big] = \mathbf{g} \quad \text{for } x_1 = 0, \tag{2.2b}$$

where $\widehat{B} = B(\widehat{\mathbf{U}}^+,\widehat{\mathbf{U}}^-)$, $\widehat{S}^{\pm} = S(\widehat{\mathbf{U}}^{\pm})$, $\widehat{A}^{\pm}_\nu = A_\nu(\widehat{\mathbf{U}}^{\pm},\boldsymbol{\sigma})$, $\widehat{A}^{\pm}_\alpha = A_\alpha(\widehat{\mathbf{U}}^{\pm})$ are constant coefficients matrices, and $\big[\widehat{S}^{-1}\widehat{A}_\nu\mathbf{U}\big] = (\widehat{S}^+)^{-1}\widehat{A}^+_\nu\mathbf{U}^+ - (\widehat{S}^-)^{-1}\widehat{A}^-_\nu\mathbf{U}^-$, etc.

ASSUMPTION 2.1. *The functions* $\Psi^i_j(\mathbf{U})$ *in* (1.3) *are linear.*

Assumption 2.1 is satisfied for most examples of constrained hyperbolic systems we know. At the same time, this assumption is made only for simplicity of arguments below and can be easily removed. Assuming the linearity of $\Psi^i_j(\mathbf{U})$, the linearized constraints (1.18) after the change of unknowns (1.22) (below we drop the bars) read:

$$\operatorname{div}\boldsymbol{\psi}_j = 0, \quad \mathbf{x} \in \mathbb{R}^n_\pm, \quad j = \overline{1, K}, \tag{2.3}$$

where $\boldsymbol{\psi}_j = \boldsymbol{\psi}_j(\mathbf{U}, \widehat{\mathbf{F}}') = ((\boldsymbol{\Psi}_j, \widehat{\mathbf{N}}), \Psi^2_j, \dots, \Psi^n_j)$, $\widehat{\mathbf{N}} = (1, -\widehat{\mathbf{F}}')$. For the case of constant coefficients (2.2), $\widehat{\mathbf{N}} = (1, -\boldsymbol{\sigma}')$.

2.1. The main definition. Given a nonnegative integer number p, we introduce the notation

$$\mathbf{W}_p := (\partial^{\alpha^1}\mathbf{U}, \dots, \partial^{\alpha^d}\mathbf{U}),$$

with

$$d = C^p_{n+p}, \quad |\alpha^i| = p, \quad i = \overline{1, d}, \quad \alpha^i \neq \alpha^j \quad \text{for } i \neq j,$$

where $\partial^\alpha := \partial_t^{\alpha_0}\partial_1^{\alpha_1}\cdots\partial_n^{\alpha_n}$, $\alpha = (\alpha_0, \dots, \alpha_n)$. In particular, $\mathbf{W}_0 = \mathbf{U}$, $\mathbf{W}_1 = (\mathbf{U}_t, \mathbf{U}_{x_1}, \dots, \mathbf{U}_{x_n})$, $\mathbf{W}_2 = (\mathbf{U}_{tt}, \mathbf{U}_{tx_1}, \dots, \mathbf{U}_{x_{n-1}x_n}, \mathbf{U}_{x_nx_n})$. Below we will usually omit the index p, i.e., $\mathbf{W} := \mathbf{W}_p$.

Differentiating systems (2.2a) (if $p \neq 0$) and taking into account relations (2.3), one gets

$$P^\pm \widetilde{L}(\widehat{\mathbf{U}}^\pm, \boldsymbol{\sigma})\mathbf{W} + \sum_{j=1}^{K}\sum_{|\alpha|=p} \mathbf{R}^\pm_{j,\alpha} \operatorname{div}\left(\boldsymbol{\psi}_j(\partial^\alpha\mathbf{U}, \boldsymbol{\sigma}')\right) = P^\pm \widetilde{\mathbf{f}}^\pm, \quad \mathbf{x} \in \mathbb{R}^n_\pm, \tag{2.4}$$

where

$$\widetilde{L}(\widehat{\mathbf{U}}^\pm, \boldsymbol{\sigma}) = I_d \otimes L(\widehat{\mathbf{U}}^\pm, \boldsymbol{\sigma}), \quad \widetilde{\mathbf{f}}^\pm = (\partial^{\alpha^1}\mathbf{f}^\pm, \dots, \partial^{\alpha^d}\mathbf{f}^\pm),$$

$\mathbf{R}^\pm_{j,\alpha} = \mathbf{R}_{j,\alpha}(\widehat{\mathbf{U}}^\pm, \boldsymbol{\sigma})$ and $P^\pm = P(\widehat{\mathbf{U}}^\pm, \boldsymbol{\sigma})$ are, respectively, vectors and nonsingular matrices of order Nd. Here and below the subscript in I_j indicates the order of the unit matrix $I_j = I$ (sometimes we omit it). Systems (2.4) can be rewritten as follows:

$$\mathcal{L}(\widehat{\mathbf{U}}^\pm, \boldsymbol{\sigma})\mathbf{W} = \boldsymbol{\mathcal{F}}^\pm, \quad \mathbf{x} \in \mathbb{R}^n_\pm, \tag{2.5}$$

where

$$\mathcal{L} = \mathcal{L}(\widehat{\mathbf{U}}^\pm, \boldsymbol{\sigma}) = \widehat{\mathcal{A}}^\pm_0\partial_t + \sum_{j=1}^{n}\widehat{\mathcal{A}}^\pm_j\partial_j, \quad \boldsymbol{\mathcal{F}}^\pm = \boldsymbol{\mathcal{F}}^\pm(\widehat{\mathbf{U}}^\pm, \boldsymbol{\sigma}) = P^\pm\widetilde{\mathbf{f}}^\pm,$$

and $\widehat{\mathcal{A}}^\pm_i = \mathcal{A}_i(\widehat{\mathbf{U}}^\pm, \boldsymbol{\sigma})$, $i = \overline{0, n}$, are matrices of order Nd (their explicit form is determined from (2.4)).

Systems (2.5) are a kind of "secondary higher order" symmetrization of the symmetric systems (2.2a) if the matrices $\widehat{\mathcal{A}}^\pm_i$ are again symmetric. Moreover, to make these matrices symmetric when writing out (2.5) one can take into account the trivial relations like

$$\partial_i\partial_j\mathbf{U} = \partial_j\partial_i\mathbf{U}, \quad \partial_i\partial_j\partial_k\mathbf{U} = \partial_j\partial_i\partial_k\mathbf{U} = \ldots = \partial_k\partial_j\partial_i\mathbf{U}, \quad \text{etc.},$$

which follow from the smoothness hypothesis (in $\mathbb{R}^n_\pm$). Clearly, systems (2.5) are not uniquely written out. It follows that a corresponding p-symmetrizer (see the definition below) is not uniquely determined and, in principle, the problem can have different p-symmetrizers. It is however quite natural and does not, of course, imply the nonuniqueness of solutions to the problem.

The boundary conditions for systems (2.5) are obtained by the tangential differentiation (with respect to t and $\mathbf{x}'$) of conditions (2.2b), and, furthermore, systems (2.4) themselves differentiated $p-1$ times and considered at $x_1 = 0$ can be used as boundary conditions. Unfortunately, the boundary conditions are difficult to present in a concrete form, but it is clear that the right-hand parts in them depend on $\partial^{\alpha}_{t,\mathrm{x}'}\mathbf{g}$, with $|\alpha| = p$, and $\partial^{\beta}\mathbf{f}^{\pm}|_{x_1=\pm 0}$, with $|\beta| = p-1$, where $\partial^{\alpha}_{t,\mathrm{x}'} := \partial_t^{\alpha_0}\partial_2^{\alpha_2}\cdots\partial_n^{\alpha_n}$, $\alpha = (\alpha_0, \alpha_2, \dots, \alpha_n)$. The corresponding vector of right-hand parts formed by $\partial^{\alpha}_{t,\mathrm{x}'}\mathbf{g}$ and $\partial^{\beta}\mathbf{f}^{\pm}|_{x_1=\pm 0}$ is below denoted by $\boldsymbol{\mathcal{G}}$.

We are now in a position to give the definition of the (strictly) dissipative p-symmetrizer.

Definition 2.1. *The set of matrices and vectors*

$$\mathbb{S} = \mathbb{S}(\widehat{\mathbf{U}}^+, \widehat{\mathbf{U}}^-, \boldsymbol{\sigma}) := \left\{P^+, P^-, \{\mathbf{R}^+_{j,\alpha}\}_{j=\overline{1,K},|\alpha|=p}, \{\mathbf{R}^-_{j,\alpha}\}_{j=\overline{1,K},|\alpha|=p}\right\}$$

($\mathbb{S} := \{P^+, P^-\}$ if system (1.1) *has no divergent constraints) is called the dissipative p-symmetrizer of problem* (2.2) *if the matrices $\widehat{\mathcal{A}}_i^{\pm}$ in* (2.5) *are symmetric and there is an open subset D of the state space $G \subset \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^n$, a constant matrix B, and a constant γ such that*

$$\widehat{\mathcal{A}}_0^+ > 0, \quad \widehat{\mathcal{A}}_0^- > 0, \tag{2.6}$$

and (cf. (1.12)*)*

$$-\left[(\widehat{\mathcal{A}}_1\mathbf{W}, \mathbf{W})\right]\big|_{x_1=0} \geq (B\boldsymbol{\mathcal{G}}, \widetilde{\mathbf{W}}) - \gamma|\boldsymbol{\mathcal{G}}|^2 \tag{2.7}$$

for all $(\widehat{\mathbf{U}}^+, \widehat{\mathbf{U}}^-, \boldsymbol{\sigma}) \in D$ and all $\mathbf{W}$ satisfying the boundary conditions for systems (2.5)*, where $\widetilde{\mathbf{W}} = (\widetilde{\mathbf{W}}^+, \widetilde{\mathbf{W}}^-)$ and $\widetilde{\mathbf{W}}^{\pm}$ is the projection of $\mathbf{W}^{\pm}$ orthogonal to* $\ker\widehat{\mathcal{A}}_1^{\pm}$ *(for shock waves $\widetilde{\mathbf{W}}^{\pm} = \mathbf{W}^{\pm} = \mathbf{W}|_{x_1=\pm 0}$).*

The set $\mathbb{S}$ is called the strictly dissipative p-symmetrizer of problem (2.2) *if it is a dissipative p-symmetrizer of this problem and there is a fixed constant $\delta > 0$ such that*

$$-\left[(\widehat{\mathcal{A}}_1\mathbf{W}, \mathbf{W})\right]\big|_{x_1=0} \geq \delta|\widetilde{\mathbf{W}}|^2 - \delta^{-1}|\boldsymbol{\mathcal{G}}|^2. \tag{2.8}$$

Remark 2.1. For 1-shocks $\widehat{A}_{\nu}^- > 0$ (cf. (1.20)) and, therefore, the strictly dissipative p-symmetrizer can be taken in the form

$$\mathbb{S} = \left\{P^+, \gamma I, \{\mathbf{R}^+_{j,\alpha}\}_{j=\overline{1,K},|\alpha|=p}, 0, \dots, 0\right\},$$

where the constant $\gamma > 0$ is large enough, P^+, $\mathbf{R}^+_{j,\alpha}$ are such that the matrices $\widehat{\mathcal{A}}_{\alpha}^+$ are symmetric, $\widehat{\mathcal{A}}_0^+ > 0$, and the relaxed condition (2.8),

$$-(\widehat{\mathcal{A}}_1^+\mathbf{W}, \mathbf{W})|_{x_1=+0} \geq \delta|\mathbf{W}^+|^2 - \delta^{-1}\left\{|\boldsymbol{\mathcal{G}}|^2 + |\mathbf{W}^-|^2\right\},$$

is fulfilled. Thanks to the choice of γ and the condition $\widehat{\mathcal{A}}_1^- = \gamma(I_d \otimes \widehat{A}_{\nu}^-) > 0$, the last inequality implies (2.8) with an appropriate (and different) δ. That is, systems (2.5) are symmetric hyperbolic and the boundary conditions for them are strictly dissipative (cf. (1.11)).

Remark 2.2. The definition of the p-symmetrizer for the case of standard boundary conditions is analogous to Definition 2.1. Note only that for problem (1.7) (with "frozen" coefficients) the p-symmetrizer has the form $\mathbb{S} = \left\{P, \{\mathbf{R}_{j,\alpha}\}_{j=\overline{1,K},|\alpha|=p}\right\}$, where P is a nonsingular matrix of order NC^p_{n+p}, etc.

We now give some concrete examples of dissipative and strictly dissipative p-symmetrizers. Without loss of generality we will consider homogenous interior equations and homogenous boundary conditions.

2.2. Example 1: The wave equation. Consider the initial-boundary-value problem in the half-plane $\mathbb{R}^2_+$ for the 2D wave equation:

$$u_{tt} = u_{x_1x_1} + u_{x_2x_2} \quad \text{for } x_1 > 0, \tag{2.9a}$$

$$u_t + au_{x_1} + bu_{x_2} = 0 \quad \text{for } x_1 = 0, \tag{2.9b}$$

where a and b are real constants. As is known, the boundary conditions (2.9b) satisfy the uniform Lopatinski condition in the half-strip

$$|b| < 1, \quad a < 0. \tag{2.10}$$

Problem (2.9) is easily reduced in the following problem for a symmetric hyperbolic system for the vector $\mathbf{U} = (u_1, u_2, u_3) = (u_t, u_{x_1}, u_{x_2})$:

$$\mathbf{U}_t + A_1\mathbf{U}_{x_1} + A_2\mathbf{U}_{x_2} = 0 \quad \text{for } x_1 > 0, \tag{2.11a}$$

$$M\mathbf{U} = 0 \quad \text{for } x_1 = 0, \tag{2.11b}$$

with

$$A_1 = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad M = \begin{pmatrix} 1 & a & b \end{pmatrix}.$$

In terms of the components of the vector $\mathbf{U}$ the trivial relation $u_{x_1x_2} = u_{x_2x_1}$ satisfied by classical solutions of (2.9) reads:

$$\operatorname{div} \mathbf{\Psi}(\mathbf{U}) = 0, \quad \mathbf{\Psi} = (u_3, -u_2). \tag{2.12}$$

Let us now forget about the connection between problems (2.9) and (2.11) (it is only important that for (2.11) the uniform Lopatinski condition is also given by (2.10)). Then (2.12) should be considered as a divergent constraint for the initial data for (2.11). Indeed, one can easily show that if (2.12) is satisfied initially, it holds for solutions of (2.11a) for all $t > 0$.

We now prove that (2.11) has a strictly dissipative 0-symmetrizer which can be taken in the form $\mathbb{S} = \{P, \mathbf{R}\}$, with

$$P = \begin{pmatrix} p_1 & p_2 & p_3 \\ p_2 & p_1 & 0 \\ p_3 & 0 & p_1 \end{pmatrix}, \quad R = \begin{pmatrix} 0 \\ -p_3 \\ p_2 \end{pmatrix}, \quad p_i \in \mathbb{R},$$

and the parameter domain D (see Definition 2.1 and Remark 2.2) coincides with the the domain of fulfillment of the uniform Lopatinski condition, (2.10). Indeed, applying $\mathbb{S}$ to (2.11), (2.12) leads to the system

$$P\mathbf{U}_t + PA_1\mathbf{U}_{x_1} + PA_2\mathbf{U}_{x_2} + \mathbf{R}\operatorname{div} \mathbf{\Psi} = \mathcal{A}_0\mathbf{U}_t + \mathcal{A}_1\mathbf{U}_{x_1} + \mathcal{A}_2\mathbf{U}_{x_2} = 0, \tag{2.13}$$

where $\mathcal{A}_0 = P > 0$ if $p_1 > 0$ and $p_1^2 - p_2^2 - p_3^2 > 0$. Also,

$$\mathcal{A}_1 = \begin{pmatrix} -p_2 & -p_1 & 0 \\ -p_1 & -p_2 & -p_3 \\ 0 & -p_3 & p_2 \end{pmatrix}, \quad \mathcal{A}_2 = \begin{pmatrix} -p_3 & 0 & -p_1 \\ 0 & p_3 & -p_2 \\ -p_1 & -p_2 & -p_3 \end{pmatrix}.$$

Note that $\mathcal{A}_0 = \mathcal{T}^*\{I_2 \otimes \mathcal{H}\}\mathcal{T}$ ($\mathcal{A}_0 > 0$ if $\mathcal{H} > 0$),

$$\mathcal{A}_1 = \mathcal{T}^*\left\{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \otimes \mathcal{H}\right\}\mathcal{T}, \quad \mathcal{A}_2 = \mathcal{T}^*\left\{\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes \mathcal{H}\right\}\mathcal{T},$$

$$\mathcal{T} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 0 & -1 \\ 0 & -1 & 0 \\ 0 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad \mathcal{H} = \begin{pmatrix} p_1 - p_3 & -p_2 \\ -p_2 & p_1 + p_3 \end{pmatrix}.$$

Omitting calculations, one has

$$-(\mathcal{A}_1\mathbf{U}, \mathbf{U})|_{x_1=0} = -\left(\{\mathcal{S}^*\mathcal{H} + \mathcal{H}\mathcal{S}\}\mathbf{V}_2, \mathbf{V}_2\right)|_{x_1=0}, \tag{2.14}$$

where

$$\mathbf{V} = \mathcal{T}\mathbf{U} = \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{pmatrix}, \quad \mathbf{V}_1|_{x_1=0} = \mathcal{S}\mathbf{V}_2|_{x_1=0}, \quad \mathcal{S} = \begin{pmatrix} \dfrac{2a}{1-b} & \dfrac{b+1}{b-1} \\ 1 & 0 \end{pmatrix}.$$

All the eigenvalues of the matrix $\mathcal{S}$ lie in the left half-plane ($\Re\lambda_i(\mathcal{S}) < 0$), provided that the uniform Lopatinski condition (2.10) holds. In this case the Lyapunov matrix equation [5]

$$\mathcal{S}^*\mathcal{H} + \mathcal{H}\mathcal{S} = -G \tag{2.15}$$

has a unique solution $\mathcal{H}$ for any symmetric matrix G, and if $G > 0$, then $\mathcal{H} = \mathcal{H}^* > 0$. Assuming that $G = G^* > 0$ and taking into account the relation $\mathbf{V} = \mathcal{T}\mathbf{U}$ and the boundary conditions $\mathbf{V}_1|_{x_1=0} = \mathcal{S}\mathbf{V}_2|_{x_1=0}$, we have

$$-(\mathcal{A}_1\mathbf{U}, \mathbf{U})|_{x_1=0} = (G\mathbf{V}_2, \mathbf{V}_2)|_{x_1=0} \geq \delta|\mathbf{U}|_{x_1=0}|^2,$$

where $\delta > 0$ is a constant depending on the norms of the matrices G, $\mathcal{S}$, and $\mathcal{T}$. Thus, $\mathbb{S}$ is the strictly dissipative 0-symmetrizer.

Remark 2.3. The constants p_1, p_2, and p_3 are found explicitly from (2.15) through the elements of the matrix $G = \{g_{ij}\}_{i,j=1,2}$. In particular,

$$p_2 = -g_{22}\frac{1-b}{2(1+b)} < 0$$

($g_{22} > 0$ since $G > 0$). The condition $p_2 \neq 0$ and the inequality $p_1^2 - p_2^2 - p_3^2 > 0$ imply that $\det \mathcal{A}_1 \neq 0$, i.e., the boundary $x_1 = 0$ is noncharacteristic for system (2.13). This could seem strange because for system (2.11a) the boundary is characteristic, but systems (2.11a) and (2.13) are equivalent (since $\det P \neq 0$). This is, however, quite natural because we should take into account the divergent constraint (2.12). Indeed, with condition (2.12) problem (2.11) has not a so-called loss of control on derivatives in the normal direction. Namely, the x_1-derivative of the "characteristic part" u_3 of $\mathbf{U}$ is estimated from (2.12). That is, problem (2.11) being formally a hyperbolic problem with characteristic boundary has the features of noncharacteristic problems.

Remark 2.4. The symmetrization of the multidimensional wave equation [8, 10] ($n \geq 3$) with strictly dissipative boundary conditions can be formalized in terms of the construction of a strictly dissipative 0-symmetrizer as well. In particular, for the case $n = 3$ the matrix P has the form

$$P = \begin{pmatrix} p_1 & p_2 & p_3 & p_4 \\ p_2 & p_1 & 0 & 0 \\ p_3 & 0 & p_1 & 0 \\ p_4 & 0 & 0 & p_1 \end{pmatrix}, \quad p_i \in \mathbb{R}, \quad p_1^2 - p_2^2 - p_3^2 - p_4^2 > 0.$$

We refer also to [13] where such a 0-symmetrizer for the 3D wave equation is used for constructing a strictly dissipative 2-symmetrizer for relativistic gas dynamical shock waves. For simplicity, in Examples 3 and 4 below we consider the 2D case for shock waves in gas dynamics and MHD, and the structure of 2-symmetrizers for them is based on a symmetrization of the 2D wave equation which is different from (2.13). For the 3D case for gas dynamical and MHD shock waves we refer to [8, 10, 14] (see also further discussion in the end of section 2).

2.3. Example 2: Compressible current-vortex sheets. For tangential discontinuities (current-vortex sheets) in MHD of ideal compressible fluid, the constant coefficients linearized problem has the form of problem (2.2) (see [45]) with

$$\widehat{A}_0^{\pm} = \operatorname{diag}\left(\frac{1}{\hat{\rho}^{\pm}(\hat{c}^{\pm})^2}, \hat{\rho}^{\pm}, \hat{\rho}^{\pm}, \hat{\rho}^{\pm}, 1, 1, 1, 1\right),$$

$$\widehat{A}_1^{\pm} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & \widehat{H}_2^{\pm} & \widehat{H}_3^{\pm} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \widehat{H}_2^{\pm} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \widehat{H}_3^{\pm} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\widehat{A}_2^{\pm} = \begin{pmatrix} \dfrac{\hat{v}_2^{\pm}}{\hat{\rho}^{\pm}(\hat{c}^{\pm})^2} & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \hat{\rho}^{\pm}\hat{v}_2^{\pm} & 0 & 0 & -\widehat{H}_2^{\pm} & 0 & 0 & 0 \\ 1 & 0 & \hat{\rho}^{\pm}\hat{v}_2^{\pm} & 0 & 0 & 0 & \widehat{H}_3^{\pm} & 0 \\ 0 & 0 & 0 & \hat{\rho}^{\pm}\hat{v}_2^{\pm} & 0 & 0 & -\widehat{H}_2^{\pm} & 0 \\ 0 & -\widehat{H}_2^{\pm} & 0 & 0 & \hat{v}_2^{\pm} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \hat{v}_2^{\pm} & 0 & 0 \\ 0 & 0 & \widehat{H}_3 & -\widehat{H}_2^{\pm} & 0 & 0 & \hat{v}_2^{\pm} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\widehat{A}_3^{\pm} = \begin{pmatrix} \dfrac{\hat{v}_3^{\pm}}{\hat{\rho}^{\pm}(\hat{c}^{\pm})^2} & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & \hat{\rho}^{\pm}\hat{v}_3^{\pm} & 0 & 0 & -\widehat{H}_3^{\pm} & 0 & 0 & 0 \\ 0 & 0 & \hat{\rho}^{\pm}\hat{v}_3^{\pm} & 0 & 0 & -\widehat{H}_3^{\pm} & 0 & 0 \\ 1 & 0 & 0 & \hat{\rho}^{\pm}\hat{v}_3^{\pm} & 0 & \widehat{H}_2^{\pm} & 0 & 0 \\ 0 & -\widehat{H}_3^{\pm} & 0 & 0 & \hat{v}_3^{\pm} & 0 & 0 & 0 \\ 0 & 0 & -\widehat{H}_3^{\pm} & \widehat{H}_2^{\pm} & 0 & \hat{v}_3^{\pm} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \hat{v}_3^{\pm} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$\mathbf{U} = (p, \mathbf{v}, \mathbf{H}, S)$, $\widehat{\mathbf{U}}^{\pm} = (\hat{p}^{\pm}, 0, \hat{v}_2^{\pm}, \hat{v}_3^{\pm}, 0, \widehat{H}_2^{\pm}, \widehat{H}_3^{\pm}, \widehat{S}^{\pm})$, and the boundary conditions

$$f_t = v_1^{\pm} - \hat{v}_2^{\pm} f_{x_2} - \hat{v}_3^{\pm} f_{x_3}, \quad [q] = 0, \tag{2.16a}$$

$$H_1^{\pm} = \widehat{H}_2^{\pm} f_{x_2} + \widehat{H}_3^{\pm} f_{x_3} \tag{2.16b}$$

at $x_1 = 0$. Here $q = p + (\widehat{\mathbf{H}}^{\pm}, \mathbf{H})$ for $\mathbf{x} \in \mathbb{R}^3_{\pm}$, $\widehat{\mathbf{H}}^{\pm} = (0, \widehat{H}_2^{\pm}, \widehat{H}_3^{\pm})$, $\hat{p}^{\pm} = p(\hat{\rho}^{\pm}, \widehat{S}^{\pm})$, $(\hat{c}^{\pm})^2 = p_{\rho}(\hat{\rho}^{\pm}, \widehat{S}^{\pm}) > 0$, and $p = p(\rho, S)$ is the state equation of gas, $\hat{\rho}^{\pm} > 0$ is the unperturbed density for $x_1 \gtrless 0$, etc. (see [45]). As in [45], without loss of generality we suppose that $\boldsymbol{\sigma} = 0$ (see (2.1)), i.e., $\widehat{A}_{\nu}^{\pm} = \widehat{A}_1^{\pm}$ and consider the homogenous problem ($\mathbf{f}^{\pm} = 0$ and $\mathbf{g} = 0$).

Remark 2.5. Since $\det \widehat{A}_1^{\pm} = 0$, the boundary $x_1 = 0$ is characteristic, i.e., current-vortex sheet is a characteristic discontinuity. For shock waves the correct number of boundary conditions is guaranteed by the Lax conditions (1.20). For the general case, the number of boundary conditions should be equal to

$$1 + \sum_{\pm} \# \text{ positive eigenvalues of } \pm A_{\nu}^{\pm} \text{ counting multiplicity.}$$

That is, for current-vortex sheets the correct number of boundary conditions is three. At first sight, problem (2.2a), (2.16) is overdetermined. On the other hand, one can show that for the original nonlinear problem the boundary conditions $(\mathbf{H}^{\pm}, \mathbf{N}) = 0$ (see [45]) can be regarded as the restrictions only on the initial data. This was shown in [46] for the case of incompressible MHD, but this proposition can be easily proved for compressible current-vortex sheets as well. Of course, this fact can be analogously (and easier) proved for the linear problem. That is, for problem (2.2a), (2.16) the boundary conditions (2.16b) are just the restrictions on the initial data.

Unlike, for example, MHD shock waves or Alfvén discontinuities [13, 44], for current-vortex sheets the Lopatinski determinant can be explicitly written out. At the same time, it is reduced to an algebraic equation of the tenth degree depending on seven dimensionless parameters and one more inner parameter determining the wave vector (see [45]). Moreover, the squaring was applied under the reduction of the Lopatinski determinant to this algebraic equation and, therefore, it can introduce spurious roots. For all these reasons both the analytical analysis and the *full* numerical study of the Lopatinski determinant are unacceptable for finding the Lopatinski condition. Although, one can analytically show that the uniform Lopatinski condition is never satisfied for problem (2.2a), (2.16), i.e., planar current-vortex sheets can be either violently unstable or weakly (neutrally) stable (see [45]). The alternative energy method suggested in [45] has first enabled one to find sufficient conditions for their weak stability.

The method in [45] can be now described in terms of the notation of dissipative symmetrizer. In fact, in [45] the dissipative 0-symmetrizer $\mathbb{S} = \{P^+, P^-, \mathbf{R}^+, \mathbf{R}^-\}$ was suggested for problem (2.2a), (2.16), where $P^{\pm} = P(\widehat{\mathbf{U}}^{\pm})$, $\mathbf{R}^{\pm} = \lambda(\widehat{\mathbf{U}}^{\pm})\mathbf{R}(\widehat{\mathbf{U}}^{\pm})$,

$$P = \begin{pmatrix} 1 & \dfrac{\lambda H_1}{\rho c^2} & \dfrac{\lambda H_2}{\rho c^2} & \dfrac{\lambda H_3}{\rho c^2} & 0 & 0 & 0 & 0 \\ \lambda H_1 \rho & 1 & 0 & 0 & -\rho\lambda & 0 & 0 & 0 \\ \lambda H_2 \rho & 0 & 1 & 0 & 0 & -\rho\lambda & 0 & 0 \\ \lambda H_3 \rho & 0 & 0 & 1 & 0 & 0 & -\rho\lambda & 0 \\ 0 & -\lambda & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -\lambda & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -\lambda & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{R} = -\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ H_1 \\ H_2 \\ H_3 \\ 0 \end{pmatrix},$$

$\lambda = \lambda(\mathbf{U})$ is a function, and the constants $\lambda^{\pm} = \lambda(\widehat{\mathbf{U}}^{\pm})$ are chosen below. Actually, the application of $\{P, \mathbf{R}\}$ to the original nonlinear MHD system gives a new symmetric form [45] of the MHD equations with the hyperbolicity condition, $\mathcal{A}_0(\mathbf{U}) = PA_0 > 0$,

$$\rho\lambda^2 < \frac{1}{1 + c_{\mathrm{A}}^2/c^2}, \tag{2.17}$$

where $c_{\mathrm{A}} = |\mathbf{H}|/\sqrt{\rho}$ (see [45]).

Note that for current sheets, i.e., for the case when $[\hat{\mathbf{v}}] = 0$ ($\hat{\mathbf{v}}^{\pm} = (0, \hat{v}_2^{\pm}, \hat{v}_3^{\pm})$) the boundary conditions (2.16) are dissipative:

$$\big[(\widehat{A}_1\mathbf{U}, \mathbf{U})\big]\big|_{x_1=0} = 2q^+[v_1] = 2q^+\big([\hat{\mathbf{v}}'], \nabla_{\mathrm{x}'} f\big) = 0,$$

where $\hat{\mathbf{v}}'^{\pm} = (\hat{v}_2^{\pm}, \hat{v}_3^{\pm})$. That is, for current sheets one has the identical 0-symmetrizer $\mathbb{S} = \{I, 0, I, 0\}$ ($\lambda^{\pm} = 0$). Suppose now that $[\hat{\mathbf{v}}] \neq 0$ and $\widehat{\mathbf{H}}^+ \times \widehat{\mathbf{H}}^- \neq 0$ (for the particular case $\widehat{\mathbf{H}}^+ \times \widehat{\mathbf{H}}^- = 0$ we refer to [45]). The matrices $\widehat{\mathcal{A}}_1^{\pm}$ (cf. (2.5)) have the form

$$\widehat{\mathcal{A}}_1^{\pm} = \begin{pmatrix} 0 & 1 & 0 & 0 & -\lambda^{\pm} & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & \widehat{H}_2^{\pm} & \widehat{H}_3^{\pm} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -\lambda^{\pm} & 0 & 0 & 0 & 0 & -\lambda^{\pm}\widehat{H}_2^{\pm} & -\lambda^{\pm}\widehat{H}_3^{\pm} & 0 \\ 0 & \widehat{H}_2^{\pm} & 0 & 0 & -\lambda^{\pm}\widehat{H}_2^{\pm} & 0 & 0 & 0 \\ 0 & \widehat{H}_3^{\pm} & 0 & 0 & -\lambda^{\pm}\widehat{H}_3^{\pm} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Then, by virtue of (2.16),

$$\big[(\widehat{\mathcal{A}}_1\mathbf{U}, \mathbf{U})\big]\big|_{x_1=0} = 2q^+[v_1 - \lambda H_1] = 2q^+\big([\hat{\mathbf{v}}' - \lambda\widehat{\mathbf{H}}'], \nabla_{\mathrm{x}'} f\big),$$

where $\widehat{\mathbf{H}}'^{\pm} = (\widehat{H}_2^{\pm}, \widehat{H}_3^{\pm})$.

The constants $\lambda^{\pm}$ are chosen so that $[\hat{\mathbf{v}}' - \lambda\widehat{\mathbf{H}}'] = 0$:

$$\lambda^{\pm} = -\frac{|[\hat{\mathbf{v}}]|\,\sin\varphi^{\mp}}{|\widehat{\mathbf{H}}^{\pm}|\,\sin(\varphi^+ - \varphi^-)}, \quad \cos\varphi^{\pm} = \frac{([\hat{\mathbf{v}}], \widehat{\mathbf{H}}^{\pm})}{|[\hat{\mathbf{v}}]|\,|\widehat{\mathbf{H}}^{\pm}|}.$$

For such $\lambda^{\pm}$ the boundary conditions for system (2.5) are dissipative and, therefore, $\mathbb{S}$ is the dissipative (but *not* strictly dissipative) 0-symmetrizer, provided that $\widehat{\mathcal{A}}_0^{\pm} = \mathcal{A}_0(\widehat{\mathbf{U}}^{\pm}) > 0$. In view of (2.17), the last conditions for the chosen $\lambda^{\pm}$ read:

$$|[\hat{\mathbf{v}}]| < |\sin(\varphi^+ - \varphi^-)| \min\left\{\frac{\gamma^+}{|\sin\varphi^-|}, \frac{\gamma^-}{|\sin\varphi^+|}\right\}, \tag{2.18}$$

where $\gamma^{\pm} = \hat{c}^{\pm}\hat{c}_{\mathrm{A}}^{\pm}/\big((\hat{c}^{\pm})^2 + (\hat{c}_{\mathrm{A}}^{\pm})^2\big)^{1/2}$. Inequality (2.18) represents the sufficient condition for the neutral stability of compressible current-vortex sheets. This condition is of importance for various astrophysical applications such as, for example, the heliopause model [4]. As was shown in [45, 46], in the incompressibility limit inequality (2.18) describes exactly half of the parameter domain of neutral stability.

2.4. Example 3: Gas dynamical shock waves. The system of gas dynamics is an unconstrained hyperbolic system, and gas dynamical shock waves are known to be 1-shocks. Therefore, while constructing a strictly dissipative symmetrizer one can suppose that $P^- = \gamma I$ (see Remark 2.1) and, clearly, $\mathbf{R}^{\pm}_{j,\alpha} = 0$. Keeping in mind the observation about P^-, without loss of generality the perturbation ahead the planar 1-shock can be assumed to be equal to zero: $\mathbf{U} = 0$ for $x_1 < 0$. That is, for gas dynamical shocks the constant coefficients linearized problem (see (2.2)) is formulated in the half-space $\mathbb{R}^n_+$ and the strictly dissipative p-symmetrizer for it is just a matrix: $\mathbb{S} = P^+$.

For the 2D case ($n = 2$) and in dimensionless values the constant coefficients linearized problem for gas dynamical shock waves has the form of (2.2) with $\mathbf{U}|_{x_1<0} = 0$ and $\widehat{A}_0^+ = \mathrm{diag}(1, M^2, M^2, 1)$,

$$\widehat{A}_\nu^+ = \widehat{A}_1^+ = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & M^2 & 0 & 0 \\ 0 & 0 & M^2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \widehat{A}_2^+ = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

(without loss of generality $\boldsymbol{\sigma} = 0$, $\mathbf{f}^{\pm} = 0$, and $\mathbf{g} = 0$). Here $M = \hat{v}_1^+/\hat{c}^+$ is the Mach number behind the shock (in view of the Lax shock conditions, $M < 1$), $\widehat{\mathbf{U}}^{\pm} = (\hat{p}^{\pm}, \hat{v}_1^{\pm}, 0, \widehat{S}^{\pm})$, $(\hat{c}^{\pm})^2 = (\rho^2 E_\rho)_\rho(\hat{\rho}^{\pm}, \widehat{S}^{\pm}) > 0$, and $E = E(\rho, S)$ is the state equation of gas, etc. (see, e.g., [13]). The vector $\mathbf{U} = (p, v_1, v_2, S)$ is the vector of perturbations in dimensionless values [13] and $\widehat{A}_\alpha^+ = D(\widehat{\mathbf{U}}^+)^* A_\alpha(\widehat{\mathbf{U}}^+) D(\widehat{\mathbf{U}}^+)$, where $D(\widehat{\mathbf{U}}^+)$ is a diagonal matrix reducing system $(2.2\mathrm{a})|_{x_1>0}$ to a dimensionless form.

The boundary conditions in a dimensionless form and after eliminating the function $f(t, x_2)$ read [13]:

$$v_1 + b_1 p = 0, \quad (v_2)_t = b_2 p_{x_2}, \quad S = b_3 p, \tag{2.19}$$

where

$$b_1 = \frac{a+1}{2M^2}, \quad b_2 = \frac{(a-1)R}{2M^2}, \quad b_3 = 1 - \frac{a}{M^2}, \quad a = \frac{h - R + 1}{h/M^2 - R + 1},$$

$R = \hat{\rho}^+/\hat{\rho}^-$, $h = (2E_S/(\rho E_{\rho S}))(\hat{\rho}^+, \widehat{S}^+)$. As is known (see [13, 34] and references therein), the boundary conditions (2.19) satisfy the uniform Lopatinski condition, i.e., planar gas dynamical shock waves are *uniformly stable* if and only if

$$\frac{M^2(R+1) - 1}{M^2(R-1) + 1} < a < 1. \tag{2.20}$$

The energy method enabling one to deduce an a priori estimate with no loss of derivatives for (2.2a), (2.19), provided that the uniform Lopatinski condition (2.20) holds, was suggested by Blokhin [6]. We now formalize this method by writing out a strictly dissipative 2-symmetrizer for this problem (for the 3D case we refer to [13] and references therein). The construction of this symmetrizer is based on a certain symmetrization of the wave equation for the pressure perturbation p implied by the acoustics system:

$$\tilde{\partial}_t^2 p - \tilde{\partial}_1^2 p - \tilde{\partial}_2^2 p = 0, \tag{2.21}$$

where

$$\tilde{\partial}_t = \frac{M}{b^2}\,\partial_t, \quad \tilde{\partial}_1 = \partial_1 - \frac{M^2}{b^2}\,\partial_t, \quad \tilde{\partial}_2 = \frac{1}{b}\,\partial_2, \quad b = \sqrt{1-M^2} \in \mathbb{R}_+.$$

Using the boundary conditions (2.19) and the acoustics system itself, one can obtain the following boundary condition for (2.21):

$$M^2(1+b_1)p_{tt} - b^2 p_{tx_1} + M^2 b_2 p_{x_2x_2} = 0, \quad x_1 = 0. \tag{2.22}$$

Problem (2.21), (2.22) is "symmetrized" as follows (see [13] for details):

$$B_0\mathbf{Y}_t + B_1\mathbf{Y}_{x_1} + B_2\mathbf{Y}_{x_2} = 0 \quad \text{for } x_1 > 0, \tag{2.23}$$

$$\mathcal{M}\mathbf{Y} = 0 \quad \text{for } x_1 = 0, \tag{2.24}$$

where

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \mathbf{Y}_3 \end{pmatrix}, \quad \mathbf{Y}_1 = \tilde{\partial}_t \begin{pmatrix} \tilde{\partial}_t p \\ \tilde{\partial}_1 p \\ \tilde{\partial}_2 p \end{pmatrix}, \quad \mathbf{Y}_i = \tilde{\partial}_{i-1} \begin{pmatrix} \tilde{\partial}_t p \\ \tilde{\partial}_1 p \\ \tilde{\partial}_2 p \end{pmatrix}, \quad i = 2, 3,$$

$$B_0 = \frac{M}{b^2}\mathcal{T}^*\left\{\begin{pmatrix} 1 & -M \\ -M & 1 \end{pmatrix} \otimes \mathcal{H}\right\}\mathcal{T}, \quad B_1 = \mathcal{T}^*\left\{\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \otimes \mathcal{H}\right\}\mathcal{T},$$

$$B_2 = \frac{1}{b}\mathcal{T}^*\left\{\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes \mathcal{H}\right\}\mathcal{T}, \quad \mathcal{T} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 0 & -1 \\ 0 & -1 & 0 \\ 0 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \otimes I_3,$$

$$\mathcal{H} = \begin{pmatrix} P_1 - P_3 & -P_2 - P_4 \\ -P_2 + P_4 & P_1 + P_3 \end{pmatrix}, \quad \mathcal{M} = \begin{pmatrix} \mathcal{M}_1 & \mathcal{M}_2 & \mathcal{M}_3 \end{pmatrix},$$

$$\mathcal{M}_1 = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathcal{M}_2 = \begin{pmatrix} -2 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & -Mb_1 & 0 \end{pmatrix}, \quad \mathcal{M}_3 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & -b_0 \end{pmatrix},$$

P_k ($k = \overline{1,3}$) are arbitrary symmetric matrices of order 3, P_4 is an arbitrary antisymmetric matrix of order 3, and $b_0 = Mb_1 + (M^3b_2/b^2)$.

Referring for detailed arguments to [13], one gets (cf. (2.14))

$$-(B_1\mathbf{Y}, \mathbf{Y})|_{x_1=0} = -\left(\{\mathcal{S}^*\mathcal{H} + \mathcal{H}\mathcal{S}\}\mathbf{V}_2, \mathbf{V}_2\right)|_{x_1=0},$$

where

$$\mathbf{V} = \mathcal{T}\mathbf{Y} = \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{pmatrix}, \quad \mathbf{V}_1|_{x_1=0} = \mathcal{S}\mathbf{V}_2|_{x_1=0}, \quad \mathcal{S} = \begin{pmatrix} \mathcal{S}_1 & -\mathcal{S}_2 \\ I_3 & 0 \end{pmatrix},$$

$$\mathcal{S}_1 = 2(\mathcal{M}_1 - \mathcal{M}_3)^{-1}\mathcal{M}_2, \quad \mathcal{S}_2 = (\mathcal{M}_1 - \mathcal{M}_3)^{-1}(\mathcal{M}_1 + \mathcal{M}_3).$$

One can show that all the eigenvalues of the matrix $\mathcal{S}$ lie strictly in the left half-plane ($\Re\lambda_i(\mathcal{S}) < 0$, $i = \overline{1,6}$) if and only if the uniform Lopatinski condition (2.20) holds. In this case the Lyapunov matrix equation in the form of (2.15) has the unique solution

$\mathcal{H} = \mathcal{H}^* > 0$ for any symmetric and positive definite matrix G. Moreover, since $\mathcal{H} > 0$, the matrix $B_0 > 0$. The assumption $G > 0$ yields

$$-(B_1\mathbf{Y}, \mathbf{Y})|_{x_1=0} = (G\mathbf{V}_2, \mathbf{V}_2)\,|_{x_1=0} \geq C_1|\mathbf{X}^+|^2,$$

where

$$\mathbf{X} = (p_{tt}, p_{tx_1}, p_{tx_2}, p_{x_1x_1}, p_{x_1x_2}, p_{x_2x_2}), \quad \mathbf{X}^+ = \mathbf{X}|_{x_1=0}, \quad \mathbf{Y} = \mathcal{K}\mathbf{X},$$

$\mathcal{K}$ is a 9×6 matrix which can be explicitly written out, and $C_1 = C_1(G) > 0$ is a constant depending on the norm of the matrix G (as well as on the norms of the matrices $\mathcal{S}$, $\mathcal{T}$, and $\mathcal{K}$).

In fact, we have constructed the strictly dissipative 1-symmetrizer for the subproblem, (2.21), (2.22), for the vector (p_t, p_{x_1}, p_{x_2}). We are now ready to describe the strictly dissipative 2-symmetrizer for the whole problem, (2.2a), (2.19). It has the form

$$\mathbb{S} = P^+ = I_{24} + \mathcal{N}^*\mathcal{K}^*B_0\mathcal{K}\mathcal{N}\big(I_6 \otimes (\widehat{A}_0^+)^{-1}\big),$$

where $\mathcal{N}$ is the projector of $\mathbf{W} = \mathbf{W}_2 = (\mathbf{U}_{tt}, \mathbf{U}_{tx_1}, \ldots, \mathbf{U}_{x_2x_2})$ on $\mathbf{X}$, i.e., $\mathbf{X} = \mathcal{N}\mathbf{W}$.

It is clear that $\widehat{\mathcal{A}}_0^+ = P^+(I_6 \otimes \widehat{A}_0^+) > 0$. Indeed, $(\widehat{\mathcal{A}}_0^+\mathbf{W}, \mathbf{W}) = ((I_6 \otimes \widehat{A}_0^+)\mathbf{W}, \mathbf{W}) + (B_0\mathbf{Y}, \mathbf{Y}) > 0$ for all $\mathbf{W} \neq 0$. Concerning condition (2.8), one has

$$-(\widehat{\mathcal{A}}_1^+\mathbf{W}, \mathbf{W})|_{x_1=0} = -((I_6 \otimes \widehat{A}_1^+)\mathbf{W}, \mathbf{W})|_{x_1=0} - (B_1\mathbf{Y}, \mathbf{Y})|_{x_1=0}.$$

Using the boundary conditions (2.19) and the acoustics system it is not difficult to show that $\mathbf{W}|_{x_1=0} = B\mathbf{X}|_{x_1=0}$, where B is a 24×6 matrix with elements depending on the coefficients of (2.19) and the matrices $\widehat{A}_\alpha^+$. Therefore, there exists a positive constant C_2 such that

$$-((I_6 \otimes \widehat{A}_1^+)\mathbf{W}, \mathbf{W})|_{x_1=0} \geq -C_2|\mathbf{X}^+|^2.$$

Analogous arguments show that $|\mathbf{X}^+|^2 \geq C_3|\mathbf{W}^+|^2$ with a constant $C_3 > 0$. On the other hand, by an appropriate choice of the matrix G (i.e., the choice of matrices P_i) one can achieve that $C_1 - C_2 > 0$. Hence,

$$-(\widehat{\mathcal{A}}_1^+\mathbf{W}, \mathbf{W})|_{x_1=0} \geq (C_1 - C_2)|\mathbf{X}^+|^2 \geq \delta|\mathbf{W}^+|^2,$$

where $\delta = (C_1 - C_2)C_3 > 0$. Thus, $\mathbb{S}$ is indeed the strictly dissipative 2-symmetrizer.

2.5. Example 4: Fast MHD shock waves. As gas dynamical shock waves, fast MHD shock waves are also 1-shocks (see, e.g., [13]). That is, the constant coefficient linearized problem for fast MHD shocks in 2D (for 3D see [14]) and in dimensionless values has the form of the problem for gas dynamical shock waves formulated above, with $\widehat{A}_0^+ = \mathrm{diag}(1, M^2, M^2, 1, 1, 1)$,

$$\widehat{A}_1^+ = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & M^2 & 0 & 0 & h_2 & 0 \\ 0 & 0 & M^2 & 0 & -h_1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & h_2 & -h_1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \widehat{A}_2^+ = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -h_2 & 0 & 0 \\ 1 & 0 & 0 & h_1 & 0 & 0 \\ 0 & -h_2 & h_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$\widehat{\mathbf{U}}^{\pm} = (\hat{p}^{\pm}, \hat{v}_1^{\pm}, \hat{v}_2^{\pm}, \widehat{H}_1^{\pm}, \widehat{H}_2^{\pm}, \widehat{S}^{\pm})$, $\hat{v}_2^{+} = 0$, $h_i = \widehat{H}_i^{+}/(\hat{c}\sqrt{\hat{\rho}})$, $\mathbf{U} = (p, v_1, v_2, H_1, H_2, S)$, etc. (see [12, 13]). The boundary conditions read

$$\begin{gathered} v_1 + b_1 p = q_1 f_{x_2}, \quad f_t = b_2 p + q_2 f_{x_2}, \quad v_2 = b_3 f_{x_2} + q_3 p, \\ H_2 = h_2 b_0 f_t - h_2 v_1 + h_1 v_2, \quad H_1 = h_2 b_0 f_{x_2}, \quad S = b_4 p + q_4 f_{x_2}, \end{gathered} \tag{2.25}$$

where the coefficients b_i are explicitly written out in [12, 13] for the case of a polytropic gas and a weak magnetic field, $q = \sqrt{h_1^2 + h_2^2} \ll 1$ (for the general case, $q \in (0, +\infty)$, see [44]), the coefficients $q_i = O(q^2)$ for $q \ll 1$, and $b_0 = 1 - (\widehat{H}_2^{-}/\widehat{H}_2^{+}) \in (0,1)$ (for parallel shocks, $\widehat{H}_2^{\pm} = 0$, $b_0 := 0$).

The energy method suggested in [12] is based on the fact that the magnetoacoustics system implies the wave equation with an additional "magnetic" term:

$$\tilde{\partial}_t^2 p - \tilde{\partial}_1^2 p - \tilde{\partial}_2^2 p + \frac{q}{b^2} \triangle Q = 0 \tag{2.26}$$

(notations are the same as in (2.21)), where $Q = (\mathbf{b}, \mathbf{H})$, $\mathbf{H} = (H_1, H_2)$, and $\mathbf{b} = (-h_1/q, h_2/q)$ ($|\mathbf{b}| = 1$). Then, the counterpart of system (2.23) reads

$$B_0 \mathbf{Y}_t + B_1 \mathbf{Y}_{x_1} + B_2 \mathbf{Y}_{x_2} + \frac{q}{b^2} \begin{pmatrix} P_1 \\ P_2 \\ P_3 \end{pmatrix} \triangle \mathbf{Q} = 0 \quad \text{for } x_1 > 0, \tag{2.27}$$

where $\mathbf{Q} = (\tilde{\partial}_t Q, \tilde{\partial}_1 Q, \tilde{\partial}_2 Q)$. Moreover, for the function p we can obtain a counterpart of the boundary condition (2.22) (see [12, 14]) which implies (2.24) with the matrices $\mathcal{M}_i$ being slightly different from the corresponding matrices in gas dynamics for the case $q \ll 1$ (the norms of the differences are of order $O(q^2)$).

The crucial role in deducing the a priori estimate [12] for (2.2a), (2.25) is played by the important fact that the term

$$\left(\mathbf{Y}, \begin{pmatrix} P_1 \\ P_2 \\ P_3 \end{pmatrix} \triangle \mathbf{Q} \right) = \sum_{i=1}^{3} (\mathbf{Y}_i, P_i \triangle \mathbf{Q})$$

can be represented in a divergent form,

$$\sum_{i=1}^{3} (\mathbf{Y}_i, P_i \triangle \mathbf{Q}) = (R_0 \mathbf{W}, \mathbf{W})_t + (R_1 \mathbf{W}, \mathbf{W})_{x_1} + (R_2 \mathbf{W}, \mathbf{W})_{x_2}, \tag{2.28}$$

where $\mathbf{W} = \mathbf{W}_2$, and the quadratic forms $(R_\alpha \mathbf{W}, \mathbf{W})$ are explicitly written out in [12, 13] (if necessary, the symmetric matrices R_α of order 36 can be written out as well). While obtaining representation (2.28) the divergent constraint $\operatorname{div} \mathbf{H} = 0$ and the magnetoacoustics system itself were essentially used.

The strictly dissipative 2-symmetrizer for fast MHD shock waves has the form

$$\mathbb{S} = \left\{ P^{+}, \{\mathbf{R}_{1,\alpha}^{+}\}_{|\alpha|=2} \right\},$$

with

$$P^{+} = I_{36} + \mathcal{N}^* \mathcal{K}^* B_0 \mathcal{K} \mathcal{N} \big(I_6 \otimes (\widehat{A}_0^{+})^{-1} \big) + P_0^{+}, \quad \{\mathbf{R}_{1,\alpha}^{+}\}_{|\alpha|=2} = \{\mathbf{R}_1, \ldots, \mathbf{R}_6\}.$$

Here $\mathcal{N}$ is the projector of $\mathbf{W}$ on $\mathbf{X}$, the matrices $\mathcal{K}$ and B_0 and the vector $\mathbf{X}$ are the same as for gas dynamical shocks, the matrix P_0^{+} and the vectors $\mathbf{R}_i$, which are of order $O(q)$ for $q \ll 1$, can be explicitly written out by analyzing representation (2.28). Moreover,

$$(\widehat{\mathcal{A}}_0^+\mathbf{W},\mathbf{W}) = ((I_6 \otimes \widehat{A}_0^+)\mathbf{W},\mathbf{W}) + (B_0\mathbf{Y},\mathbf{Y}) + \frac{2q}{b^2}(R_0\mathbf{W},\mathbf{W}),$$

$$-(\widehat{\mathcal{A}}_1^+\mathbf{W},\mathbf{W})|_{x_1=0} = -\left(\left(I_6 \otimes \widehat{A}_1^+ + \frac{2q}{b^2}R_0\right)\mathbf{W},\mathbf{W}\right)\Bigg|_{x_1=0} -(B_1\mathbf{Y},\mathbf{Y})|_{x_1=0}.$$

It is clear that for $\mathcal{H} > 0$ and $q \ll 1$ the matrix $\widehat{\mathcal{A}}_0^+ > 0$. For the case of a weak magnetic field ($q \ll 1$) and a polytropic gas, one can show that all the eigenvalues of the matrix $\mathcal{S}$ lie strictly in the left half-plane. That is, $-(\widehat{\mathcal{A}}_1^+\mathbf{W},\mathbf{W})|_{x_1=0} \geq \delta|\mathbf{W}^+|^2$, where the constant $\delta > 0$ for $q \ll 1$ (see [12, 13] for more details). Thus, $\mathbb{S}$ is indeed the strictly dissipative 2-symmetrizer of (2.2a), (2.25) for the case of a weak magnetic field and a polytropic gas. Concerning the case of a general equation of state, the same is true if we require the fulfillment of (2.20) that is in fact the uniform Lopatinski condition for fast MHD shocks for $q \ll 1$.

Remark 2.6. In principle, for the general case, $q \in (0, +\infty)$, we can try to find the conditions for $\widehat{\mathbf{U}}^\pm$ guaranteeing the fulfillment of the requirements $\widehat{\mathcal{A}}_0^+ > 0$ and $-B^*\widehat{\mathcal{A}}_1^+B > 0$, where $\mathbf{W}^+ = B\mathbf{X}^+$. Then, $\mathbb{S}$ is the strictly dissipative 2-symmetrizer, provided that these conditions hold. However, rather cumbersome calculations should be unfortunately performed to find the mentioned conditions. At the same time, if they are possible to be found at least numerically, for fixed parameters $\widehat{\mathbf{U}}^\pm$, then it is interesting to compare them with the uniform stability domain for fast MHD shock waves. This domain was found in [44] by numerical testing of the uniform Lopatinski condition with the help of an algorithm suggested for 1-shocks.

2.6. Further examples. For gas dynamical shock waves we have presented a strictly dissipative 2-symmetrizer for the 2D case. For the 3D case the structure of the 2-symmetrizer is different, but the process of construction of this symmetrizer is also based on using a symmetrization of the wave equation. The same structure has the strictly dissipative 2-symmetrizer for shock waves in relativistic gas dynamics (see [13]) and in nonrelativistic radiation hydrodynamics [3]. For shock waves in relativistic radiation hydrodynamics [15] the structure of the strictly dissipative 2-symmetrizer is a little bit more complicated; however, it refers to a symmetrization of the wave equation as well.

It is interesting to note that, for instance, the system of Landau's equations of superfluid [29] is a constrained hyperbolic system, but the relations $\nabla \times \mathbf{v}_\mathrm{s} = 0$ are not used under the construction of the strictly dissipative 2-symmetrizer for shock waves in this model (see [11] and references therein), i.e., $\mathbf{R}_{j,\alpha}^\pm = 0$. Unfortunately, there is not a general method to construct a (strictly) dissipative p-symmetrizer. For most concrete examples $p = 2$ and they are based on using different symmetrizations of the wave equation. So, Definition 2.1 was given for $p \geq 0$, but we will privately suppose that $p = 0$ or $p = 1$ or $p = 2$ (actually, we do not know examples with $p \geq 3$). Moreover, for the physical cases $n = 2$ and $n = 3$ (2D and 3D) the existence of a strictly dissipative p-symmetrizer with $p \geq 3$ implies a weaker local existence theorem for the nonlinear problem (in the generic case, for a symmetrizer with $p \geq [n/2] + 2$ one obtains a weaker nonlinear result, see section 4).

Observe that in [10] also a strictly dissipative 1-symmetrizer was in fact constructed for gas dynamical shock waves. But in this case the domain D (see Definition

2.1) is only a subdomain of the whole domain of the uniform Lopatinski condition. Although, it should be noted that for nonlinear analysis (see section 4) a 1-symmetrizer has no advantages in comparison with a 2-symmetrizer.

3. The constant coefficients linear analysis. In the rest of the paper we consider *shock waves* and only make certain remarks concerning the case of characteristic discontinuities.

ASSUMPTION 3.1. *For the boundary conditions* (2.2b), rank $\widehat{B} = n$, *i.e., the vectors* $[\boldsymbol{\mathcal{P}}^0(\widehat{\mathbf{U}})]$, $[\boldsymbol{\mathcal{P}}^k(\widehat{\mathbf{U}})]$, $k=\overline{2,n}$, *are linearly independent.*

It follows from Assumption 3.1 that $n \le N$ and there is a nonsingular matrix $\mathcal{M} = \mathcal{M}(\widehat{\mathbf{U}}^+, \widehat{\mathbf{U}}^-)$ of order N such that

$$\mathcal{M}\widehat{B} = \begin{pmatrix} I_n \\ 0 \end{pmatrix}.$$

Let

$$\mathcal{M} = \begin{pmatrix} \mathcal{M}^{\mathrm{I}} \\ \mathcal{M}^{\mathrm{II}} \end{pmatrix},$$

where $\mathcal{M}^{\mathrm{I}}$ and $\mathcal{M}^{\mathrm{II}}$ are matrices of order $n \times N$ and $(N-n)\times N$, respectively. Then the boundary conditions (2.2b) can be divided into the two groups

$$\mathbf{F} = \mathcal{M}^{\mathrm{I}}\big[\widehat{S}^{-1}\widehat{A}_\nu \mathbf{U}\big] + \mathcal{M}^{\mathrm{I}}\mathbf{g}, \quad x_1 = 0, \tag{3.1}$$

$$-\mathcal{M}^{\mathrm{II}}\big[\widehat{S}^{-1}\widehat{A}_\nu \mathbf{U}\big] = \mathcal{M}^{\mathrm{II}}\mathbf{g}, \quad x_1 = 0. \tag{3.2}$$

Note that by cross differentiation one can, in principle, eliminate the front f from relations (3.1). Such a procedure results in first order boundary conditions (see, e.g., (2.19)).

Assumption 3.1 is quite natural because, as was proved in [37], it is fulfilled for uniformly stable shock waves, i.e., when (2.2) satisfies the uniform Lopatinski condition. Moreover, as was shown in [17], Assumption 3.1 is also fulfilled for weakly stable shocks under some additional supposition. At the same time, for example, for gas dynamical and MHD shock waves Assumption 3.1 is always satisfied if only $R = \hat{\rho}^+/\hat{\rho}^- \neq 1$.

THEOREM 3.1. *Suppose the Lax shock conditions* (1.20) *and all the assumptions above are fulfilled. Suppose also that* (2.2) *has a strictly dissipative p-symmetrizer. Then, the a priori estimate*

$$\begin{aligned}
&\sum_{\pm}\Big\{|||\mathbf{U}(t)|||_{W_2^p(\mathbb{R}^n_\pm)} + \|\mathbf{U}^\pm\|_{W_2^p([0,T]\times\mathbb{R}^{n-1})}\Big\} + \|f\|_{W_2^{p+1}([0,T]\times\mathbb{R}^{n-1})}\\
&\qquad \le C\Bigg\{\sum_{\pm}\Big\{\|\mathbf{f}^\pm\|_{W_2^p([0,T]\times\mathbb{R}^n_\pm)} + |||\mathbf{U}_0|||_{W_2^p(\mathbb{R}^n_\pm)}\Big\}\\
&\qquad\qquad + \|\mathbf{g}\|_{W_2^p([0,T]\times\mathbb{R}^{n-1})} + \|f_0\|_{W_2^{p+1}(\mathbb{R}^{n-1})}\Bigg\}
\end{aligned} \tag{3.3}$$

holds for any $t \in (0,T)$. *Here* T *is a positive constant,* $C = C(T)$ *is a positive constant independent of the initial data and the source terms,*

$$|||(\cdot)(t)|||^2_{W_2^k} := \sum_{j=0}^{k} \|\partial_t^j(\cdot)(t)\|^2_{W_2^{k-j}}.$$

If (2.2) *has a dissipative (but not strictly dissipative) p-symmetrizer, the following weaker a priori estimate holds*

$$\begin{aligned}\sum_{\pm}\Big\{|||\mathbf{U}(t)|||_{W_2^r(\mathbb{R}^n_\pm)}+\|\mathbf{U}^\pm\|_{W_2^{r-1}([0,T]\times\mathbb{R}^{n-1})}\Big\}+\|f\|_{W_2^r([0,T]\times\mathbb{R}^{n-1})}\\ \leq C\Bigg\{\sum_{\pm}\Big\{\|\mathbf{f}^\pm\|_{W_2^r([0,T]\times\mathbb{R}^n_\pm)}+|||\mathbf{U}_0|||_{W_2^r(\mathbb{R}^n_\pm)}\Big\}\\ +\|\mathbf{g}\|_{W_2^{r+1}([0,T]\times\mathbb{R}^{n-1})}+\|f_0\|_{W_2^r(\mathbb{R}^{n-1})}\Bigg\},\end{aligned}\tag{3.4}$$

where $r=1$ *for* $p=0$ *and* $r=p$ *for* $p\geq 1$.

Proof. We will not prove estimate (3.3) in detail since arguments to do this are quite standard. By virtue of (2.6), (2.8), it follows from (2.5) that

$$\begin{aligned}I_1(t)+\int_0^t\int_{\mathbb{R}^{n-1}}\left(|\mathbf{W}^+|^2+|\mathbf{W}^-|^2\right)d\mathbf{x}'dt\\ \leq C_1\left\{I_1(0)+J(T)+\int_0^t I_1(s)ds\right\},\end{aligned}\tag{3.5}$$

where

$$I_1(t)=\sum_{\pm}\|\mathbf{W}(t)\|^2_{L_2(\mathbb{R}^n_\pm)},\quad J(T)=\|\mathbf{g}\|^2_{W_2^p([0,T]\times\mathbb{R}^{n-1})}+\sum_{\pm}\|\mathbf{f}^\pm\|^2_{W_2^p([0,T]\times\mathbb{R}^n_\pm)}.$$

Here and below $C_i=C_i(\widehat{\mathbf{U}}^+,\widehat{\mathbf{U}}^-,\boldsymbol{\sigma})$, $i=1,2,3,\dots$, appearing under the analysis of (2.2) are positive constants.

If $p\neq 0$, we use the elementary inequality

$$I_0(t)\leq I_0(0)+\int_0^t I(s)ds\tag{3.6}$$

coming from the trivial identity

$$\frac{d}{dt}I_0(t)=2\sum_{\pm}\int_{\mathbb{R}^n_\pm}(\mathbf{Y},\mathbf{Y}_t)\,dt,$$

where $I(t)=I_0(t)+I_1(t)$, $\mathbf{Y}=(\mathbf{W}_0,\dots,\mathbf{W}_{p-1})$, $p\geq 1$,

$$I_0(t)=\sum_{\pm}|||\mathbf{U}(t)|||^2_{W_2^{p-1}(\mathbb{R}^n_\pm)}.$$

Inequalities (3.5) and (3.6) yield

$$I(t)\leq C_1\left\{I(0)+J(T)+\int_0^t I(s)ds\right\}.$$

Applying Gronwall's lemma, one gets

$$I(t)\leq C_2(I(0)+J(T)),\quad 0\leq t\leq T.\tag{3.7}$$

From trace's property one has

$$\int_0^t\int_{\mathbb{R}^{n-1}}\left(|\mathbf{Y}^+|^2+|\mathbf{Y}^-|^2\right)d\mathbf{x}'dt\leq\int_0^t I(s)ds.\tag{3.8}$$

Adding up (3.5) and (3.8) and taking into account estimate (3.7), we obtain

$$\sum_{\pm} \|\mathbf{U}^{\pm}\|^2_{W_2^p([0,T]\times\mathbb{R}^{n-1})} \leq C_3(I(0)+J(T)). \tag{3.9}$$

Using (3.9), the boundary conditions (3.1), and an elementary inequality for f like that in (3.6) for $\mathbf{U}$, we estimate the front perturbation f:

$$\|f\|^2_{W_2^{p+1}([0,T]\times\mathbb{R}^{n-1})} \leq C_4\left\{I(0)+J(T)+\|f_0\|^2_{W_2^{p+1}(\mathbb{R}^{n-1})}\right\}. \tag{3.10}$$

Estimates (3.7), (3.9), and (3.10) imply the desired a priori estimate (3.3).

Let us now assume (2.2) has a dissipative (but not strictly dissipative) p-symmetrizer. If $p=0$ we differentiate system (2.5) with respect to t and $\mathbf{x}'$ and obtain a symmetric hyperbolic system for the vector $(\mathbf{U}_t, \mathbf{U}_{x_2}, \dots, \mathbf{U}_{x_n})$. Using this system for the case $p=0$ or system (2.5) for $p\geq 1$ and taking into account (2.6) and (2.7), we obtain the inequality

$$I_1(t) + \int_0^t \int_{\mathbb{R}^{n-1}} (\widehat{B}\boldsymbol{\mathcal{G}}, \widetilde{\mathbf{Z}})d\mathbf{x}'dt \leq C_4\left\{I_1(0)+J(T)+\int_0^t I_1(s)ds\right\}, \tag{3.11}$$

where $\widehat{B}$ is a constant matrix, $\widetilde{\mathbf{Z}} = (\mathbf{Z}^+, \mathbf{Z}^-)$, $\mathbf{Z} = (\partial_t \mathbf{W}_{r-1}, \partial_2\mathbf{W}_{r-1}, \dots, \partial_n \mathbf{W}_{r-1})$ ($r=1$ for $p=0$ and $r=p$ for $p\geq 1$). Other notations are the same as in (3.5), but for the case of 0-symmetrizer $W := W_1$, $p=1$ in J, and the vector $\boldsymbol{\mathcal{G}}$ is formed by $\partial^{\alpha}_{t,\mathrm{x}'}\mathbf{g}$ ($|\alpha|=1$) and $\mathbf{f}^{\pm}|_{x_1=\pm 0}$. While obtaining (3.11) we used the relations

$$\mathbf{U}_{x_1} = -(\widehat{A}^{\pm}_{\nu})^{-1}\widehat{A}^{\pm}_0\mathbf{U}_t - \sum_{k=2}^n (\widehat{A}^{\pm}_{\nu})^{-1}\widehat{A}^{\pm}_k \mathbf{U}_{x_k} + (\widehat{A}^{\pm}_{\nu})^{-1}\mathbf{f}^{\pm}, \quad \mathbf{x}\in\mathbb{R}^n_{\pm}$$

(recall that we consider shock waves and, therefore, $\det \widehat{A}^{\pm}_{\nu} \neq 0$).

To estimate the boundary integral in the left-hand side of inequality (3.11) we carry out standard manipulations with derivatives. For example, with the terms like $u^+_{x_2}g$ and $u^+_{x_2}h^+$ appearing in this integral we proceed as follows:

$$\int_{\mathbb{R}^{n-1}} u^+_{x_2} g\, d\mathbf{x}' = -\int_{\mathbb{R}^{n-1}} u^+ g_{x_2}\, d\mathbf{x}',$$
$$\int_{\mathbb{R}^{n-1}} u^+_{x_2} h^+\, d\mathbf{x}' = -\int_{\mathbb{R}^n_+} (u_{x_2}h)_{x_1}\, d\mathbf{x} = \int_{\mathbb{R}^n_+} (u_{x_1}h_{x_2} - u_{x_2}h_{x_1})\, d\mathbf{x},$$

where $u=\partial^{\alpha}u_j$, $h=\partial^{\beta}f_k$, $g=\partial^{\gamma}_{t,\mathrm{x}'}g_k$, $|\alpha|=|\beta|=r-1$, $|\gamma|=r$, and g_k and f_k are, respectively, components of the vectors $\mathbf{g}$ and $\mathbf{f}^+$. Analogous standard arguments were also applied in [45] to treat lower order terms in the boundary integral for the variable coefficients linearized problem for current-vortex sheets. Observe that while estimating integrals like $\int_0^t\int_{\mathbb{R}^{n-1}} u^+_t h^+\, d\mathbf{x}'dt$ we should be more careful because terms in the form $\partial_t\{\cdots\}$ do not disappear under the integration over the domain $[0,t]\times(\mathbb{R}^n_+ \cup \mathbb{R}^n_-)$ (for corresponding simple arguments see [45]). As a result, omitting details, from (3.11) we deduce estimate (3.4). □

Remark 3.1. Using

$$\mathbf{U}_t = -(\widehat{A}^{\pm}_0)^{-1}\widehat{A}^{\pm}_{\nu}\mathbf{U}_{x_1} - \sum_{k=2}^n (\widehat{A}^{\pm}_0)^{-1}\widehat{A}^{\pm}_k \mathbf{U}_{x_k} + (\widehat{A}^{\pm}_0)^{-1}\mathbf{f}^{\pm}, \quad \mathbf{x}\in\mathbb{R}^n_{\pm},$$

one can reduce estimates (3.3) and (3.4) to those with the norms $\|(\cdot)(t)\|$ instead of the norms $|||(\cdot)(t)|||$.

COROLLARY 3.2. *If the symmetric hyperbolic system* (1.5) *meets either the block structure condition* [1, 35] *or the conditions of Métivier and Zumbrun* [38] *and problem* (2.2) *for the case of Lax shocks has a strictly dissipative p-symmetrizer, then in the parameter domain D the boundary conditions* (2.2b) *satisfy the uniform Lopatinski condition.*

Proof. First of all, following arguments like those used in [34] for strictly dissipative boundary value problems, one can easily obtain an analogue of estimate (3.3) when the exponentially weighted $W^s_{2,\eta}$-norms (with $s = p$ and $s = p + 1$) are used instead of the usual Sobolev norms, where

$$\|\cdot\|_{W^s_{2,\eta}} := \sum_{|\alpha|\leq s} \eta^{s-|\alpha|}\|e^{-\eta t}\partial^\alpha(\cdot)\|_{L_2}.$$

For the case when system (1.5) satisfies the block structure condition [1, 35, 32], it was proved in [32] (see also [37]) that such an estimate (with $s = 0$ and $s = 1$) holds for problem (2.2) *if and only if* this problem meets the uniform Lopatinski condition. This result was recently extended by Métivier and Zumbrun to the case of variable multiplicities provided that some additional conditions [38] hold. It is clear that the $L_{2,\eta}$-estimate (with the $W^1_{2,\eta}$-norm for f) implies $W^s_{2,\eta}$-estimates (see [37]). Hence, the boundary conditions (2.2b) satisfy the uniform Lopatinski condition. ∎

Remark 3.2. If the linear problem (2.2) meets the uniform Lopatinski condition and the symmetric hyperbolic system (1.5) satisfies either the block structure condition [1, 35] or the conditions of Métivier and Zumbrun [38], then the solution to (2.2) obeys an a priori L_2-estimate [32, 37]. That is, if $p > 0$, the result of Theorem 3.1 obtained for the linearized problem by the energy method is weaker than that in [32, 37] obtained by Kreiss' symmetrizer analysis in the sense that in estimate (3.3) we require more regularity for $\mathbf{U}$. However, if $p < [n/2] + 2$, for the original nonlinear problem the energy method gives the same result (see Theorem 4.1) as the technique used in [32, 37]. Since $p < [n/2] + 2$ for all the known concrete examples of p-symmetrizers, we will suppose that this condition is satisfied.

Remark 3.3. In [45] the a priori estimates for the linearized problem for current-vortex sheets were written out for the case of the homogenous problem ($\mathbf{f}^{\pm} = 0$ and $\mathbf{g} = 0$). For the case of the nonhomogenous problem (2.2a), (2.16) (with the source term $\mathbf{g}$ in (2.16)), the a priori estimate

$$\begin{aligned}\sum_{\pm}\Big\{|||\mathbf{U}(t)|||_{\widetilde{W}^1_2(\mathbb{R}^3_{\pm})} + \|\mathbf{V}^{\pm}\|_{L_2([0,T]\times\mathbb{R}^2)}\Big\} + \|f\|_{W^1_2([0,T]\times\mathbb{R}^2)} \\ \leq C\Bigg\{\sum_{\pm}\Big\{\|\mathbf{f}^{\pm}\|_{W^1_2([0,T]\times\mathbb{R}^3_{\pm})} + |||\mathbf{U}_0|||_{\widetilde{W}^1_2(\mathbb{R}^3_{\pm})}\Big\} \\ +\|\mathbf{g}\|_{W^2_2([0,T]\times\mathbb{R}^2)} + \|f_0\|_{W^1_2(\mathbb{R}^{n-1})}\Bigg\}\end{aligned}$$

can be deduced, provided that the sufficient neutral stability condition (2.18) holds. Here $\mathbf{V}$ ($= (q, v_1, H_1)$, see [45]) is the "noncharacteristic part" of $\mathbf{U}$,

$$|||\mathbf{U}(t)|||^2_{\widetilde{W}^s_2(\mathbb{R}^n_{\pm})} = |||\mathbf{V}(t)|||^2_{W^s_2(\mathbb{R}^n_{\pm})} + \sum_{|\alpha|\leq s}\|(\partial^\alpha_{t,\mathrm{x}'}\mathbf{U})(t)\|^2_{L_2(\mathbb{R}^n_{\pm})}.$$

For the general case of characteristic discontinuities, if we require that in (3.1)

$$\mathcal{M}^{\mathrm{I}}\big[\widehat{S}^{-1}\widehat{A}_\nu \mathbf{U}\big] = \mathcal{M}_1\mathbf{V}^+ + \mathcal{M}_2\mathbf{V}^- \tag{3.12}$$

with corresponding matrices $\mathcal{M}_i$, i.e., the vector-function $\mathbf{F}$ can be expressed by the "noncharacteristic parts" $\mathbf{V}^\pm$ of the traces $\mathbf{U}^\pm$, then in the counterparts of estimates (3.3) and (3.4) the $\widetilde{W}_2^s(\mathbb{R}^n_\pm)$-norms are used instead of the $W_2^s(\mathbb{R}^n_\pm)$-norms, and we can control only the "noncharacteristic" traces $\mathbf{V}^\pm$. If assumption (3.12) does not hold, we have weaker a priori estimates. For example, this is so for current-vortex sheets if $\widehat{\mathbf{H}}^+ \times \widehat{\mathbf{H}}^- = 0$. For this case, the a priori estimate indicates already the loss of two derivatives from the front f (see [45]).

4. Local existence of shock-front solutions. The local existence theorem for the nonlinear problem (1.17) has been first proved by Blokhin [7, 9] for uniformly stable gas dynamical shock waves by the direct energy method. Recall that the linearized constant coefficients problem for them has a strictly dissipative 2-symmetrizer (see section 2). The functional setting in the theorem from [7, 9] (see also [13]) is provided by the usual Sobolev spaces W_2^s, where $s \geq 3$. The analogous theorem, but in the exponentially weighted Sobolev spaces $W^s_{2,\eta}$, where s is large enough, was proved by Majda [33] for Lax shocks by Kreiss' symmetrizer technique [27] and using pseudo-differential calculus, provided that the symmetric hyperbolic system satisfies the block structure condition [1, 35, 32].

Recently, the theorem from [33] (see also [34]) was considerably improved by Métivier in [37], where the nonlinear local existence theorem was formulated in the form of Blokhin's theorem from [7, 9, 13] (see below). Actually, the theorem proved in [37] is valid for shock waves for which the linearized problem admits constructing Kreiss' symmetrizer. That is, the class of hyperbolic symmetrizable systems covered by this theorem is wider than that of systems satisfying the block structure condition. Moreover, taking into account the recent result in [38] mentioned above, the local existence theorem from [37] (see also [7, 9, 13]) takes place for the hyperbolic symmetrizable systems satisfying either the block structure condition or Métivier and Zumbrun's conditions [38]. That is, for Lax shock waves for which the assumption of Corollary 3.2 is fulfilled we have the following theorem (cf. [7, 9, 13, 37]).

THEOREM 4.1. *Let that the linearized constant coefficients problem* (2.2) *has a strictly dissipative p-symmetrizer. Suppose the initial data* (1.17c) *satisfy the hyperbolicity condition* $A_0 > 0$ *(for* $\mathbf{x} \in \mathbb{R}^n_\pm$*), the Lax shock conditions* (1.20)*, and the compatibility conditions (see* [37]*). Suppose also that* $(\mathbf{U}_0|_{x_1>0}, \mathbf{U}_0|_{x_1<0}, f_0) \in D$ *for all* $\mathbf{x} \in \mathbb{R}^n_\pm$ *(see Definition* 2.1*). Then, for all*

$$(\mathbf{U}_0, f_0) \in \big\{W_2^s(\mathbb{R}^n_+) \cap W_2^s(\mathbb{R}^n_-)\big\} \times W_2^{s+1}(\mathbb{R}^{n-1}),$$

where $s \geq [n/2] + 2$*, there is a sufficiently short time* $T > 0$ *such that* (1.17) *has a unique solution*

$$(\mathbf{U}, f) \in Z_T^s = \big\{X_s([0,T], \mathbb{R}^n_+) \cap X_s([0,T], \mathbb{R}^n_-)\big\} \times W_2^{s+1}([0,T] \times \mathbb{R}^{n-1}),$$

where

$$X_k([0,T], \mathbb{R}^n_\pm) := \bigcap_{j=0}^{k} C^j([0,T], W_2^{k-j}(\mathbb{R}^n_\pm))$$

with the norm $\|\cdot\|_{X_k} = \max\limits_{t\in[0,T]} |||(\cdot)(t)|||_{W_2^k}$.

It seems that for most physical examples of hyperbolic symmetrizable systems either the block structure condition or the "nonglancing" condition of Métivier and Zumbrun [38] is satisfied. Moreover, we still do not know any concrete example of a strictly dissipative p-symmetrizer for a hyperbolic system for which both of these conditions are violated. Therefore, in view of Corollary 3.2, construction of a strictly dissipative p-symmetrizer can be considered as an indirect test of the uniform Lopatinski condition. That is, as soon as such a symmetrizer is found, we have Theorem 4.1. In the light of this, there is now no practical sense for proving Theorem 4.1 directly by the energy method (as was earlier done in [7, 9, 10] for gas dynamical shocks), i.e., without referring to [37, 38] and Corollary 3.2.

We connect further perspectives of the method of p-symmetrizers with "nonstandard" problems, in particular, for characteristic discontinuities for which the structure of the Lopatinski determinant cannot be analyzed for technical reasons (see discussion in section 5). At the same time, to demonstrate how the energy method works for Lax shock waves for the case of variable coefficients and for the original nonlinear problem, we now outline the proof of Theorem 4.1. The main attention will be given to the deduction of an a priori estimate for the variable coefficients linearized problem (1.24), (1.17c). After that, to show the existence of solutions to (1.24), (1.17c) we comment how to go back from the system for p-derivatives with strictly dissipative boundary conditions to the original problem (1.24), (1.17c). At last, the proof of the existence of solutions to the nonlinear problem (1.17) follows from a fixed-point argument and we sketch it in the end of this section.

In the following we suppose that $p < [n/2]+2$ (see Remark 3.2). We just observe that if we prove Theorem 4.1 by the energy method, then for the case $p \geq [n/2]+2$ we have to assume that $s \geq \max\{[n/2]+2, p+1\}$. We underline once more that we do not know concrete examples of p-symmetrizers with $p \geq [n/2]+2$.

We first analyze the variable coefficients linear problem (1.24). We introduce the norm of $(\mathbf{u}(t,\mathbf{x}), \varphi(t,\mathbf{x}')) \in \mathbb{R}^N \times \mathbb{R}$:

$$\mathcal{N}_T^k(\mathbf{u},\varphi) := \sum_{\pm} \left\{ \|\mathbf{u}\|_{X_k([0,T],\mathbb{R}^n_\pm)} + \|\mathbf{u}^\pm\|_{W_2^k([0,T]\times\mathbb{R}^{n-1})} \right\} + \|\varphi\|_{W_2^{k+1}([0,T]\times\mathbb{R}^{n-1})},$$

where k is a nonnegative integer number. Fix an integer $s \geq [n/2]+2$ and consider $(\widehat{\mathbf{U}}, \hat{f}) \in Z_T^s$ with a time $T > 0$. Assume that there is a constant $M > 0$ such that

$$\mathcal{N}_T^s(\widehat{\mathbf{U}}, \hat{f}) \leq M. \tag{4.1}$$

THEOREM 4.2. *Given an integer $m \geq p$, suppose that problem* (1.24) *with "frozen" coefficients*

$$(\widehat{\mathbf{U}}|_{x_1>0}, \widehat{\mathbf{U}}|_{x_1<0}, \widehat{\mathbf{F}}) = (\widehat{\mathbf{U}}^+, \widehat{\mathbf{U}}^-, \boldsymbol{\sigma})$$

has a strictly dissipative p-symmetrizer ($p < [n/2]+2$) and $(\widehat{\mathbf{U}}, \hat{f}) \in Z_T^s$, with $s = \max\{m, [n/2]+2\}$. Suppose also that the Lax shock conditions (1.20) *and inequality* (4.1) *are fulfilled. Then, the following a priori estimate holds for the initial-boundary-value problem* (1.24), (1.17c)*:*

$$\begin{aligned} \mathcal{N}_T^m(\mathbf{U}, f) \leq C(T,M) \Bigg\{ & \sum_{\pm} \left\{ \|\mathbf{f}^\pm\|_{W_2^m([0,T]\times\mathbb{R}^n_\pm)} + |||\mathbf{U}_0|||_{W_2^m(\mathbb{R}^n_\pm)} \right\} \\ & + \|\mathbf{g}\|_{W_2^m([0,T]\times\mathbb{R}^{n-1})} + \|f_0\|_{W_2^{m+1}(\mathbb{R}^{n-1})} \Bigg\}. \end{aligned} \tag{4.2}$$

Here and below $C = C(T, M)$, $C_i = C_i(T, M)$, $i = 1, 2, 3, \dots$, are positive constants independent of the data and depending on T and M.

Proof. The methods for deducing the a priori estimate (4.2) are standard and based on the application of the Gagliardo–Nirenberg inequalities (see, e.g., [34])

$$\|\partial^\alpha u\|_{L_{2p}(\Omega)} \le c_k \|u\|^{1-1/p}_{L_\infty(\Omega)} \|u\|^{1/p}_{W_2^k(\Omega)}, \quad \frac{1}{p} = \frac{|\alpha|}{k}, \tag{4.3}$$

$$\|\partial^\alpha u\|_{L_p(\Omega)} \le c_k \|u\|^{1-r}_{L_2(\Omega)} \|u\|^r_{W_2^k(\Omega)}, \quad \frac{|\alpha|}{k} < r < 1, \quad \frac{1}{p} = \frac{1}{2} + \frac{|\alpha| - rk}{\dim\Omega}, \tag{4.4}$$

where $c_k > 0$ is a constant, $2 < p < \infty$. The domain Ω can be, for example, $\mathbb{R}^n$, $\mathbb{R}^n_\pm$, $[0, T] \times \mathbb{R}^n_\pm$, or $[0, T] \times \mathbb{R}^{n-1}$ (in general, Ω is a Lipschitz domain).

Inequalities (4.3) and (4.4) imply a number of calculus inequalities (see, e.g., [47, 34]). In particular, using (4.3) and (4.4), one can obtain the inequality

$$\|uv\|_{W_2^k(\Omega)} \le c_k \|u\|_{W_2^q(\Omega)} \|v\|_{W_2^k(\Omega)}, \quad q = \max\left\{\left[\frac{n}{2}\right] + 1, k\right\} \tag{4.5}$$

(here $\dim\Omega = n$). In Appendix B of [41] the following generalization of the last inequality was proved:

$$|||(uv)(t)|||_{W_2^k(\Omega)} \le c_k |||u(t)|||_{W_2^q(\Omega)} |||v(t)|||_{W_2^k(\Omega)}, \tag{4.6}$$

where Ω is a space domain (e.g., $\Omega = \mathbb{R}^n_\pm$). It is clear that the analogous inequality holds when x_1 is fixed instead of t:

$$\langle\!\langle\!\langle uv(x_1) \rangle\!\rangle\!\rangle_k \le c_k \langle\!\langle\!\langle u(x_1) \rangle\!\rangle\!\rangle_q \, \langle\!\langle\!\langle v(x_1) \rangle\!\rangle\!\rangle_k, \tag{4.7}$$

where

$$\langle\!\langle\!\langle (\cdot)(x_1) \rangle\!\rangle\!\rangle_k := \sum_{j=0}^{k} \|\partial_1^j (\cdot)(x_1)\|^2_{W_2^{k-j}([0,T]\times\mathbb{R}^{n-1})}.$$

One can also get the more special inequality

$$\begin{aligned} &\|(\partial^{\alpha_1} u_1 \cdots \partial^{\alpha_l} u_l)(\partial^\beta_{t,\mathrm{x}'} v)(x_1)\|_{L_2([0,T]\times\mathbb{R}^{n-1})} \\ &\qquad \le c_k \|v(x_1)\|_{W_2^k([0,T]\times\mathbb{R}^{n-1})} \prod_{i=1}^{l} \langle\!\langle\!\langle u_i(x_1) \rangle\!\rangle\!\rangle_q, \end{aligned} \tag{4.8}$$

where $|\alpha_1| + \dots + |\alpha_l| + |\beta| = k$. To prove (4.8) we should follow arguments analogous to those from Appendix B of [41], and the proof is based mainly on the application of (4.4).

Let us obtain the system satisfied by the vector $\mathbf{W}_\beta = \partial^\beta_{t,\mathrm{x}'} \mathbf{W}$, with $|\beta| \le m - p$. It follows from (1.24a) that

$$L(\widehat{\mathbf{U}}, \widehat{\mathbf{F}})(\partial^{\alpha^i} \partial^\beta_{t,\mathrm{x}'} \mathbf{U}) = \mathbf{f}^\pm_{i\beta} \quad \text{if} \quad \mathbf{x} \in \mathbb{R}^n_\pm, \tag{4.9}$$

where

$$\mathbf{f}^\pm_{i\beta} = \partial^{\alpha^i} \partial^\beta_{t,\mathrm{x}'} \mathbf{f}^\pm - [\partial^{\alpha^i} \partial^\beta_{t,\mathrm{x}'}, L(\widehat{\mathbf{U}}, \widehat{\mathbf{F}})]\mathbf{U}.$$

Here and below we use the notation of commutator: $[a,b]c := a(bc) - b(ac)$. From systems (4.9) with $i = \overline{1,d}$ we construct the system for $\mathbf{W}_\beta$ (cf. (2.4), (2.5)):

$$\mathcal{L}(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\mathbf{W}_\beta = \boldsymbol{\mathcal{F}}_\beta^{\pm}\,, \quad \mathbf{x}\in\mathbb{R}^n_{\pm}\,, \tag{4.10}$$

where

$$\mathcal{L} = \mathcal{L}(\widehat{\mathbf{U}},\widehat{\mathbf{F}}) = \mathcal{A}_0(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\partial_t + \sum_{j=1}^{n}\mathcal{A}_j(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\partial_j\,, \quad \widetilde{\mathbf{f}}_\beta^{\pm} = (\mathbf{f}_{1\beta}^{\pm},\dots,\mathbf{f}_{d\beta}^{\pm}),$$

$$\boldsymbol{\mathcal{F}}_\beta^{\pm} = P(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\widetilde{\mathbf{f}}_\beta^{\pm} - \sum_{j=1}^{K}\sum_{|\alpha|=p}\mathbf{R}_{j,\alpha}(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\big[\partial^\alpha\partial^\beta_{t,\mathrm{x}'},\widehat{\mathbf{N}}\big]\partial_1\boldsymbol{\Psi}_j(\mathbf{U}),$$

the matrices $P(\widehat{\mathbf{U}},\widehat{\mathbf{F}})$ and the vectors $\mathbf{R}_{j,\alpha}(\widehat{\mathbf{U}},\widehat{\mathbf{F}})$ form the strictly dissipative p-symmetrizer $\mathbb{S}$ if we "freeze" their coefficients, and the matrices $\mathcal{A}_i$ $(i=\overline{0,n})$ with "frozen" coefficients are the same as in (2.5).

For system (4.10) with variable coefficients, the counterpart of condition (2.8) is

$$-\big[(\mathcal{A}_1(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\mathbf{W}_\beta,\mathbf{W}_\beta)\big]\big|_{x_1=0} \ge \delta\big(|\mathbf{W}_\beta^+|^2 + |\mathbf{W}_\beta^-|^2\big) - \delta^{-1}g^2, \tag{4.11}$$

where g^2 is a sum of terms in the form

$$|G_k(\widehat{\mathbf{U}}^+,\widehat{\mathbf{U}}^-,\widehat{\mathbf{F}})\partial^\alpha_{t,\mathrm{x}'}\mathbf{g}|^2, \qquad |G_l(\widehat{\mathbf{U}}^+,\widehat{\mathbf{U}}^-,\widehat{\mathbf{F}})\partial^\gamma\mathbf{f}^{\pm}|_{x_1=\pm0}|^2,$$

$$|G_{l_0}(\widehat{\mathbf{U}}^+,\widehat{\mathbf{U}}^-,\widehat{\mathbf{F}})\partial^\gamma\mathbf{U}^{\pm}|^2, \qquad \text{and} \quad |G_{l_0}(\widehat{\mathbf{U}}^+,\widehat{\mathbf{U}}^-,\widehat{\mathbf{F}})\partial^\gamma_{t,\mathrm{x}'}\mathbf{F}|^2,$$

with

$$G_i(\widehat{\mathbf{U}}^+,\widehat{\mathbf{U}}^-,\widehat{\mathbf{F}}) = \partial^{\alpha_1}\hat{u}^+_{i_1}\cdots\partial^{\alpha_j}\hat{u}^+_{i_j}\partial^{\alpha_{j+1}}\hat{u}^-_{i_{j+1}}\cdots\partial^{\alpha_r}\hat{u}^-_{i_r}$$
$$\times\partial^{\alpha_{r+1}}_{t,\mathrm{x}'}\widehat{F}_{i_{r+1}}\cdots\partial^{\alpha_q}_{t,\mathrm{x}'}\widehat{F}_{i_q}H(\widehat{\mathbf{U}}^+,\widehat{\mathbf{U}}^-,\widehat{\mathbf{F}}),$$

$$|\alpha_1| + \ldots + |\alpha_q| = i\,, \quad 0\le q\le 2N+n,$$

$$k+|\alpha|\le m, \quad l+|\gamma|\le m-1, \quad l_0+|\gamma|\le m, \quad |\gamma|\le m-1$$

(for constant coefficients, cf. (2.8), $k=l=0$ and there are no lower order terms, i.e., $G_{l_0}\equiv 0$). Here $\widehat{F}_{i_j}$ is a component of the vector $\widehat{\mathbf{F}}$ ($\widehat{F}_{i_j} = \hat{f}_t$ or $\widehat{F}_{i_j} = \hat{f}_{x_k}$), $H(\widehat{\mathbf{U}}^+,\widehat{\mathbf{U}}^-,\widehat{\mathbf{F}})$ is a matrix which elements are determined by the elements of the matrices A_0, A_ν, and A_k $(k=\overline{2,n})$ and their derivatives up to order m with respect to $\mathbf{U}$ and $\mathbf{F}$.

Since arguments below are standard we are quite brief in the rest of the proof. In view of (4.11), using arguments as in (3.6), (3.8) and applying energy methods to (4.10), we deduce the inequality

$$I_{\tan}(t) + \int_0^t I^{\mathrm{tr}}_{\tan}(s)ds \le C_1(T,M)\left\{I(0) + J(t) + \int_0^t I(s)ds\right\}, \tag{4.12}$$

where

$$I_{\tan}(t)=\sum_{\pm}\left\{|||\mathbf{U}(t)|||^2_{W_2^p(\mathbb{R}^n_\pm)}+\sum_{|\beta|\le m-p}\|\mathbf{W}_\beta\|^2_{L_2(\mathbb{R}^n_\pm)}\right\},$$

$$I^{\rm tr}_{\tan}(t)=\sum_{\pm}\left\{\sum_{j=0}^{p}|||\partial_1^j\mathbf{U}^\pm(t)|||^2_{W_2^{p-j}(\mathbb{R}^{n-1})}+\sum_{|\beta|\le m-p}\|\mathbf{W}^\pm_\beta\|^2_{L_2(\mathbb{R}^{n-1})}\right\},$$

$$\partial_1^j\mathbf{U}^\pm:=\partial_1^j\mathbf{U}|_{x_1=\pm0},\quad I(t)=\sum_{\pm}|||\mathbf{U}(t)|||^2_{W_2^m(\mathbb{R}^n_\pm)},$$

$$J(t)=\|g\|^2_{L_2([0,t]\times\mathbb{R}^{n-1})}+\sum_{\pm}\sum_{|\beta|\le m-p}\|\boldsymbol{\mathcal{F}}^\pm_\beta\|^2_{L_2([0,t]\times\mathbb{R}^n)}.$$

The commutator $[\partial^{\alpha^i}\partial^\beta_{t,\mathrm{x}'},L(\widehat{\mathbf{U}},\widehat{\mathbf{F}})]\mathbf{U}$ is a sum of terms $G_k(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\partial^\alpha\mathbf{U}$, where G_k are determined as G_i above (but not on the boundary), $k+|\alpha|\le m+1$, $k\ge1$, and $|\alpha|\ge1$. Since $k\ge1$ and $|\alpha|\ge1$, applying (4.6) with $k=m-1$ and using elementary inequalities like

$$\sum_{|\alpha_1|+\ldots+|\alpha_l|\le k}\|\partial^{\alpha_1}v_1\cdots\partial^{\alpha_l}v_l\|_{L_2}\le\mathrm{const}\|v_1\cdots v_l\|_{W_2^k},$$

one estimates the commutator

$$\sum_{\pm}\left\|\,[\partial^{\alpha^i}\partial^\beta_{t,\mathrm{x}'},L(\widehat{\mathbf{U}},\widehat{\mathbf{F}})]\mathbf{U}\,\right\|^2_{L_2([0,t]\times\mathbb{R}^{n-1})}\le C_2(T,M)\int_0^t I(s)ds.$$

Then, estimating analogously other terms in $\boldsymbol{\mathcal{F}}^\pm_\beta$, one gets

$$\sum_{\pm}\sum_{|\beta|\le m-p}\|\boldsymbol{\mathcal{F}}^\pm_\beta\|^2_{L_2([0,t]\times\mathbb{R}^{n-1})}\le C_3(T,M)\left\{\sum_{\pm}\|\mathbf{f}^\pm\|^2_{W_2^m([0,T]\times\mathbb{R}^n_\pm)}+\int_0^t I(s)ds\right\}.$$

To estimate the L_2-norm of g we use inequality (4.7) at $x_1=\pm0$ and trace's property. As a result, one has

$$\|g\|^2_{L_2([0,t]\times\mathbb{R}^{n-1})}\le C_4(T,M)\Big\{\|\mathbf{g}\|^2_{W_2^m([0,T]\times\mathbb{R}^{n-1})}$$
$$+\sum_{\pm}\|\mathbf{f}^\pm\|^2_{W_2^m([0,T]\times\mathbb{R}^n_\pm)}+\int_0^t\Big(I(s)+\|\mathbf{F}(s)\|^2_{W_2^{m-1}(\mathbb{R}^{n-1})}\Big)ds\Big\}.$$

Expressing $\mathbf{F}$ by $\mathbf{U}^\pm$ and $\mathbf{g}$ ((3.1) for variable coefficients is applied) and using trace's property, from (4.12) one obtains

$$I_{\tan}(t)+\int_0^t I^{\rm tr}_{\tan}(s)ds\le C_5(T,M)\left\{I(0)+J_1(T)+\int_0^t I(s)ds\right\},\tag{4.13}$$

where

$$J_1(T)=\|\mathbf{g}\|^2_{W_2^m([0,T]\times\mathbb{R}^{n-1})}+\sum_{\pm}\|\mathbf{f}^\pm\|^2_{W_2^m([0,T]\times\mathbb{R}^n_\pm)}.$$

Applying energy methods to systems for the vectors $\partial^\alpha \mathbf{U}$ with $|\alpha| \le m$ and reasoning as above one can easily obtain the inequality

$$I(t) - \int_0^t I^{\mathrm{tr}}(s)ds \le C_6(T,M)\left\{I(0) + J_1(T) + \int_0^t I(s)ds\right\}, \tag{4.14}$$

where

$$I^{\mathrm{tr}}(t) = \sum_{\pm}\sum_{j=0}^{m} |||\partial_1^j \mathbf{U}^{\pm}(t)|||^2_{W_2^{m-j}(\mathbb{R}^{n-1})}.$$

To get an inequality for $I(t)$ when "+" stands in (4.14) instead of "−", we use the great advantage that the boundary conditions are strictly dissipative and, therefore, one has the positive integral in the left-hand side of inequality (4.13). We should now estimate the "full trace" $\int_0^t I^{\mathrm{tr}}(s)ds$ by $\int_0^t I^{\mathrm{tr}}_{\tan}(s)ds$.

Using the equations

$$\mathbf{U}_{x_1}|_{x_1=\pm 0} = A_\nu^{-1}(\widehat{\mathbf{U}}, \widehat{\mathbf{F}})\left\{\mathbf{f}^{\pm} - A_0(\widehat{\mathbf{U}})\mathbf{U}_t - \sum_{k=2}^{n} A_k(\widehat{\mathbf{U}})\mathbf{U}_{x_k}\right\}\Bigg|_{x_1=\pm 0},$$

one has that $\partial^\alpha \mathbf{U}^{\pm}$ with $|\alpha| = m$ is a sum of terms $G_k(\widehat{\mathbf{U}}^{\pm}, \widehat{\mathbf{F}})\partial^\beta_{t,\mathbf{x}'}\mathbf{U}^{\pm}$ and $G_l(\widehat{\mathbf{U}}^{\pm}, \widehat{\mathbf{F}})$ $\times \partial^\gamma \mathbf{f}^{\pm}|_{x_1=\pm 0}$, where $k + |\beta| \le m$ and $l + |\gamma| \le m-1$. Applying to these terms inequalities (4.8) and (4.7) (at $x_1 = \pm 0$), respectively, one gets the desired estimate

$$\int_0^t I^{\mathrm{tr}}(s)ds \le C_7(T,M)\int_0^t I^{\mathrm{tr}}_{\tan}(s)ds. \tag{4.15}$$

Summing up (4.13) multiplied by $2C_7$ with (4.14) and using (4.15), we obtain

$$I(t) + \int_0^t I^{\mathrm{tr}}(s)ds \le C_8(T,M)\left\{I(0) + J_1(T) + \int_0^t I(s)ds\right\}, \tag{4.16}$$

where $C_8 = C_6 + 2C_5C_7$.

Throwing away the positive integral in the left-hand side of (4.16) and applying Gronwall's lemma yield

$$I(t) \le C_9(T,M)\big(I(0) + J_1(T)\big), \quad 0 \le t \le T. \tag{4.17}$$

It follows from (4.16) and (4.17) that

$$\sum_{\pm} \|\mathbf{U}^{\pm}\|^2_{W_2^m([0,T]\times\mathbb{R}^{n-1})} \le C_{10}(T,M)\big(I(0) + J_1(T)\big). \tag{4.18}$$

At last, using the boundary conditions (3.1) (for variable coefficients) and applying (4.5) with $\Omega = [0,T]\times\mathbb{R}^{n-1}$ and $k = m$, we get from (4.18) that

$$\|f\|^2_{W_2^{m+1}([0,T]\times\mathbb{R}^{n-1})} \le C_{11}(T,M)\left\{I(0) + J_1(T) + \|f_0\|^2_{W_2^{m+1}(\mathbb{R}^{n-1})}\right\}. \tag{4.19}$$

Estimates (4.17)–(4.19) imply (4.2). □

Remark 4.1. In [7, 9, 10], the estimate in the form of (4.2) was obtained for gas dynamical shock waves by an accurate use of various Sobolev's imbedding theorems.

Unlike [7, 9, 10], the proof of Theorem 4.2 above relies on the Gagliardo–Nirenberg inequalities and is, therefore, closer to arguments of Metivier [37]. But, if $\Omega = [0,T]\times\mathbb{R}^n_\pm$ or $\Omega = [0,T]\times\mathbb{R}^{n-1}$ the constants in (4.3), (4.4) blow up as $T\to 0$. This unpleasant fact can prevent the proof of existence for the nonlinear problem. To overcome this difficulty it was suggested in [37] to use some substitutes of the Gagliardo–Nirenberg inequalities for which the constants are uniform with respect to T as $T\to 0$ (we do not want to go into details and just refer to [37]). Using such substitutes allows one to prove some modifications of inequalities (4.5)–(4.8) which now include norms of $u(0)$ and $v(0)$ (see [37]). Further arguments in the proof of Theorem 4.2 remain valid (with little modification), and the constant $C(T,M)$ in (4.2) is uniform with respect to T as $T\to 0$.

Consider the system

$$\widetilde{L}(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\mathbf{Y} + C(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\mathbf{Y} = \widetilde{\mathbf{f}}^{\pm}_{p-1}, \tag{4.20a}$$

$$\mathcal{L}(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\mathbf{W} = \boldsymbol{\mathcal{F}}^{\pm}, \quad \mathbf{x}\in\mathbb{R}^n_\pm, \tag{4.20b}$$

where $\mathbf{Y} = (\mathbf{U},\mathbf{W}_1,\dots,\mathbf{W}_{p-1})$, system (4.20a) is formed by (1.24a) and systems obtained by the differentiation of (1.24a) with respect to t and $\mathbf{x}$; the matrix C can be explicitly written out, $\widetilde{\mathbf{f}}^{\pm}_{p-1} = (\partial^{\alpha^1}\mathbf{f}^{\pm},\dots,\partial^{\alpha^{d_0}}\mathbf{f}^{\pm})$, $d_0 = C^{p-1}_{n+p-1}$, etc. (see section 2); system (4.20b) coincides with (4.10) for $|\beta| = 0$. We supplement system (4.20) with the boundary conditions (1.24b). All other boundary conditions follow from (1.24b) and system (4.20) itself at $x_1 = 0$.

System (4.20a) is equivalently rewritten as

$$A_0(\widehat{\mathbf{U}})\mathbf{Y}_t \mp \mathbf{Y}_{x_1} + \sum_{k=2}^{n} A_k(\widehat{\mathbf{U}})\mathbf{Y}_{x_k} + C^{\pm}(\widehat{\mathbf{U}},\widehat{\mathbf{F}})\mathbf{W} = \widetilde{\mathbf{f}}^{\pm}_{p-1}, \quad \mathbf{x}\in\mathbb{R}^n_\pm, \tag{4.21}$$

where $C^{\pm}\mathbf{W} = C\mathbf{Y} + (A_\nu \pm I)\mathbf{Y}_{x_1}$. The boundary matrix for system (4.21), (4.20b) is $\mathrm{diag}(-I, A_\nu)$ for $x_1 > 0$ and $\mathrm{diag}(I, A_\nu)$ for $x_1 < 0$. Clearly, the boundary conditions for system (4.21), (4.20b) are strictly dissipative (see (4.11) for $|\beta| = 0$).

Thus, system (4.21), (4.20b), that is equivalent to (4.20), has strictly dissipative boundary conditions. The initial-boundary-value problem for system (4.21), (4.20b) differs from one studied in Appendix A of [43] only by the presence of the unknown function f in the boundary conditions. The compatibility conditions for (1.24) can be written by analogy with those for standard boundary conditions in [40, 43]. The existence of a smooth solution $(\mathbf{U},\mathbf{W}_1,\dots,\mathbf{W}_{p-1},\mathbf{W})\in W_2^{m-p}$ to the problem for system (4.21), (4.20b) is proved exactly in the same manner as in Appendix A of [43] for linear hyperbolic problems with strictly dissipative boundary conditions. Moreover, the component $\mathbf{U}$ of this solution satisfies the original problem (1.24). Note also that in [10] the existence of smooth solutions to the linearized problem for gas dynamical shock waves was proved by approximation by grid functions. Such an approach suggested by Godunov [25] for linear hyperbolic problems with strictly dissipative boundary conditions and applied by Blokhin [10] to gas dynamical shock waves can also be used for general Lax shocks under consideration. So, we have the following existence theorem.

THEOREM 4.3. *Let that all the assumptions of Theorem* 4.2 *are satisfied. Then, for all the data*

$$\mathbf{U}_0 \in W_2^m(\mathbb{R}^n_+) \cap W_2^m(\mathbb{R}^n_-), \quad f_0 \in W_2^{m+1}(\mathbb{R}^{n-1}),$$

$$\mathbf{f}^{\pm} \in W_2^m([0,T] \times \mathbb{R}^n_{\pm}), \quad \mathbf{g} \in W_2^m([0,T] \times \mathbb{R}^{n-1})$$

satisfying the compatibility conditions up to order $m-1$, *the initial-boundary-value problem* (1.24), (1.17c) *has a unique solution* $(\mathbf{U}, f) \in Z_T^m$ *that obeys the a priori estimate* (4.2).

Sketch of the proof of Theorem 4.1. The proof follows from a fixed-point argument and we are quite brief here. For a time $T > 0$, a constant $M > 0$, and an integer $s \geq [n/2] + 2$, we define

$$\mathcal{K} = \Big\{ (\widehat{\mathbf{U}}, \hat{f}) \in Z_T^s \;\big|\; \mathcal{N}_T^s(\widehat{\mathbf{U}}, \hat{f}) \leq M,\ \widehat{\mathbf{U}}(0,\mathbf{x}) = \mathbf{U}_0(\mathbf{x}),\ \hat{f}(0,\mathbf{x}') = f_0(\mathbf{x}')\,,$$
$$(\mathbf{U}_0, f_0) \in \big\{\textstyle\bigcap_{\pm} W_2^s(\mathbb{R}^n_{\pm})\big\} \times W_2^{s+1}(\mathbb{R}^{n-1}) \text{ is compatible to order } s-1 \Big\}.$$

We do not specify here the compatibility conditions and just refer to [37].

Consider now the mapping $\Lambda : (\widehat{\mathbf{U}}, \hat{f}) \to (\mathbf{U}, f)$, where $(\mathbf{U}, f)$ satisfies the initial-boundary-value problem (1.24), (1.17c) with $\mathbf{f}^{\pm} \equiv 0$ and

$$\mathbf{g} = \big[\boldsymbol{\mathcal{P}}^1(\widehat{\mathbf{U}})\big] - \big[S^{-1}(\widehat{\mathbf{U}})A_\nu(\widehat{\mathbf{U}}, \widehat{\mathbf{F}})\widehat{\mathbf{U}}\big].$$

Actually, with such a choice of $\mathbf{g}$ the linear conditions in (1.24b) are Newton's approximation of the nonlinear boundary conditions (1.17b) (see discussion in [34]). Theorem 4.3 guarantees the existence of $(\mathbf{U}, f) \in Z_T^s$. Moreover, it follows from estimate (4.2) that $\Lambda(\mathcal{K}) \subset \mathcal{K}$ for appropriate choices of T and M (see Remark 4.1).

Consider $(\widehat{\mathbf{U}}^i, \hat{f}^i) \in \mathcal{K}$ and let $(\mathbf{U}, f) = \Lambda(\widehat{\mathbf{U}}^i, \hat{f}^i)$, $i = 1, 2$. For the differences $\mathbf{U}^1 - \mathbf{U}^2$ and $f^1 - f^2$ we obtain problem (1.24) with the trivial initial data, the coefficients $(\widehat{\mathbf{U}}, \hat{f}) = (\widehat{\mathbf{U}}^1, \hat{f}^1)$,

$$\mathbf{f}^{\pm} = \big(L(\widehat{\mathbf{U}}^2, \widehat{\mathbf{F}}^2) - L(\widehat{\mathbf{U}}^1, \widehat{\mathbf{F}}^1)\big)\mathbf{U}^2, \quad \mathbf{x} \in \mathbb{R}^n_{\pm},$$

and a corresponding $\mathbf{g}$ (it can be easily written down as well). Applying estimate (4.2) with $m = s-1$ to this problem and using the mean-value theorem for $\mathbf{f}^{\pm}$ and $\mathbf{g}$, one gets

$$\mathcal{N}_T^{s-1}(\mathbf{U}^1 - \mathbf{U}^2, f^1 - f^2) \leq \delta\, \mathcal{N}_T^{s-1}(\widehat{\mathbf{U}}^1 - \widehat{\mathbf{U}}^2, \hat{f}^1 - \hat{f}^2),$$

where the positive constant $\delta = \delta(T, M) < 1$ for T sufficiently small (we do not describe in detail the choice of T and M and just refer to standard arguments, for example, in [34] for the Cauchy problem or in [41, 42] for initial-boundary-value problems).

That is, the mapping Λ is a contraction in the low norm $\mathcal{N}_T^{s-1}$. Hence, there exists a unique fixed point $(\mathbf{U}, f) = (\widehat{\mathbf{U}}, \hat{f}) \in \mathcal{K}$ which solves (1.17). ∎

5. Concluding remarks. By introducing the notations of dissipative and strictly dissipative p-symmetrizers we have formalized the energy method applied earlier to strong discontinuities for concrete hyperbolic systems of conservation laws. We have proved that if the constant coefficients linearized problem for Lax shocks has a strictly dissipative p-symmetrizer, then under natural assumptions this implies the local in time existence of shock-front solutions of the original nonlinear system. This result recovers Blokhin's local existence theorem for gas dynamical shock waves [7, 9] and enables one to conclude the local existence of shock-front solutions for various concrete models (MHD [12], radiation hydrodynamics [3, 15], Landau's equations of superfluid [11], etc.) for which a priori estimates with no loss of derivatives for constant coefficients linearized problems were earlier deduced by the energy method.

It seems that the result of Theorem 4.1 could be extended, under appropriate assumptions, to the case of characteristic discontinuities. Note, however, that we do not know any concrete example of a characteristic discontinuity for which one can construct a strictly dissipative p-symmetrizer. Evidently, this is because all the known characteristic discontinuities (vortex sheets, current-vortex sheets, Alfvén discontinuities, etc.) can be only neutrally stable, i.e., the uniform Lopatinski condition is never satisfied for them.

With regard to the case where the loss of derivatives phenomenon takes place, i.e., when we are able to construct only a dissipative (but not strictly dissipative) p-symmetrizer, a theorem for the variable coefficients linearized problem like Theorem 4.3 could be proved both for Lax shocks and characteristic discontinuities. In the generic case, for characteristic discontinuities the functional setting is provided by the anisotropic weighted Sobolev spaces H^s_* (see [42] and references therein). Concerning the proof of a local existence theorem, it seems that the only way to overcome difficulties connected with the loss of derivatives phenomenon is the use of the Nash–Moser method (see discussion in Remark 1.2). Note that the existence of a dissipative (but not strictly dissipative) p-symmetrizer implies the fulfillment of the (weak) Lopatinski condition, i.e., the weak stability of a corresponding strong discontinuity. In particular, the existence of a dissipative p-symmetrizer for a planar Lax shock implies the weak stability of this shock wave and, in view of the recent result of Coulombel and Secchi [20] (see Remark 1.2), the nonlinear existence of nonplanar shock waves that are close to the planar shock under consideration.

Unfortunately, there is not a general procedure to construct a p-symmetrizer. At the same time, if it was somehow constructed, we do not need to examine the Lopatinski condition, which is often untestable analytically (numerical testing is usually not so simple either). The requirements for a set $\mathbb{S}$ to be a (strictly) dissipative p-symmetrizer suggest sufficient or, sometimes, necessary and sufficient conditions for the fulfillment of the (uniform) Lopatinski condition. In this connection, the best example for illustration is the construction of the dissipative 0-symmetrizer for current-vortex sheets [45] that first enabled the finding of wide sufficient conditions for their neutral stability (i.e., sufficient conditions of the macroscopic stability of the heliopause [4]).

For Lax shock waves, the construction of a strictly dissipative p-symmetrizer can be interpreted as an indirect test of the uniform Lopatinski condition and, referring then to [37, 38], we have at once the local existence theorem for the nonlinear problem. However, to construct Kreiss' symmetrizer for the case of characteristic discontinuities it is necessary to know not only that the Lopatinski condition is satisfied but also *how* it is satisfied, i.e., to know a detailed structure of the Lopatinski determinant (see [19]). For example, for current-vortex sheets [45] we know sufficient conditions for the fulfillment of the Lopatinski condition, but the structure of the Lopatinski determinant cannot be analyzed because of insuperable technical difficulties. That is, the only way to achieve a nonlinear result is to follow the energy method in the variable coefficients and nonlinear analysis as well. In the light of this, we think that future perspectives of the method of p-symmetrizers are connected with “nonstandard” problems for which either Kreiss' symmetrizer technique does not work for technical reasons or the general theory is still not developed (as, for example, for nonhyperbolic problems appearing for incompressible fluids; see [46]).

Acknowledgments. The author gratefully thanks Alexander Blokhin for many helpful discussions. The author is indebted to Paolo Secchi for providing him with the reference to [43]. Also, the author thanks Konstantin Ilin and Vladimir Vladimirov for their kind hospitality during his stay at the Hull Institute of Mathematical Sciences and Applications of the University of Hull.

REFERENCES

[1] M. S. Agranovich, *Theorem on matrices depending on parameters and its applications to hyperbolic systems*, Funct. Anal. Appl., 6 (1972), pp. 85–93.

[2] S. Alinhac, *Existence d'ondes de raréfaction pour des systèmes quasi-linéaires hyperboliques multidimensionnels*, Comm. Partial Differential Equations, 14 (1989), pp. 173–230.

[3] A. M. Anile, A. M. Blokhin, and Yu. L. Trakhinin, *Investigation of a mathematical model for radiation hydrodynamics*, Z. Angew. Math. Phys., 50 (1999), pp. 677–697.

[4] V. B. Baranov, K. V. Krasnobaev, and A. G. Kulikovsky, *A model of interaction of the solar wind with the interstellar medium*, Soviet Phys. Dokl., 15 (1970), pp. 791–793.

[5] R. Bellman, *Introduction to Matrix Analysis*, Classics Appl. Math. 19, SIAM, Philadelphia, 1997.

[6] A. M. Blokhin, *A mixed problem for a system of equations of acoustics with boundary conditions on a shock wave*, Izv. Sibirsk. Otdel. Akad. Nauk SSSR Ser. Tekhn. Nauk, 13 (1979), pp. 25–33 (in Russian).

[7] A. M. Blokhin, *Estimation of the energy integral of a mixed problem for gas dynamics equations with boundary conditions on the shock wave*, Siberian Math. J., 22 (1981), pp. 501–523.

[8] A. M. Blokhin, *A mixed problem for symmetric t-hyperbolic systems of acoustic type*, Dinamika Sploshn. Sredy, 52 (1981), pp. 11–29 (in Russian).

[9] A. M. Blokhin, *Uniqueness of the classical solution of a mixed problem for equations of gas dynamics with boundary conditions on a shock wave*, Siberian Math. J., 23 (1982), pp. 604–615.

[10] A. M. Blokhin, *Energy Integrals and Their Applications in Problems in Gas Dynamics*, Nauka, Novosibirsk, Russia, 1986 (in Russian).

[11] A. M. Blokhin and V. N Dorovsky, *Mathematical Modelling in the Theory of Multivelocity Continuum*, Nova Science, New York, 1995.

[12] A. Blokhin and Yu. Trakhinin, *Investigation of the well-posedness of the mixed problem on the stability of fast shock waves in magnetohydrodynamics*, Matematiche (Catania), 49 (1994), pp. 123–141.

[13] A. Blokhin and Yu. Trakhinin, *Stability of strong discontinuities in fluids and MHD*, in Handbook of Mathematical Fluid Dynamics, Vol. 1, S. Friedlander and D. Serre, eds., North-Holland, Amsterdam, 2002, pp. 545–652.

[14] A. Blokhin and Yu. Trakhinin, *Stability of Strong Discontinuities in Magnetohydrodynamics and Electrohydrodynamics*, Nova Science, New York, 2003.

[15] A. M. Blokhin, V. Romano, and Yu. L. Trakhinin, *Stability of shock waves in relativistic radiation hydrodynamics*, Ann. Inst. H. Poincaré Phys. Théor., 67 (1997), pp. 145–180.

[16] E. Casella, P. Secchi, and P. Trebeschi, *Non-homogeneous linear symmetric hyperbolic systems with characteristic boundary*, Differential Integral Equations, 19 (2006), pp. 51–74.

[17] J.-F. Coulombel, *Weakly stable multidimensional shocks*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 21 (2004), pp. 401–443.

[18] J.-F. Coulombel, *Well-posedness of hyperbolic initial boundary value problems*, J. Math. Pures Appl., 84 (2005), pp. 786–818.

[19] J.-F. Coulombel and P. Secchi, *The stability of compressible vortex sheets in two space dimensions*, Indiana Univ. Math. J., 53 (2004), pp. 941–1012.

[20] J.-F. Coulombel and P. Secchi, *Nonlinear compressible vortex sheets in two space dimensions*, Seminario Matematico, Brescia, preprint, 2005.

[21] J. Francheteau and G. Métivier, *Existence de chocs faibles pour des systèmes quasi-linéaires hyperboliques multidimensionnels*, Astérisque, 268, Soc. Math. France, Paris, 2000.

[22] K. O. Friedrichs, *Symmetric positive linear differential equations*, Comm. Pure Appl. Math., 11 (1958), pp. 333–418.

[23] S. K. Godunov, *An interesting class of quasi-linear systems*, Soviet Math. Dokl., 2 (1961), pp. 947–948.

[24] S. K. Godunov, *Symmetric form of the equations of magnetohydrodynamics*, in Numerical Methods for Continuum Mechanics, Vol. 3, Computer Center of the Siberian Branch of the USSR Academy of Sciences, Novosibirsk, 1972, pp. 26–34 (in Russian). English translation at http://citeseer.ist.psu.edu/37221.html.

[25] S. K. Godunov, *Equations of Mathematical Physics*, Nauka, Moscow, 1979 (in Russian).

[26] T. Kato, *The Cauchy problem for quasi-linear symmetric hyperbolic systems*, Arch. Ration. Mech. Anal., 58 (1975), pp. 181–205.

[27] H.-O. Kreiss, *Initial boundary value problems for hyperbolic systems*, Comm. Pure Appl. Math., 23 (1970), pp. 277–296.

[28] L. D. Landau and E. M. Lifshiz, *Electrodynamics of Continuous Media. Course of Theoretical Physics*, Vol. 8, Pergamon Press, Oxford, 1984.

[29] L. D. Landau and E. M. Lifshiz, *Fluid Mechanics. Course of Theoretical Physics*, Vol. 6, Pergamon Press, New York and Oxford, 1987.

[30] P. D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CMBS-NSF Regional Conf. Ser. in Appl. Math. 11, SIAM, Philadelphia, 1973.

[31] D. Li, *Rarefaction and shock waves for multi-dimensional hyperbolic conservation laws*, Comm. Partial Differential Equations, 16 (1991), pp. 425–450.

[32] A. Majda, *The Stability of Multi-Dimensional Shock Fronts — A New Problem for Linear Hyperbolic Equations*, Mem. Amer. Math. Soc. 41, no. 275, Amer. Math. Soc., Providence, RI, 1983.

[33] A. Majda, *The Existence of Multi-Dimensional Shock Fronts*, Mem. Amer. Math. Soc. 43, no. 281, Amer. Math. Soc., Providence, RI, 1983.

[34] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Springer-Verlag, New York, 1984.

[35] A. Majda and S. Osher, *Initial-boundary value problems for hyperbolic equations with uniformly characteristic boundary*, Comm. Pure Appl. Math., 28 (1975), pp. 607–675.

[36] G. Métivier, *The block structure condition for symmetric hyperbolic systems*, Bull. London Math. Soc., 32 (2000), pp. 689–702.

[37] G. Métivier, *Stability of multidimensional shocks*, in Advances in the Theory of Shock Waves, Progr. Nonlinear Differential Equations Appl. 47, Birkhäuser, Boston, 2001, pp. 25–103.

[38] G. Métivier and K. Zumbrun, *Hyperbolic boundary value problems for symmetric systems with variable multiplicities*, J. Differential Equations, 211 (2005), pp. 61–134.

[39] J. Rauch, *Symmetric positive systems with boundary characteristic of constant multiplicity*, Trans. Amer. Math. Soc., 291 (1985), pp. 167–187.

[40] J. B. Rauch and F. J. Massey, *Differentiability of solutions to hyperbolic initial-boundary value problems*, Trans. Amer. Math. Soc., 189 (1974), pp. 303–318.

[41] S. Schochet, *The compressible Euler equations in a bounded domain: Existence of solutions and the incompressible limit*, Comm. Math. Phys., 104 (1986), pp. 49–75.

[42] P. Secchi, *Well-posedness of characteristic symmetric hyperbolic systems*, Arch. Ration. Mech. Anal., 134 (1996), pp. 155–197.

[43] P. Secchi, *Inflow-outflow problems for inviscid compressible fluids*, Commun. Appl. Anal., 2 (1998), pp. 81–110.

[44] Yu. Trakhinin, *A complete 2D stability analysis of fast MHD shocks in an ideal gas*, Comm. Math. Phys., 236 (2003), pp. 65–92.

[45] Yu. Trakhinin, *On existence of compressible current-vortex sheets: Variable coefficients linear analysis*, Arch. Ration. Mech. Anal., 177 (2005), pp. 331–366.

[46] Yu. Trakhinin, *On the existence of incompressible current-vortex sheets: Study of a linearized free boundary value problem*, Math. Methods Appl. Sci., 28 (2005), pp. 917–945.

[47] A. I. Vol'pert and S. I. Khudyaev, *On the Cauchy problem for composite systems of nonlinear differential equations*, Math. USSR-Sb., 16 (1972), pp. 517–544.

SIAM J. MATH. ANAL.
Vol. 37, No. 6, pp. 2025–2048

SURFACTANT SPREADING ON THIN VISCOUS FILMS: NONNEGATIVE SOLUTIONS OF A COUPLED DEGENERATE SYSTEM*

HARALD GARCKE† AND SANDRA WIELAND‡

Abstract. We consider the Navier–Stokes system for an incompressible fluid coupled with a convection-diffusion equation for surfactant molecules on the free surface. The lubrication approximation leads to a coupled system of parabolic equations, consisting of a degenerate fourth-order equation for the film height and a second-order equation for the surfactant concentration. A proof based on energy estimates shows the existence of global weak solutions which in addition fulfill an integral inequality (entropy condition) which ensures positivity properties for the solution.

Key words. partial differential equations, degenerate parabolic equation, thin liquid film, surfactant spreading, free surface, fluid interface.

AMS subject classifications. 35K55, 35K65, 35K35, 76A20, 76D08

DOI. 10.1137/040617017

1. Introduction.

The aim of this paper is to prove the existence of nonnegative, weak solutions of the following system of coupled nonlinear, degenerate parabolic partial differential equations

$$h_t + \left(\frac{1}{3}\, h^3\, h_{xxx} + \frac{1}{2}\, h^2\, \sigma(\Gamma)_x\right)_x = 0, \tag{1.1}$$

$$\Gamma_t + \left(\frac{1}{2}\, h^2\, \Gamma\, h_{xxx} + h\, \Gamma\, \sigma(\Gamma)_x\right)_x = D\, \Gamma_{xx} \tag{1.2}$$

with suitable initial and boundary conditions. The above system appears in the lubrication theory for thin films on which surfactant molecules diffuse (see [BG88], [DeG94], [GG90], [JG92], [MT99], [WJ01]). In (1.1)–(1.2) the function h describes the height of the film and Γ is the concentration of the surfactants. The monotone decreasing function σ models that the surfactant molecules lower the surface tension. The first equation follows from mass conservation for the fluid and the term in brackets is the total horizontal velocity. The second equation is a convection-diffusion equation describing mass balance for the surfactants. The term in brackets in (1.2) is the horizontal velocity on top of the film times Γ and hence this term accounts for transport of Γ induced by the velocity field. We will discuss these issues in more detail in section 2.

The analysis for (1.1)–(1.2) is difficult due to the fact that the system degenerates

*Received by the editors October 14, 2004; accepted for publication (in revised form) November 8, 2005; published electronically March 24, 2006. This work has been supported by the DFG (Deutsche Forschungsgemeinschaft) through the Schwerpunktprogramm 611 Singular Phenomena and Scaling in Mathematical Models.

http://www.siam.org/journals/sima/37-6/61701/html

†NWF I–Mathematik, Universität Regensburg, 93040 Regensburg, Germany (harald.garcke@mathematik.uni-regensburg.de).

‡Department of Mathematics, Duke University, Box 90320, Durham, NC 27708-0320 (sandra@math.duke.edu). The research of this author was partially supported by NSF grants NSF FRG DMS 0073841 and NSF FRG DMS 0244498.

as h tends to zero. Equation (1.1) with $\sigma =$ constant is the thin film equation

$$h_t + \left(\frac{1}{3}h^3\, h_{xxx}\right)_x = 0 \tag{1.3}$$

which has been studied by many authors (see [BF90], [B96], [BBD95], [BMS99], [BP96], [DG01], [DGG98], [EG96], [G03], [O98] and the references therein). Due to the fact that the equation is of fourth order and since no maximum principle is valid, it is difficult to analyze the thin film equation. For example, it is not clear how to show nonnegativity of solutions. For the thin film equation (1.3) a priori estimates follow from the identity

$$\frac{d}{dt}\int_\Omega \frac{1}{2}h_x^2 + \int_\Omega \frac{1}{3}h^3\, h_{xxx}^2 = 0 \tag{1.4}$$

which holds under appropriate boundary conditions. But also further integral estimates, of which

$$\frac{d}{dt}\int_\Omega G(h) + \frac{1}{3}\int_\Omega h\, h_{xx}^2 = 0 \tag{1.5}$$

with $G''(h) = h^{-2}$ is the easiest example, hold and give further a priori estimates (see [BF90], [BBD95], [BP96], [DGG98] for details). By now many deep results such as nonnegativity of solutions, finite speed of propagation, results on the long time behavior of solutions, waiting time behavior and regularity results have been shown by using global and local versions of the above-mentioned a priori estimates (see [BF90], [BBD95], [B96], [DGG98]). The mathematical analysis of the system (1.1)–(1.2) is even more involved. Renardy ([R1-96], [R2-96], [R97]) studied this system and variants of it. He showed local existence results and studied shock profiles in certain singular perturbed variants of (1.1)–(1.2). Barrett, Garcke and Nürnberg [BGN03] studied and analyzed a finite element method for (1.1)–(1.2) and they present several numerical simulations showing an extreme thinning of the film due to convection resulting from surface tension gradients. We also refer to Grün, Lenz and Rumpf [GLR02] for numerical simulations based on a finite volume method.

In this paper we will show global existence of weak solutions to (1.1)–(1.2). Fundamental for our approach is a proper generalization of the energy identity (1.4) to the case of surfactants. This is not straightforward and, therefore, we will reconsider the derivation for (1.1)–(1.2) from the full free boundary problem for the Navier–Stokes equations. In particular we will derive an energy inequality for the full problem and taking the scaling of the lubrication approximation into account we can derive an energy estimate for (1.1)–(1.2). This part of the paper is formal and will be presented in section 2. The formally derived energy estimates will then be used in sections 3 and 4 to derive rigorous a priori estimates which are basic ingredients of the existence theory for (1.1)–(1.2). Generalization of the integral identities ("entropy" identities) for the thin film equation in general do not seem to hold for (1.1)–(1.2). But in section 4 we will show that at least one of the "entropy" estimates still can be generalized to (1.1)–(1.2) and this will allow us to show that solutions to positive initial data can only become zero on a set of measure zero. This shows that the influence of surfactants does not lead to dead cores, i.e., sets with positive measure on which h becomes zero.

2. The models. The motion of an incompressible viscous fluid on a bounded solid substrate is governed by the Navier–Stokes equations

$$\rho_0\{\mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{u}\} = \mu\,\Delta\mathbf{u} - \nabla p, \tag{2.1}$$

$$\operatorname{div}\,\mathbf{u} = 0, \tag{2.2}$$

where $\mathbf{u} = (u_1, u_2, u_3)$ is the velocity field, p the pressure and ρ_0 the constant density. We assume that the fluid does not penetrate the substrate given by $\{(x, y, z) \in \mathbb{R}^3 | z = 0\}$ and that there is no flux across the lateral boundary. We also impose no-slip boundary conditions for the velocity, i.e.,

$$\mathbf{u}_{,z=0} = 0. \tag{2.3}$$

We assume that the evolving free surface (i.e., the fluid/air interface) C_t is given as the graph of a smooth time-dependent height function $h = h(t, x, y)$ on a spatial domain $\Omega \subset \mathbb{R}^2$, i.e.,

$$C_t = \{(x, y, z) \big| \,(x, y) \in \Omega,\ z = h(t, x, y)\}. \tag{2.4}$$

At the interface (briefly abbreviated as $\{z = h\}$) a kinematic boundary condition dictates that the normal component of the liquid velocity balances the speed of the interface (the convective time-derivative $\frac{dh}{dt}$ of h), i.e.,

$$u_{3,z=h} = \frac{dh}{dt} = h_t + u_1\,h_x + u_2\,h_y. \tag{2.5}$$

In order to balance the shear stress tensor

$$T(\mathbf{u}, p) := -p\,\mathrm{Id} + \mu\left(\nabla\mathbf{u} + \nabla\mathbf{u}^T\right)$$

of the fluid we look at its normal and tangential component separately. The normal stress exhibits a jump equal to the surface tension times the curvature (later referred to as normal stress condition):

$$\boldsymbol{\nu}\cdot T\boldsymbol{\nu} = \sigma\kappa, \tag{2.6}$$

where $\boldsymbol{\nu}$ is the outer unit normal and κ is the sum of the principal curvatures of the interface at a given point on C_t.

For the description of the tangential shear stress on the interface it becomes important to clarify the role of the surface tension dependence on the surfactant concentration Γ more explicitly. A result of the surface agency of the surfactant monolayer is that the surface concentration gradient of surfactant molecules opposes the surface tension gradient. To restore the disturbed equilibrium at the surface (caused by inhomogeneously distributed surfactants), the tangential part of the surface tension gradient balances the tangential component of the shear stress $T\boldsymbol{\nu}$, i.e.,

$$\boldsymbol{\tau}\cdot T\boldsymbol{\nu} = \partial_{\boldsymbol{\tau}}\,\sigma(\Gamma) \tag{2.7}$$

for all tangents $\boldsymbol{\tau}$, where $\partial_{\boldsymbol{\tau}}$ denotes the gradient in direction of the tangent vector $\boldsymbol{\tau}$. This effect is known as the (solutal) Marangoni effect.

2.1. Mass balance for the surfactant concentration. We now derive the evolution equation for the surfactant concentration by looking at the mass balance for the surfactant on subsets $C'_t \subset C_t$. Let

$$C' = \bigcup_{t\in[t_1,t_2]} \{t\} \times C'_t \quad \text{and} \quad C = \bigcup_{t\in[t_1,t_2]} \{t\} \times C_t$$

be a smooth three-dimensional surface with boundary $\partial C'$, such that $C'_t \subset \mathbb{R}^2$ are smooth two-dimensional surfaces with boundary $\partial C'_t$. Let Γ be the concentration of the surfactant which is a function on C. Then the mass balance on C' reads as follows:

$$\frac{d}{dt}\left(\int_{C'_t} \Gamma \, dS^2\right) = -\int_{\partial C'_t} \Big(\Gamma \mathbf{u}_{tan} - D\nabla_s \Gamma\Big) \cdot \mathbf{n}_{\partial C'_t} \, dS^1 + \int_{\partial C'_t} \Gamma \, v_{\partial C'_t} \, dS^1. \tag{2.8}$$

Here $\mathbf{u}_{tan}$ is the tangential component of the velocity, i.e.,

$$\mathbf{u}_{tan} := \mathbf{u} - (\mathbf{u}\cdot\boldsymbol{\nu})\boldsymbol{\nu},$$

where $\boldsymbol{\nu} \in \mathbb{R}^3$ is the normal to the interface C'_t. Furthermore, $\nabla_s \Gamma$ denotes the surface gradient of Γ on C_t. The vector $\mathbf{n}_{\partial C'_t}$ is the outer unit normal to $\partial C'_t$, i.e., $\mathbf{n}_{\partial C'_t}$ lies in the tangent space to C'_t and is perpendicular to $\partial C'_t$. The quantity $v_{\partial C'_t}$ is the normal boundary velocity of $\partial C'_t$ as a subcurve of C'_t, i.e., $v_{\partial C'_t}$ describes the local decrease or increase of the surface area of C'_t due to the tangential velocity of the boundary $\partial C'_t$. To compute $v_{\partial C'_t}$ at a point $x \in \partial C'_t$ let $y(s) \in \partial C'_t$ be such that $y(t) = x$. Then $y'(t)$ denotes the velocity of the curve y. Any motion tangential to $\partial C'_t$ does not lead to an increase of area, i.e., only the component

$$v_{\partial C'_t} := y'(t) \cdot \mathbf{n}_{\partial C'_t}$$

changes the surface area of C'_t.

Let us briefly discuss what the two terms on the right-hand side of (2.8) describe. The first integral describes changes of mass due to convective transport by the velocity $\mathbf{u}$ and diffusional transport where $D > 0$ is the diffusion coefficient. The second integral takes into account the change of surfactant on C'_t due to the fact that the surface C'_t increases or decreases. Our goal is now to derive a pointwise identity for the surfactant concentration from the balance law (2.8). To reformulate (2.8) we need the transport theorem (for a proof see the appendix)

$$\frac{d}{dt}\left(\int_{C'_t} \Gamma \, dS^2\right) = \int_{C'_t} \Big[\partial_{(1,\mathbf{v}_{\boldsymbol{\nu}})}\Gamma - \Gamma(\mathbf{v}_{\boldsymbol{\nu}} \cdot \boldsymbol{\kappa}_{\boldsymbol{\nu}})\Big] dS^2 + \int_{\partial C'_t} \Gamma \, v_{\partial C'_t} \, dS^1. \tag{2.9}$$

In the above $\partial_{(1,\mathbf{v}_{\boldsymbol{\nu}})}\Gamma$ denotes the partial derivative of Γ in the direction $(1, \mathbf{v}_{\boldsymbol{\nu}})$. Here $\mathbf{v}_{\boldsymbol{\nu}}$ is the uniquely determined vector lying in the tangent space to C'_t such that $(1, \mathbf{v}_{\boldsymbol{\nu}})$ is tangent to C'. The quantity $\mathbf{v}_{\boldsymbol{\nu}}$ is the normal velocity vector and $\partial_{(1,\mathbf{v}_{\boldsymbol{\nu}})}\Gamma$ is the normal time-derivative in the notation of Gurtin [Gur93]. Furthermore $\boldsymbol{\kappa}_{\boldsymbol{\nu}}$ is the mean curvature vector, i.e., $\boldsymbol{\kappa}_{\boldsymbol{\nu}}$ has the direction of the normal $\boldsymbol{\nu}$ and its length is equal to the sum of the principal curvatures. Combining (2.8) and (2.9) and using the Gauss theorem for the first integral on the right-hand side of (2.8) we obtain

$$0 = \int_{C'_t} \Big[\partial_{(1,\mathbf{v}_{\boldsymbol{\nu}})}\Gamma - \Gamma(\mathbf{v}_{\boldsymbol{\nu}} \cdot \boldsymbol{\kappa}_{\boldsymbol{\nu}}) + \nabla_s \cdot (\Gamma \, \mathbf{u}_{tan} - D\nabla_s \Gamma)\Big] dS^2. \tag{2.10}$$

Since C'_t is arbitrary we obtain

$$\partial_{(1,\mathbf{v}_{\boldsymbol{\nu}})}\Gamma - \Gamma(\mathbf{v}_{\boldsymbol{\nu}} \cdot \boldsymbol{\kappa}_{\boldsymbol{\nu}}) + \nabla_s \cdot (\Gamma \mathbf{u}_{tan} - D\nabla_s \Gamma) = 0 \tag{2.11}$$

pointwise on C'. A simple computation shows

$$\mathbf{v}_{\boldsymbol{\nu}} = \frac{h_t}{1+|\nabla h|^2}(-\nabla h, 1) = \frac{h_t}{\sqrt{1+|\nabla h|^2}}\boldsymbol{\nu},$$

where the normal $\boldsymbol{\nu}$ is given by

$$\boldsymbol{\nu} = \frac{1}{\sqrt{1+|\nabla h|^2}}(-\nabla h, 1).$$

Hence we obtain from (2.5) that $\mathbf{u} \cdot \boldsymbol{\nu} = \mathbf{v}_{\boldsymbol{\nu}} \cdot \boldsymbol{\nu}$ and therefore

$$\mathbf{v}_{\boldsymbol{\nu}} \cdot \boldsymbol{\kappa}_{\boldsymbol{\nu}} = \mathbf{u} \cdot \boldsymbol{\kappa}_{\boldsymbol{\nu}} = (\mathbf{u} \cdot \boldsymbol{\nu})\kappa.$$

As a result (2.11) can be rewritten as

$$\partial_{(1,\mathbf{v}_{\boldsymbol{\nu}})}\Gamma - \Gamma(\mathbf{u} \cdot \boldsymbol{\nu})\kappa + \nabla_s \cdot (\Gamma \mathbf{u}_{tan} - D\,\nabla_s \Gamma) = 0. \tag{2.12}$$

Using $\mathbf{u} = \mathbf{u}_{tan} + (\mathbf{u} \cdot \boldsymbol{\nu})\boldsymbol{\nu}$ and the fact that $\nabla_s(\Gamma\mathbf{u} \cdot \boldsymbol{\nu}) \cdot \boldsymbol{\nu} = 0$ we obtain using the following sign convention for the curvature $\kappa = -\nabla_s \cdot \boldsymbol{\nu}$ the identity

$$\nabla_s \cdot (\Gamma \mathbf{u}) = \nabla_s \cdot (\Gamma \mathbf{u}_{tan}) - \Gamma(\mathbf{u} \cdot \boldsymbol{\nu})\kappa$$

and finally (2.12) takes the form

$$\partial_{(1,\mathbf{v}_{\boldsymbol{\nu}})}\Gamma + \nabla_s \cdot (\Gamma \mathbf{u}) = D\,\Delta_s \Gamma. \tag{2.13}$$

2.2. Energy for the free surface problem. Assuming constant temperature, a fundamental equation of chemical thermodynamics relates the concentration-dependent surface tension σ to the free energy $g(\Gamma)$ and the chemical potential $g'(\Gamma)$ (see [V00]), where both functions depend on the surfactant concentration Γ:

$$\sigma(\Gamma) = g(\Gamma) - \Gamma\, g'(\Gamma). \tag{2.14}$$

Convexity of the free energy implies a monotone decrease of surface tension for non-negative concentration, which is consistent with the surface agency of the surfactant molecules we described before.

The total energy decomposes into the kinetic energy of the flow and the free energy of the surface:

$$E(t) = \int_{\Omega_t} \frac{\rho_0}{2}\mathbf{u}^2(t,\cdot)\, dV^3 + \int_{C_t} g(\Gamma(t,\cdot))\, dS^2, \tag{2.15}$$

where $\Omega_t := \left\{(x,y,z) \in \mathbb{R}^3 \middle| \; (x,y) \in \Omega,\; 0 < z < h(t,x,y) \right\}$ is the domain occupied by the liquid. A careful computation (see appendix 5.2) reveals the dissipation of the total energy:

$$\frac{d}{dt}E(t) = -\int_{\Omega_t} \nabla \mathbf{u} : T\, dV^3 - D \int_{C_t} g''(\Gamma)\, |\nabla_s \Gamma|^2 dS^2. \tag{2.16}$$

Here the imposed boundary conditions (2.3), (2.5), (2.6) and (2.7), a no-slip boundary condition at the lateral boundary and a 90° angle condition at points where the free surface intersects the lateral boundary have been used. To avoid additional boundary contributions we choose a 90° angle condition. In a more general situation the free energy has to be supplemented by terms representing the surface energy of the boundary. The importance of (2.16) for the analysis lies in the fact that with integration over a bounded time interval $[0, T]$ a priori estimates for $\mathbf{u}, \Gamma$ and their spatial gradients result.

2.3. Lubrication approximation. As our subsequent analysis will be restricted to two spatial dimensions (x, z) we state the following derivation in the two-dimensional setting and introduce the variable $\mathbf{u} = (u, w)$ where u represents the horizontal velocity in the direction of x and w the vertical velocity in the direction of z.

Scalings appropriate for a lubrication approximation of surfactant driven thin films (see [GG90]) are

$$\hat{x} = \frac{1}{L}\, x, \quad \hat{z} = \frac{1}{H}\, z, \quad \hat{t} = \frac{\epsilon U}{H}\, t,$$

$$\hat{u} = \frac{1}{U}\, u, \quad \hat{w} = \frac{1}{\epsilon U}\, w, \quad \hat{h} = \frac{1}{H}\, h, \quad \hat{\Gamma} = \frac{1}{\Gamma_m}\, \Gamma,$$

where L represents the typical horizontal length scale, H the typical film height, U the typical horizontal velocity and Γ_m the critical surfactant concentration. Considering thin films we assume that the parameter

$$\epsilon = \frac{H}{L}$$

is small. We also need to scale the pressure and the surface energy density and in this context the scaling

$$\hat{p} = \frac{H}{S}\, p \quad \text{and} \quad \hat{\sigma} = \frac{1}{S}\big(\sigma - \sigma_m\big) \tag{2.17}$$

has been chosen (see [GG90], [JG92]). Here S is the spreading coefficient, i.e., the surface tension difference at the surfactant monolayer edge and σ_m the surface tension of the saturated surface. Defining the typical horizontal velocity as

$$U = \frac{\epsilon S}{\mu},$$

with μ being the dynamic viscosity of the fluid, ensures that the Marangoni force at the free surface remains as a dominant force in the tangential stress boundary condition.

Lubrication theory now gives

$$p(t,x,z) = -\mathcal{S}\, h_{xx}(t,x), \tag{2.18}$$

$$u(t,x,z) = z\, u_z(t,x,h(t,x)) + p_x(t,x,z)\left(\frac{1}{2}z^2 - h(t,x)\, z\right), \tag{2.19}$$

where

$$u_z(t,x,h(t,x)) = \sigma(\Gamma(t,x))_x$$

and $\mathcal{S}$ denotes the rescaled capillary constant. The incompressibility of the flow and the evolution equation for Γ imply to leading order

$$h_t = -\partial_x \left(\int_0^{h(t,x)} u(t,x,z)\, dz \right), \tag{2.20}$$

$$\Gamma_t = -\partial_x \Big(\Gamma(t,x)\, u(t,x,h(t,x)) \Big) + D\Gamma_{xx} \tag{2.21}$$

(see [GG90] for details). Using the representation of u in (2.19) we get the system (1.1)–(1.2) which we are going to analyze in what follows. Furthermore we take the scaling of the lubrication approximation to approximate the energy inequality (2.16) which gives

$$\frac{d}{dt} \int_\Omega \left(\frac{\mathcal{S}}{2} h_x^2 + g(\Gamma) \right) + \int_\Omega \int_0^{h(t,x)} u_z^2 + \mathcal{D} \int_\Omega g''(\Gamma)\, \Gamma_x^2 = 0, \tag{2.22}$$

where $\mathcal{D}$ denotes the rescaled diffusion constant. Making use of the representation (2.19) we obtain

$$\begin{aligned} \frac{d}{dt} \int_\Omega \left(\frac{\mathcal{S}}{2} h_x^2 + g(\Gamma) \right) + \int_\Omega h \left\{ (\sigma(\Gamma)_x)^2 + \mathcal{S}\, h\, h_{xxx}\, \sigma(\Gamma)_x + \frac{1}{3} h^2 \left| \mathcal{S} h_{xxx} \right|^2 \right\} \\ + \mathcal{D} \int_\Omega g''(\Gamma)\, \Gamma_x^2 = 0. \end{aligned} \tag{2.23}$$

We remark that the second term in brackets is positive since the middle term can be absorbed by the first and the third term using Young's inequality. As a result the above identity gives a priori estimates for h and Γ. These estimates will be crucial for the analysis presented in the following sections.

We observe that through lubrication approximation the complexity of the free boundary problem is reduced—compensated by higher-order spatial derivatives for h.

3. Existence of weak solutions. We are interested in solutions of the system of nonlinear partial differential equations we derived in section 2.3, namely,

$$h_t + \partial_x \Big(a_2(h)\, \sigma(\Gamma)_x + a_3(h)\, h_{xxx} \Big) = 0 \qquad \text{in } \Omega_T, \tag{3.1}$$

$$\Gamma_t + \partial_x \Big(\Gamma\, a_1(h)\, \sigma(\Gamma)_x + \Gamma\, a_2(h)\, h_{xxx} \Big) = \mathcal{D}\, \Gamma_{xx} \quad \text{in } \Omega_T, \tag{3.2}$$

where we impose the no-flux and initial boundary conditions

$$h_x = h_{xxx} = \Gamma_x = 0 \quad \text{on } (0,T) \times \partial\Omega\,, \tag{3.3}$$

$$h(0,\cdot) = h_0 \quad \text{in } \Omega\,, \tag{3.4}$$

$$\Gamma(0,\cdot) = \Gamma_0 \quad \text{in } \Omega. \tag{3.5}$$

In sections 3 and 4 the set Ω will always be assumed to be a bounded interval. For later use we introduce the coefficient functions a_i for $i = 1, 2, 3$, which will be such that the special case $a_i(s) = \frac{1}{i} s^i$ for $s \geq 0$ leads to (1.1)–(1.2). It turns out that solutions of appropriate regularity exist in a weak sense. This result is formulated in the following theorem. Before we state the result we need to formulate certain assumptions on the coefficients. In detail we assume:

(A1) The functions $a_i : \mathbb{R} \to \mathbb{R}_0^+$ are continuous for $i = 1, 2, 3$ and $a_i(s) = 0$ if $s \le 0$.

(A2) For $A : \mathbb{R} \to Mat_{2,2}(\mathbb{R}_0^+)$ there exist $d_1, d_3 \in C(\mathbb{R}, \mathbb{R}_0^+)$ with $d_i(s) = 0 <=> s \le 0$ such that

$$A : s \mapsto \begin{pmatrix} a_3(s) & a_2(s) \\ a_2(s) & a_1(s) \end{pmatrix}$$

has the property

$$\xi^T A(s)\, \xi \ge d_3(s)\, \xi_1^2 + d_1(s)\, \xi_2^2$$

for all $\xi = (\xi_1, \xi_2)^T \in \mathbb{R}^2$.

(A3) There exist $k, l > 0$ such that for all $s \in \mathbb{R}$

$$a_2(s) \le C\, s^k \sqrt{d_1(s)} \text{ and } a_3(s) \le C\, s^k \sqrt{d_3(s)},$$
$$a_2(s) \le C\, s^l \sqrt{d_3(s)} \text{ and } a_1(s) \le C\, s^l \sqrt{d_1(s)}.$$

(A4) The function g lies in $C^{2,1}_{loc}(\mathbb{R})$ where $C^{2,1}_{loc}(\mathbb{R})$ is the space of functions with locally Lipschitz continuous second derivatives.

(A5) There exists a $c_g > 0$ such that $g''(s) \ge c_g$ for all $s \in \mathbb{R}$.

(A6) There exists a $C_g > 0$ such that $g''(s) \le C_g(|s|^r + 1)$ for all $s \in \mathbb{R}$ and with $r \in (0, 2)$.

(A7) The functions σ and g are related by $\sigma(s) = g(s) - sg'(s)$.

(A8) The diffusion coefficient $\mathcal{D}$ is positive.

The above assumptions do not allow for an affine linear relation between σ and Γ which is often assumed in applications. In the following $\langle \cdot, \cdot \rangle$ denotes the dual product between a linear functional and a point in the corresponding normed space.

THEOREM 3.1. *Let the initial data fulfill $h_0 \ge 0$, $h_0 \in H^{1,2}(\Omega)$, $\Gamma_0 \in L^2(\Omega)$ and $g'(\Gamma_0) \in L^2(\Omega)$. Assume also that* (A1)–(A8) *hold. Then there exists a weak solution (h, Γ) of problem* (3.1)–(3.5) *such that*

$$\int_0^T \langle h_t, \zeta \rangle - \int_{\Omega_T} a_2(h)\, \sigma(\Gamma)_x\, \zeta_x - \int_{\Omega_T \setminus \Omega_T^0} \mathcal{S} a_3(h)\, h_{xxx}\, \zeta_x = 0, \tag{3.6}$$

$$\int_0^T \langle \Gamma_t, \zeta \rangle - \int_{\Omega_T} \Gamma\, a_1(h)\, \sigma(\Gamma)_x\, \zeta_x - \int_{\Omega_T \setminus \Omega_T^0} \mathcal{S}\Gamma\, a_2(h)\, h_{xxx}\, \zeta_x + \mathcal{D} \int_{\Omega_T} \Gamma_x\, \zeta_x = 0 \tag{3.7}$$

for all $\zeta \in L^3\big(0, T; H^{2,2}(\Omega)\big)$ with $\zeta_x = 0$ on $(0, T) \times \partial\Omega$ and

$$\Omega_T^0 := \{(t, x) \in \Omega_T;\ h(t, x) = 0\}.$$

The solutions h and Γ have the following regularity properties

$$h \in L^2\big(0, T; H^{1,2}(\Omega)\big) \cap H^{1,2}\big(0, T; (H^{1,2}(\Omega))^*\big) \cap C^{\frac{1}{8}, \frac{1}{2}}(\overline{\Omega_T}),$$

$$\Gamma \in L^\infty\big(0, T; L^2(\Omega)\big) \cap L^2\big(0, T; H^{1,2}(\Omega)\big) \cap H^{1, \frac{3}{2}}\big(0, T; (H^{1,3}(\Omega))^*\big) \cap L^6(\Omega_T),$$

and the initial conditions for h and Γ are attained in the sense of traces in the spaces $H^{1,2}(0, T; (H^{1,2}(\Omega))^*)$ and $H^{1,\frac{3}{2}}(0, T; (H^{1,3}(\Omega))^*)$, respectively. Furthermore

$h \in L^2\big(0,T;H^{3,2}(\Omega\backslash\{h<\delta\})\big)$ *for all* $\delta>0$, $h^3 h_{xxx} \in L^2(\Omega\setminus\Omega_T^0)$ *and in addition it holds* $h \geq 0$ *almost everywhere.*

To prove the above theorem we are following a two-step approach: First we *regularize the degeneracy* which is apparent for $h=0$. For this purpose we approximate the equation for h by a family of nondegenerate equations

$$h_t^\delta + \partial_x\left(a_2(h^\delta)\,\sigma(\Gamma^\delta)_x + [a_3(h^\delta)+\delta]\,\mathcal{S}h^\delta_{xxx}\right) = 0,$$

where $\delta>0$. The surfactant concentration Γ^δ is still subject to the nondegenerate equation

$$\Gamma_t^\delta + \partial_x\left(\Gamma^\delta\, a_1(h^\delta)\,\sigma(\Gamma^\delta)_x + \Gamma^\delta\, a_2(h^\delta)\,\mathcal{S}h^\delta_{xxx}\right) = \mathcal{D}\,\Gamma^\delta_{xx}$$

as we assume $\mathcal{D}>0$. On the boundary $(0,T)\times\partial\Omega$ we impose again no-flux conditions that are

$$h_x^\delta = h_{xxx}^\delta = \Gamma_x^\delta = 0 \tag{3.8}$$

and for $t=0$ we require the same initial data as for the degenerate problem

$$\begin{aligned} h^\delta(0,\cdot) &= h_0 \quad \text{in } \Omega,\\ \Gamma^\delta(0,\cdot) &= \Gamma_0 \quad \text{in } \Omega. \end{aligned}$$

In a second step we use a *Galerkin approximation* which transforms the system of partial differential equations into a system of ordinary differential equations. As basis functions for the finite dimensional space E_k we select an L^2-orthonormal basis of eigenfunctions $\phi_0 = \text{const}$, ϕ_1, $\phi_2,\dots$ of $-\partial_{xx}$ with zero Neumann boundary conditions belonging to eigenvalues $0=\lambda_0<\lambda_1\leq\lambda_2\leq\cdots$. The usual approach would be to make an ansatz for h and Γ in $E_n=\text{span}(\phi_0,\dots,\phi_n)$, but in this case it turned out that it is necessary to reformulate the equation for Γ in terms of the chemical potential $g'(\Gamma)$ and to set up a Galerkin ansatz for $g'(\Gamma)$ instead or Γ. With the help of this transformation we are able to establish the necessary a priori estimates to prove global existence of Galerkin solutions.

Since $g'\in C^1(\mathbb{R})$ and $g''>0$ there exists $W:=(g')^{-1}$ with

$$(W'\circ g')(s) = \frac{1}{g''(s)},$$

so that we can easily transform (3.2) into an equation for $v := g'(\Gamma)$ using that $\sigma(\Gamma)_x = -W(v)\,v_x$:

$$\partial_t\, W(v) - \partial_x\left(W^2(v)\,a_1(h)\,v_x\right) + \partial_x\left(W(v)\,a_2(h)\,\mathcal{S}h_{xxx}\right) = \mathcal{D}\,W(v)_{xx}. \tag{3.9}$$

Due to the approximating properties of the eigenfunctions ϕ_k there exist sequences $(\beta_{kn})_{n\in\mathbb{N}}$ and $(\gamma_{kn})_{n\in\mathbb{N}}$ for $h_0\in H^{1,2}(\Omega)$ and $g'(\Gamma_0)\in L^2(\Omega)$ such that

$$\begin{aligned} h_{0n} &= \sum_{k=0}^{n}\beta_{kn}\phi_k \quad \text{with} \quad h_{0n}\to h_0 \quad \text{in} \quad H^{1,2}(\Omega),\\ v_{0n} &= \sum_{k=0}^{n}\gamma_{kn}\phi_k \quad \text{with} \quad v_{0n}\to v_0 := g'(\Gamma_0) \quad \text{in} \quad L^2(\Omega). \end{aligned}$$

Furthermore we make a Galerkin ansatz for $h^\delta(t)$ and $v^\delta(t) = g'(\Gamma^\delta(t))$ of the form

$$h_n^\delta(t) = \sum_{k=0}^{n} b_{kn}^\delta(t)\phi_k \quad \text{for all } t \in]0,T[\quad \text{with } h_n^\delta(0) = h_{0n},$$
$$v_n^\delta(t) = \sum_{k=0}^{n} c_{kn}^\delta(t)\phi_k \quad \text{for all } t \in]0,T[\quad \text{with } v_n^\delta(0) = v_{0n},$$

where according to (3.1) and (3.9) the functions $b_{kn}^\delta(t)$ and $c_{kn}^\delta(t)$ are subject to the following *Galerkin* equations which have to hold for $j = 0, \dots, n$:

(3.10)
$$\frac{d}{dt}\Big(h_n^\delta(t), \phi_j\Big) + \Big(a_2(h_n^\delta(t))\, W(v_n^\delta(t)) v_{n,x}^\delta(t), \phi_j'\Big) - \Big(a_3(h_n^\delta(t))\, \mathcal{S} h_{n,xxx}^\delta(t) + \delta \mathcal{S} h_{n,xxx}^\delta(t), \phi_j'\Big) = 0,$$

(3.11)
$$\frac{d}{dt}\Big(W(v_n^\delta(t)), \phi_j\Big) + \Big(W^2(v_n^\delta(t))\, a_1(h_n^\delta(t))\, v_{n,x}^\delta(t), \phi_j'\Big) - \Big(W(v_n^\delta(t))\, a_2(h_n^\delta(t)) \mathcal{S} h_{n,xxx}^\delta(t) - \mathcal{D}\, W'(v_n^\delta(t)) v_{n,x}^\delta(t), \phi_j'\Big) = 0,$$

where $(\cdot,\cdot)$ denotes the L^2-scalar product.

Proof. The first step to take is to prove the existence of local solutions of the Galerkin equations (3.10)–(3.11).

In the second step we derive a priori estimates which allow us to extend the local solutions established in the first step towards global solutions of the Galerkin equations.

As the Galerkin equations are an approximation to the nondegenerate system of partial differential equations we establish in the third step the convergence of the Galerkin method.

Generalizing ideas of Bernis and Friedman [BF90] we can show in the final fourth step that solutions of the nondegenerate system converge to solutions of the degenerate system. ∎

3.1. Local existence of solutions to the Galerkin system. To make use of standard theory for systems of ordinary equations we have to work out the structure of (3.10)–(3.11). Since we can write

$$\frac{d}{dt}\big(W(v_n^\delta(t)), \phi_j\big) = \Big[B(c_n^\delta(t))\, \frac{d}{dt} c_n^\delta(t)\Big]_j,$$

where $B(c_n^\delta(t)) = \big[B_{jk}(c_n^\delta(t))\big]_{j,k}$ is the matrix

$$B_{jk}(c_n^\delta(t)) := \int_\Omega W'(v_n^\delta(t,x))\phi_k(x)\phi_j(x)\, dx$$

and $c_n^\delta(t)$ is the vector $(c_{nk}^\delta(t))_{0\le k\le n}$, we simply have to make sure that $B(c_n^\delta(t))$ is symmetric and positive definite and therefore invertible. For $\xi = (\xi_0, \dots, \xi_n) \in \mathbb{R}^{n+1}$ we obtain

$$\sum_{j,k} \xi_j B_{jk}(c_n^\delta(t))\xi_k = \int_\Omega W'(v_n^\delta(t,x)) \left(\sum_j \xi_j \phi_j(x)\right)^2 dx.$$

Since $W' > 0$ and the eigenfunctions ϕ_j are linear independent we obtain that $B(c_n^\delta(t))$ is positive definite and we can multiply (3.11) with $B^{-1}(c_n^\delta(t))$. As a result we have to solve a system of first-order ordinary differential equations of the following type: for given initial data β_n, $\gamma_n \in \mathbb{R}^{n+1}$ and for any value of $\delta > 0$ we look for functions b_n^δ, $c_n^\delta : [0,T] \to \mathbb{R}^{n+1}$ which satisfy the equations

$$\frac{d}{dt}\left(b_n^\delta, c_n^\delta\right)(t) = F\bigl(t, (b_n^\delta(t), c_n^\delta(t))\bigr)^T, \tag{3.12}$$

$$\bigl(b_n^\delta, c_n^\delta\bigr)(0) = \bigl(\beta_n, \gamma_n\bigr), \tag{3.13}$$

where

$$F : [0,T] \times \mathbb{R}^{2(n+1)} \to \mathbb{R}^{2(n+1)},$$
$$(t, y_1, y_2) \mapsto \begin{bmatrix} (f_1(y_1,y_2), \Phi') \\ \bigl(f_2(y_1,y_2), B^{-1}(c_n^\delta(t))\,\Phi'\bigr) \end{bmatrix}$$

with $\Phi = (\phi_0, \dots, \phi_n)$, and

$$\begin{aligned} f_2(y_1,y_2) &:= -W^2(y_2\cdot\Phi)\,a_1(y_1\cdot\Phi)\,y_2\cdot\Phi' + \mathcal{S}W(y_2\cdot\Phi)\,a_2(y_1\cdot\Phi)\,y_1\cdot\Phi''' \\ &\quad - \mathcal{D}\,W'(y_2\cdot\Phi)\,y_2\cdot\Phi', \\ f_1(y_1,y_2) &:= -a_2\bigl(y_1\cdot\Phi\bigr)\,W(y_2\cdot\Phi)\,y_2\cdot\Phi' + \mathcal{S}\,a_3\bigl(y_1\cdot\Phi\bigr)\,y_1\cdot\Phi''' + \delta\mathcal{S}\,y_1\cdot\Phi'''. \end{aligned}$$

By assumptions (A1)–(A3) the right-hand side F satisfies a local Lipschitz condition with respect to $\boldsymbol{y}$. Therefore by the Picard–Lindelöf theorem a unique local solution of the initial value problem (3.12),(3.13) exists.

3.2. Global existence of solutions for the Galerkin system. In this section we will use a priori estimates in order to extend the local solution to a global solution. Since ϕ_k'' is a multiple of ϕ_k we can plug in $p_n^\delta(t) = -\mathcal{S}h_{n,xx}^\delta(t)$ as a test function in (3.10) and $v_n^\delta(t)$ as test function in (3.11) and get:

$$\begin{aligned} &\mathcal{S}\Bigl(\frac{d}{dt}h_{n,x}^\delta(t), h_{n,x}^\delta(t)\Bigr) - \Bigl(a_2(h_n^\delta(t))\,W(v_n^\delta(t))v_{n,x}^\delta(t), \mathcal{S}h_{n,xxx}^\delta(t)\Bigr) \\ &\qquad + \Bigl(a_3(h_n^\delta(t))\,\mathcal{S}h_{n,xxx}^\delta(t) + \delta\mathcal{S}h_{n,xxx}^\delta(t), \mathcal{S}h_{n,xxx}^\delta(t)\Bigr) = 0, \end{aligned} \tag{3.14}$$

$$\begin{aligned} &\Bigl(\frac{d}{dt}W(v_n^\delta(t)), v_n^\delta(t)\Bigr) + \Bigl(W^2(v_n^\delta(t))\,a_1(h_n^\delta(t))\,v_{n,x}^\delta(t), v_{n,x}^\delta(t)\Bigr) \\ &\qquad - \Bigl(W(v_n^\delta(t))\,a_2(h_n^\delta(t))\mathcal{S}h_{n,xxx}^\delta(t) - \mathcal{D}\,W'(v_n^\delta(t))v_{n,x}^\delta(t), v_{n,x}^\delta(t)\Bigr) = 0. \end{aligned} \tag{3.15}$$

Defining

$$\Gamma_n^\delta := (g')^{-1}(v_n^\delta)$$

we can recalculate

$$\begin{aligned} W'(v_n^\delta)\,v_{n,t}^\delta\,v_n^\delta &= \Gamma_{n,t}^\delta\,g'(\Gamma_n^\delta) = \partial_t\,g(\Gamma_n^\delta), \\ W(v_n^\delta)\,v_{n,x}^\delta &= -\sigma(\Gamma_n^\delta)_x \quad \text{and} \quad W'(v_n^\delta)\,|v_{n,x}^\delta|^2 = g''(\Gamma_n^\delta)\,|\Gamma_{n,x}^\delta|^2. \end{aligned}$$

Using these formulas we receive by adding (3.14) and (3.15):

$$\begin{aligned}
(3.16)\quad & \frac{d}{dt}\left\{\frac{\mathcal{S}}{2}\left\|h^\delta_{n,x}(t)\right\|^2_{L^2(\Omega)} + \left\|g(\Gamma^\delta_n(t))\right\|_{L^1(\Omega)}\right\} \\
& + \int_\Omega \Big\{ a_3(h^\delta_n(t))\left|\mathcal{S}h^\delta_{n,xxx}(t)\right|^2 + 2a_2(h^\delta_n(t))\,\mathcal{S}h^\delta_{n,xxx}(t)\,\sigma(\Gamma^\delta_n(t))_x \\
& + a_1(h^\delta_n(t))\left|\sigma(\Gamma^\delta_n(t))_x\right|^2 \Big\} + \delta\left\|\mathcal{S}h^\delta_{n,xxx}(t)\right\|^2_{L^2(\Omega)} + \mathcal{D}\left\|\sqrt{g''(\Gamma^\delta_n(t))}\,\Gamma^\delta_{n,x}(t)\right\|^2_{L^2(\Omega)} = 0.
\end{aligned}$$

Assumption (A2) guarantees that the integral in the second and third line of (3.16) is nonnegative, so that by integration of (3.16) over $[0,T]$ we get for any given $0 < T < \infty$:

$$\frac{\mathcal{S}}{2}\sum_{k=1}^n \left(b^\delta_{kn}(T)\right)^2\lambda_k + \int_\Omega g(W(c^\delta_n(T)\cdot\Phi)) \le C(\beta_n,\gamma_n). \tag{3.17}$$

Taking $\phi_0 = 1$ in (3.10) gives in addition that b^δ_{0n} is bounded and we can deduce that the b^δ_n are a priori bounded. The assumption (A6) implies that $g\circ W(s)\to\infty$ if $s\to\infty$ and hence we obtain that c^δ_n is bounded and therefore the solution (b^δ_n, c^δ_n) can be extended globally. As a conclusion we have shown that the Galerkin equations have global solutions

$$h^\delta_n,\ \Gamma^\delta_n \in C^1\big(0,T;C^\infty(\Omega)\big).$$

3.3. Convergence of the Galerkin method. Let $\zeta \in L^3\big(0,T;H^{2,2}(\Omega)\big)$ be arbitrarily chosen with $\zeta_x = 0$ on $(0,T)\times\partial\Omega$. Then there exist functions $\zeta_n(t,\cdot) = P_n\zeta(t,\cdot) \in \operatorname{span}\{\phi_0,\dots,\phi_n\}$ such that for $n\to\infty$: $\zeta_n(t,\cdot)\to\zeta(t,\cdot)$ in $H^{1,3}(\Omega)$ for almost all $t\in[0,T]$. Using the convergence theorem of Lebesgue we finally get

$$\zeta_n \to \zeta \ \text{ in } \ L^3\big(0,T;H^{1,3}(\Omega)\big) \ \text{ for } \ n\to\infty.$$

Plugging ζ_n into the Galerkin equations (3.10)–(3.11) we have to show that in the following weak formulation we can pass to the limit for $n\to\infty$:

$$\begin{aligned}
(3.18)\quad & \int_0^T \langle h^\delta_{n,t}(t),\zeta_n(t)\rangle - \int_{\Omega_T} a_2(h^\delta_n)\,\sigma(\Gamma^\delta_n)_x\,\zeta_{n,x} \\
& - \int_{\Omega_T} a_3(h^\delta_n)\,\mathcal{S}h^\delta_{n,xxx}\zeta_{n,x} - \delta\int_{\Omega_T}\mathcal{S}h^\delta_{n,xxx}\zeta_{n,x} = 0, \\
(3.19)\quad & \int_0^T \langle \Gamma^\delta_{n,t}(t),\zeta_n(t)\rangle - \int_{\Omega_T} \Gamma^\delta_n\,a_1(h^\delta_n)\,\sigma(\Gamma^\delta_n)_x\,\zeta_{n,x} \\
& - \int_{\Omega_T} \Gamma^\delta_n\,a_2(h^\delta_n)\,\mathcal{S}h^\delta_{n,xxx}\zeta_{n,x} + \mathcal{D}\int_{\Omega_T}\Gamma^\delta_{n,x}\zeta_{n,x} = 0.
\end{aligned}$$

To ensure convergence we have to establish appropriate convergence properties for the integrands involved. Exploiting (3.16) together with assumption (A2) and the convergence properties of the initial data (L^2-convergence of $\big(h^\delta_{n,x}(0)\big)_{n\in\mathbb{N}} \to h_{0,x}$

and L^1-convergence of $\big(g(\Gamma_n^\delta(0))\big)_{n\in\mathbb{N}} \to g(\Gamma_0))$ we can deduce the following:

$$\left(h_{n,x}^\delta\right)_{n\in\mathbb{N}} \quad \text{is uniformly bounded in } L^\infty(0,T;L^2(\Omega)), \tag{3.20}$$

$$\left(g(\Gamma_n^\delta)\right)_{n\in\mathbb{N}} \quad \text{is uniformly bounded in } L^\infty(0,T;L^1(\Omega)), \tag{3.21}$$

$$\left((g''(\Gamma_n^\delta))^{\frac{1}{2}}\,\Gamma_{n,x}^\delta\right)_{n\in\mathbb{N}} \quad \text{is uniformly bounded in } L^2(0,T;L^2(\Omega)), \tag{3.22}$$

$$\left((d_3(h_n^\delta))^{\frac{1}{2}} h_{n,xxx}^\delta\right)_{n\in\mathbb{N}} \quad \text{is uniformly bounded in } L^2(0,T;L^2(\Omega)), \tag{3.23}$$

$$\left((d_1(h_n^\delta))^{\frac{1}{2}}\,\sigma(\Gamma_n^\delta)_x\right)_{n\in\mathbb{N}} \quad \text{is uniformly bounded in } L^2(0,T;L^2(\Omega)), \tag{3.24}$$

$$\sqrt{\delta}\left(h_{n,xxx}^\delta\right)_{n\in\mathbb{N}} \quad \text{is uniformly bounded in } L^2(0,T;L^2(\Omega)). \tag{3.25}$$

Choosing $\zeta_n \equiv 1$ in (3.18) gives that $\frac{d}{dt}\int_\Omega h_n^\delta = 0$ and hence the mean value of h_n is controlled. This fact, (3.20) and the Sobolev embedding theorem lead to

$$\exists\, C>0 \quad \forall\, n\in\mathbb{N} \quad \forall\, \delta>0 \quad \operatorname*{ess\,sup}_{t\in[0,T]} \|h_n^\delta(t)\|_{L^\infty(\Omega)} \le C.$$

Using (3.21)–(3.22) together with (A4)–(A5) we are able to establish uniform bounds for Γ_n^δ with respect to n and δ as follows:

$$\begin{aligned}
&(\Gamma_n^\delta)_{n\in\mathbb{N}} \quad \text{is uniformly bounded in } L^\infty\big(0,T;L^2(\Omega)\big),\\
&(\Gamma_n^\delta)_{n\in\mathbb{N}} \quad \text{is uniformly bounded in } L^2\big(0,T;H^{1,2}(\Omega)\big).
\end{aligned}$$

Applying an embedding theorem for parabolic function spaces (see, e.g., [DiB93]) these results can be combined to:

$$(\Gamma_n^\delta)_{n\in\mathbb{N}} \quad \text{is uniformly bounded in } L^4(0,T;L^\infty(\Omega))\cap L^6(\Omega_T).$$

Finally using (3.23)–(3.24), assumption (A3) and the previous statements allow us to prove that

$$\begin{aligned}
&\mathcal{I}_n^\delta := a_2(h_n^\delta)\,\sigma(\Gamma_n^\delta)_x + a_3(h_n^\delta)\,\mathcal{S}h_{n,xxx}^\delta + \delta\mathcal{S}h_{n,xxx}^\delta \qquad \text{is bounded uniformly in } L^2(\Omega_T),\\
&\mathcal{J}_n^\delta := \Gamma_n^\delta\, a_1(h_n^\delta)\,\sigma(\Gamma_n^\delta)_x + \Gamma_n^\delta\, a_2(h_n^\delta)\,\mathcal{S}h_{n,xxx}^\delta - \mathcal{D}\,\Gamma_{n,x}^\delta \text{ is bounded uniformly in } L^{\frac{3}{2}}(\Omega_T),
\end{aligned}$$

and therefore we obtain

$$\begin{aligned}
&(h_{n,t}^\delta)_{n\in\mathbb{N}} \text{ is uniformly bounded in } L^2(0,T;(H^{1,2}(\Omega))^*),\\
&(\Gamma_{n,t}^\delta)_{n\in\mathbb{N}} \text{ is uniformly bounded in } L^{\frac{3}{2}}(0,T;(H^{1,3}(\Omega))^*).
\end{aligned}$$

To demonstrate the convergence of (3.18)–(3.19) we list the following. Since $(h_n^\delta)_{n\in\mathbb{N}}$ is uniformly bounded in $C^{\frac{1}{8},\frac{1}{2}}(\overline{\Omega_T})$ (Hölder-continuity of $h_n^\delta(t,x)$ with respect to $t\in[0,T]$ and $x\in\Omega$, see [BF90]) we conclude that a subsequence of $(h_n^\delta)_{n\in\mathbb{N}}$ converges uniformly to h^δ for $n\to\infty$. This implies together with the reflexivity of $L^2(0,T;(H^{1,2}(\Omega))^*)$ that $\partial_t h_n^\delta \overset{*}{\rightharpoonup} \partial_t\, h^\delta$ in $L^2\big(0,T;(H^{1,2}(\Omega))^*\big)$. By Poincaré's lemma we can prove that the boundedness of $(h_{n,xxx}^\delta)_{n\in\mathbb{N}}$ in $L^2(\Omega_T)$ implies for all δ the convergence $h_n^\delta \rightharpoonup h^\delta$ in $L^2(0,T;H^{3,2}(\Omega))$. We remark that we sometimes choose subsequences but keep the original index of the sequence.

The convergence results for $(\Gamma_n^\delta)_{n\in\mathbb{N}}$ are not as good as for $(h_n^\delta)_{n\in\mathbb{N}}$. Using the compactness lemma of Aubin–Lions (see [LIO69]) we get strong convergence of $\Gamma_n^\delta \to \Gamma^\delta$ in $L^2(0,T;L^q(\Omega))$ for all $q \in [1,\infty)$. Besides we have the weak convergence of $(\partial_t \Gamma_n^\delta)_{n\in\mathbb{N}}$ to $\partial_t \Gamma^\delta$ in $L^{\frac{3}{2}}(0,T;(H^{1,3}(\Omega))^*)$ and the strong convergence of $\Gamma_n^\delta \to \Gamma^\delta$ in $L^q(\Omega_T)$ for all $q \in [2,6)$ using the boundedness of $(\Gamma_n^\delta)_{n\in\mathbb{N}}$ in $L^6(\Omega_T)$ together with an interpolation estimate for L^p norms (see [E98]).

In order to get a convergence result for $g'(\Gamma_n^\delta)$ we make use of (A4) and (A6): By Lebesgue's theorem and the strong convergence of $(\Gamma_n^\delta)_{n\in\mathbb{N}}$ in $L^q(\Omega_T)$ we first conclude the convergence of $(g''(\Gamma_n^\delta))_{n\in\mathbb{N}}$ to $g''(\Gamma^\delta)$ in $L^3(\Omega_T)$. By an interpolation estimate for L^p-norms this can be extended to $g''(\Gamma_n^\delta) \to g''(\Gamma^\delta)$ in $L^q(\Omega_T)$ for all $q \in [3, \frac{6}{r})$, $r \in (0,2)$. Combining this with the weak convergence $\Gamma_{n,x}^\delta \rightharpoonup \Gamma_x^\delta$ in $L^2(\Omega_T)$ we therefore get $(g''(\Gamma_n^\delta))^{\frac{1}{2}}\Gamma_{n,x}^\delta \rightharpoonup (g''(\Gamma^\delta))^{\frac{1}{2}}\Gamma_x^\delta$ in $L^2(\Omega_T)$ for $n \to \infty$. Since $\sigma(\Gamma_n^\delta)_x$ decomposes into $\sigma(\Gamma_n^\delta)_x = -\Gamma_n^\delta (g''(\Gamma_n^\delta))^{\frac{1}{2}} \cdot (g''(\Gamma_n^\delta))^{\frac{1}{2}}\Gamma_{n,x}^\delta$ we can prove that $\sigma(\Gamma_n^\delta)_x \rightharpoonup \sigma(\Gamma^\delta)_x$ in $L^s(\Omega_T)$ for $s \in [\frac{2}{3}, \frac{6}{4+r})$.

Applying these convergence results to (3.18)–(3.19) we get that the Galerkin solutions $(h_n^\delta, \Gamma_n^\delta)$ converge for any fixed $\delta > 0$ to a weak solution $(h^\delta, \Gamma^\delta)$ of the nondegenerate problem

(3.26)
$$\int_0^T \langle h_t^\delta, \phi\rangle - \int_{\Omega_T} a_2(h^\delta)\,\sigma(\Gamma^\delta)_x \phi_x - \int_{\Omega_T} a_3(h^\delta)\,\mathcal{S}h_{xxx}^\delta \phi_x - \delta \int_{\Omega_T} \mathcal{S}h_{xxx}^\delta \phi_x = 0,$$

(3.27)
$$\int_0^T \langle \Gamma_t^\delta, \phi\rangle - \int_{\Omega_T} \Gamma^\delta a_1(h^\delta)\,\sigma(\Gamma^\delta)_x \phi_x - \mathcal{S}\int_{\Omega_T} \Gamma^\delta a_2(h^\delta)\, h_{xxx}^\delta \phi_x + \mathcal{D}\int_{\Omega_T} \Gamma_x^\delta \phi_x = 0.$$

3.4. Existence of weak solutions of the degenerate problem. When we take the limit $\delta \to 0$ we lose control over h_{xxx}^δ in $L^2(\Omega_t)$. Therefore, similar to as in [BF90], we introduce the sets $\Omega_T \backslash \Omega_T^0$ with

$$\Omega_T^0 := \{(t,x) \in \Omega_T;\, h(t,x) = 0\}$$

on which convergence of the terms involving third derivatives of h^δ holds. This is due to the fact that $h_{xxx}^\delta \rightharpoonup h_{xxx}$ in $L^2(\Omega_T \backslash \Omega_T^\eta)$ with $\Omega_T^\eta := \{(t,x) \in \Omega_T;\, |h(t,x)| \le \eta\}$ and since

$$\begin{aligned}(3.28)\qquad \int_{\Omega_T^\eta} |a_3(h^\delta)\mathcal{S}h_{xxx}^\delta \phi_x| &\le C\|h^\delta\|_{L^\infty(\Omega_T^\eta)}^k \,\|d_3(h^\delta)^{\frac{1}{2}}\mathcal{S}h_{xxx}^\delta\|_{L^2(\Omega_T)}\,\|\phi_x\|_{L^2(\Omega_T)} \\ &\le C\eta^k,\end{aligned}$$

$$\begin{aligned}(3.29)\qquad \int_{\Omega_T^\eta} |\Gamma^\delta a_2(h^\delta)\mathcal{S}h_{xxx}^\delta \phi_x| &\le C\|h^\delta\|_{L^\infty(\Omega_T^\eta)}^l \|\Gamma^\delta d_3(h^\delta)^{\frac{1}{2}}\mathcal{S}h_{xxx}^\delta\|_{L^{\frac{3}{2}}(\Omega_T)}\,\|\phi_x\|_{L^3(\Omega_T)} \\ &\le C\eta^l,\end{aligned}$$

where we made use of (A3) and the bounds (3.20)–(3.23), which are uniform with respect to δ. In particular using the bound (3.23) we obtain the integrability of $h^3 h_{xxx}\chi_{\Omega\backslash\Omega_T^0}$.

Those terms in (3.26)–(3.27) in which h_{xxx}^δ does not occur are not affected and we can pass to the limit for $\delta \to 0$ in the same way as in section 3.3. Since $\delta\|h_{xxx}^\delta\|_{L^2(\Omega_T)}^2$ is uniformly bounded (see (3.25)) we conclude furthermore

$$\delta \int_{\Omega_T} |h_{xxx}^\delta\, \phi_x| \le \delta \|h_{xxx}^\delta\|_{L^2(\Omega_T)} \cdot \|\phi_x\|_{L^2(\Omega_T)} \to 0 \quad \text{for} \quad \delta \to 0.$$

In the remaining terms involving h^δ_{xxx} we can pass to the limit as in [BF90] using the estimates (3.28) and (3.29). Taking $h_- := \min\{h, 0\}$ as a test function in (3.6) gives $h \geq 0$ almost everywhere (see [Yin92] or the discussion in [BGN03]).

Remark. In the proof of Theorem 3.1 it was not necessary to require $\Gamma_0 \geq 0$. But physically relevant initial data for the density Γ_0 are nonnegative. In this case one would expect the solutions Γ to be nonnegative as well. Unfortunately the available methods are not sufficient to show nonnegativity of Γ under the above assumptions. Testing (3.7) by $\min(\Gamma, 0)$ is not possible since the coefficients are not smooth enough. For example, the term $a_2(h)h_{xxx}\chi_{\Omega\setminus\Omega_T^0}$ is only in $L^2(\Omega_T)$ and more regularity is needed to proceed (see, e.g., Ladyzenskaya et al. [LSU68, Chapter III, section 7, assumption (7.1)]).

4. Nonnegativity. Numerical simulations (see [BGN03]) indicate that surfactants can lead to a dramatic thinning of films. In this section we will show that solutions to (1.1)–(1.2) which have positive initial data with respect to h will remain positive besides a set of zero Lebesgue measure. This shows that surfactants do not lead to a rupture of the film on large sets. For in h positive initial data we will generate strictly positive solutions of an approximation problem. In analogy to the single thin film equation we are looking for a functional $G(h)$ such that we can derive further estimates from the identity

$$\frac{d}{dt}\int_\Omega G(h) = \int_\Omega G'(h)h_t = \frac{1}{2}\int_\Omega G''(h)h_x\, h^2\sigma(\Gamma)_x + \frac{\mathcal{S}}{3}\int_\Omega G''(h)h_x\, h^3 h_{xxx}. \tag{4.1}$$

Making the ansatz $G''(h) = h^\alpha$ for any $\alpha > 0$ a formal computation leads to

$$\frac{d}{dt}\int_\Omega G(h) = \frac{1}{2}\int_\Omega h^{2+\alpha}\, h_x\, \sigma(\Gamma)_x - \frac{\mathcal{S}\,(3+\alpha)}{3}\int_\Omega h^{2+\alpha}\, h_x^2\, h_{xx} - \frac{\mathcal{S}}{3}\int_\Omega h^{3+\alpha} h_{xx}^2,$$

which suggests to take $\alpha = -2$. For this choice the first term on the right-hand side is bounded which follows from the a priori estimates we derived in section 3 and the second term vanishes due to the Neumann boundary condition for h. After integration over $[0, T]$ we therefore receive formally the entropy equation:

$$\int_\Omega G(h(T,\cdot)) + \frac{\mathcal{S}}{3}\int_{\Omega_T} h\, h_{xx}^2 = \frac{1}{2}\int_{\Omega_T} h_x\, \sigma(\Gamma)_x + \int_\Omega G(h(0,\cdot)).$$

4.1. Approximation by positive solutions. Starting with the boundary problem

$$\begin{aligned} h_t + \partial_x\Big[a_2(h)\,\sigma(\Gamma)_x + a_3(h)\,\mathcal{S}h_{xxx}\Big] &= 0 \quad \text{in } \Omega_T,\\ \Gamma_t + \partial_x\Big[\Gamma\, a_1(h)\,\sigma(\Gamma)_x + \Gamma\, a_2(h)\,\mathcal{S}h_{xxx} - \mathcal{D}\,\Gamma_x\Big] &= 0 \quad \text{in } \Omega_T,\\ h_x = h_{xxx} = \Gamma_x &= 0 \quad \text{on } (0,T)\times\partial\Omega,\\ h(0,\cdot) &= h_0 \quad \text{in } \Omega,\\ \Gamma(0,\cdot) &= \Gamma_0 \quad \text{in } \Omega, \end{aligned}$$

we will follow an idea by [BF90] and regularize the coefficient functions $a_i(s)$ and lift the initial data h_0 such that the entropy equation can be derived rigorously and the existence proof for weak solutions can be imitated. Both requirements are fulfilled by choosing the regularization

$$a_i^\epsilon(s) = \frac{s^{n+i}}{s^n + \epsilon s^3} \quad \text{for} \quad i \in \{1, 2, 3\}. \tag{4.2}$$

We see that for any fixed $\epsilon > 0$, $i \in \{1,2,3\}$ and $n > 3$:

$$\frac{a_i^\epsilon(s)}{s^{n+i-3}} = \mathcal{O}(1) \quad \text{for} \quad s \to 0 \quad \text{and} \quad \frac{s^n}{s^n + \epsilon s^3} = \mathcal{O}(1) \quad \text{for} \quad s \to \infty.$$

As a consequence the matrix $A^\epsilon(s) := \frac{s^n}{s^n+\epsilon\, s^3}\, A(s)$ of the regularized coefficients fulfills the requirements of Theorem 3.1 as long as $A(s) = \begin{pmatrix} a_3(s) & a_2(s) \\ a_2(s) & a_1(s) \end{pmatrix}$ does. Furthermore we lift the initial data for h with ϵ^θ, $0 < \theta < \frac{1}{n-3}$, so that they become strictly positive and formulate the approximation problem P^ϵ as follows:

$$h_t^\epsilon + \partial_x \left[a_2^\epsilon(h^\epsilon)\, \sigma(\Gamma^\epsilon)_x + a_3^\epsilon(h^\epsilon)\, \mathcal{S} h_{xxx}^\epsilon \right] = 0 \qquad \text{in } \Omega_T, \tag{4.3}$$

$$\Gamma_t^\epsilon + \partial_x \left[\Gamma^\epsilon\, a_1^\epsilon(h^\epsilon)\, \sigma(\Gamma^\epsilon)_x + \Gamma^\epsilon\, a_2^\epsilon(h^\epsilon)\, \mathcal{S} h_{xxx}^\epsilon \right] = \mathcal{D}\, \Gamma_{xx}^\epsilon \qquad \text{in } \Omega_T, \tag{4.4}$$

$$h_x^\epsilon = h_{xxx}^\epsilon = \Gamma_x^\epsilon = 0 \qquad \text{on } (0,T) \times \partial\Omega, \tag{4.5}$$

$$h^\epsilon(0,\cdot) = h_0^\epsilon := h_0 + \epsilon^\theta \quad \text{in } \Omega, \tag{4.6}$$

$$\Gamma^\epsilon(0,\cdot) = \Gamma_0 \qquad \text{in } \Omega. \tag{4.7}$$

For this system we can state the following theorem.

THEOREM 4.1 (existence of positive approximative solutions). *Let the assumptions of Theorem* 3.1 *hold and assume in addition that* a_1^ϵ, a_2^ϵ, a_3^ϵ *are given by* (4.2) *with* $n \geq 5$. *Then there exist for all* $\epsilon > 0$ *functions* $(h^\epsilon, \Gamma^\epsilon)$ *with* $h^\epsilon > 0$ *in* Ω_T *and*

$$h^\epsilon \in H^{1,2}\big(0,T;(H^{1,2}(\Omega))^*\big) \cap L^2\big(0,T;H^{3,2}(\Omega)\big) \cap C^{\frac{1}{8},\frac{1}{2}}(\overline{\Omega_T}),$$

$$\Gamma^\epsilon \in H^{1,\frac{3}{2}}\big(0,T;(H^{1,3}(\Omega))^*\big) \cap L^\infty\big(0,T;L^2(\Omega)\big) \cap L^2\big(0,T;H^{1,2}(\Omega)\big),$$

which attain the initial data and are such that for all $\zeta \in L^3(0,T;H^{1,3})$ *equations* (4.3)–(4.5) *are fulfilled in the following weak sense:*

$$\int_0^T \langle h_t^\epsilon(t), \zeta(t) \rangle\, dt - \int_{\Omega_T} a_2^\epsilon(h^\epsilon)\, \sigma(\Gamma^\epsilon)_x\, \zeta_x - \mathcal{S} \int_{\Omega_T} a_3^\epsilon(h^\epsilon)\, h_{xxx}^\epsilon \zeta_x = 0, \tag{4.8}$$

$$\int_0^T \langle \Gamma_t^\epsilon(t), \zeta(t) \rangle\, dt - \int_{\Omega_T} \Gamma^\epsilon a_1^\epsilon(h^\epsilon)\, \sigma(\Gamma^\epsilon)_x\, \zeta_x - \mathcal{S} \int_{\Omega_T} \Gamma^\epsilon a_2^\epsilon(h^\epsilon)\, h_{xxx}^\epsilon \zeta_x = -\mathcal{D} \int_{\Omega_T} \Gamma_x^\epsilon \zeta_x. \tag{4.9}$$

Proof. For the existence part we can imitate the proof of existence of solutions to (3.1)–(3.2). We receive functions $(h^\epsilon, \Gamma^\epsilon)$ such that for all $\zeta \in L^3(0,T;H^{1,3})$:

$$\int_0^T \langle h_t^\epsilon(t), \zeta(t) \rangle\, dt - \int_{\Omega_T} a_2^\epsilon(h^\epsilon)^*\, \sigma(\Gamma^\epsilon)_x\, \zeta_x - \int_{\Omega_T \setminus \Omega_T^{\epsilon,0}} a_3^\epsilon(h^\epsilon)^*\, \mathcal{S} h_{xxx}^\epsilon \zeta_x = 0,$$

$$\int_0^T \langle \Gamma_t^\epsilon(t), \zeta(t) \rangle\, dt - \int_{\Omega_T} \Gamma^\epsilon a_1^\epsilon(h^\epsilon)^*\, \sigma(\Gamma^\epsilon)_x\, \zeta_x - \int_{\Omega_T \setminus \Omega_T^{\epsilon,0}} \Gamma^\epsilon a_2^\epsilon(h^\epsilon)^*\, \mathcal{S} h_{xxx}^\epsilon \zeta_x = -\mathcal{D} \int_{\Omega_T} \Gamma_x^\epsilon \zeta_x$$

with $\Omega_T^{\epsilon,0} := \{(t,x) \in \Omega_T;\, h^\epsilon(t,x) = 0\}$ and $a_i^\epsilon(s)^* := a_i^\epsilon(s)\chi_{\{s>0\}}$. We already know that $h^\epsilon \geq 0$ for all $t \in [0,T]$ and almost all $x \in \Omega$. We now want to show that the set $\Omega_T^{\epsilon,0}$ is empty. Since $(h^\epsilon)_{\epsilon>0}$ is uniformly bounded in $C^0(\overline{\Omega_T})$ there exists an $A > 0$ such that $\max h^\epsilon \leq A$ for all ϵ and we define

$$g_\epsilon(s) = -\int_s^A \frac{r}{a_3^\epsilon(r)}\, dr \leq 0 \quad \text{and} \quad G_\epsilon(s) = -\int_s^A g_\epsilon(r)\, dr \geq 0. \tag{4.10}$$

Then $G'_\epsilon(s) = g_\epsilon(s)$ and

$$g'_\epsilon(s) = G''_\epsilon(s) = \frac{s}{a_3^\epsilon(s)} = \frac{s^{n+1} + \epsilon s^4}{s^{n+3}}.$$

More precisely we obtain

$$g_\epsilon(s) := -\left(\frac{1}{s} + \frac{\epsilon s^{2-n}}{n-2}\right) + c_\epsilon$$

and

$$G_\epsilon(s) := \log\frac{1}{s} + \frac{\epsilon s^{3-n}}{(n-3)(n-2)} + c_\epsilon s + d_\epsilon$$

with constants c_ϵ and d_ϵ that depend on A. We remark that c_ϵ and d_ϵ are uniformly in ϵ bounded.

Since h_0^ϵ is strictly positive and since h^ϵ is continuous we conclude that there exists a time t^* such that h^ϵ is strictly positive on $[0, t^*]$. On this time interval the system (4.8)–(4.9) is strictly parabolic. Therefore parabolic regularity implies that h is smooth on this time interval. Hence the following computations are justified. Choosing $g_\epsilon(h^\epsilon)$ as test function for (4.3) leads to

$$\int_\Omega G_\epsilon(h^\epsilon(t^*,\cdot)) - \mathcal{S}\int_{\Omega_{t^*}} h^\epsilon h_x^\epsilon h_{xxx}^\epsilon = \int_\Omega G_\epsilon(h^\epsilon(0,\cdot)) + \int_{\Omega_{t^*}} \frac{h^\epsilon\, a_2^\epsilon(h^\epsilon)}{a_3^\epsilon(h^\epsilon)} h_x^\epsilon\, \sigma(\Gamma^\epsilon)_x.$$

Using the boundary condition (4.5) the second term reduces to

$$-\int_{\Omega_{t^*}} h^\epsilon h_x^\epsilon h_{xxx}^\epsilon = \frac{1}{3}\underbrace{\int_{\Omega_{t^*}} \partial_x\,(h_x^\epsilon)^3}_{=0} + \int_{\Omega_{t^*}} h^\epsilon (h_{xx}^\epsilon)^2$$

and we get

$$\int_\Omega G_\epsilon(h^\epsilon(t^*,\cdot)) + \int_{\Omega_{t^*}} h^\epsilon (h_{xx}^\epsilon)^2 = \int_\Omega G_\epsilon(h_0^\epsilon) + \int_{\Omega_{t^*}} h_x^\epsilon\, \sigma(\Gamma^\epsilon)_x. \tag{4.11}$$

Since $h_x^\epsilon\, \sigma(\Gamma^\epsilon)_x$ is uniformly in ϵ bounded in $L^1(\Omega_T)$ we have to check the first term on the right-hand side. By definition we know that

$$\int_\Omega G_\epsilon(h_0^\epsilon) = \int_\Omega \frac{\epsilon (h_0^\epsilon)^{3-n}}{(n-3)(n-2)} + \int_\Omega G_0(h_0^\epsilon).$$

Since h_0^ϵ is bounded away from zero we conclude from the above that for all $t \in [0, t^*]$

$$\int_\Omega G_\epsilon(h^\epsilon(t,\cdot)) \le C(\epsilon). \tag{4.12}$$

Let's assume $t_* < T$ and let $t_0 \in (t_*, T]$ be the first time such that $h^\epsilon(t_0, x_0) = 0$ for $x_0 \in \Omega$. Let $(t_n)_{n\in\mathbb{N}}$ be a sequence $t_n \nearrow t_0$ for $n \to \infty$. We then conclude

$$h^\epsilon(t_n,\cdot) \to h^\epsilon(t_0,\cdot) \quad \text{uniformly in } \Omega \text{ for} \quad n \to \infty. \tag{4.13}$$

To prove that the entropy is still bounded in t_0 we apply Fatou's lemma. Since $G_\epsilon(h^\epsilon(t_n,\cdot))$ is bounded from below uniformly with respect to n we get

$$\int_\Omega \liminf_{n\to\infty} G_\epsilon(h^\epsilon(t_n,\cdot)) \le \liminf_{n\to\infty}\int_\Omega G_\epsilon(h^\epsilon(t_n,\cdot)) \le C. \tag{4.14}$$

This leads to the following contradiction: Using the Hölder-regularity of h^ϵ we conclude that for all $x\in\Omega$,

$$h^\epsilon(t_0,x) \le C|x-x_0|^{\frac{1}{2}}$$

which gives for $n>3$

$$\int_\Omega [h^\epsilon(t_0,x)]^{3-n} \ge C\int_\Omega |x-x_0|^{\frac{3-n}{2}}. \tag{4.15}$$

The right-hand side in (4.15) is unbounded since $\frac{n-3}{2}\ge 1$ for $n\ge 5$ and this contradicts the facts (4.13) and (4.14). Hence we can conclude that h^ϵ are strictly positive for all times $t\in[0,T]$. ◻

4.2. Convergence of the regularized problem. In this subsection we show that initial data with a positive height possess solutions to (1.1)–(1.2) which do not form dead cores, i.e., regimes with zero height cannot have positive measure.

THEOREM 4.2. *Let the assumptions of Theorem* 3.1 *hold with* $a_i(s)=\frac{1}{i}s^i$ *and suppose*

$$h_0\ge 0 \quad \textit{and} \quad \int_\Omega |\log h_0| < \infty. \tag{4.16}$$

Then there exists a solution of (1.1)–(1.2) *which fulfills all properties required in Theorem* 3.1 *and in addition:*

(i) $h\ge 0$ *a.e. in* Ω_T *and* $\mathcal{L}^1\big(\{x\in\Omega \,\big|\, h(t,x)=0\}\big)=0$ *for all* $t\ge 0$,

(ii) *there exists a constant* $0<C<\infty$ *such that for all* $t\in[0,T]$

$$\int_\Omega |\log h(t,\cdot)| \le C.$$

Proof. We again will make use of the energy estimates and therefore use $-\mathcal{S}\,h^\epsilon_{xx}$ as test function for (4.3) and $g'(\Gamma^\epsilon)$ as test function for (4.4). From

$$\begin{aligned}&\frac{\mathcal{S}}{2}\int_\Omega (h^\epsilon_x(T,\cdot))^2 + \int_\Omega g(\Gamma^\epsilon(T,\cdot)) + \mathcal{D}\int_{\Omega_T} g''(\Gamma^\epsilon)(\Gamma^\epsilon_x)^2\\ &+\int_{\Omega_T} d^\epsilon_3(h^\epsilon)(\mathcal{S}h^\epsilon_{xxx})^2 + \int_{\Omega_T} d^\epsilon_1(h^\epsilon)(\sigma(\Gamma^\epsilon)_x)^2 \le \frac{\mathcal{S}}{2}\int_\Omega (h^\epsilon_x(0,\cdot))^2 + \int_\Omega g(\Gamma^\epsilon(0,\cdot)),\end{aligned}$$

we then deduce in the same way as in the existence proof of weak solutions for (3.1)–(3.2) the necessary convergence results for $\epsilon\to 0$ which enable us to show that weak solutions of (4.3)–(4.7) converge to weak solutions of (3.1)–(3.5). For the evidence of this we take a closer look at one of the terms of interest:

$$\int_{\Omega_T} \Gamma^\epsilon\, a^\epsilon_2(h^\epsilon)\, h^\epsilon_{xxx}\,\zeta_x = \int_{\Omega_T} \Gamma^\epsilon\, \tilde{a}^\epsilon_2(h^\epsilon)\,(h^\epsilon)^{\frac{3}{2}} h^\epsilon_{xxx}\,\zeta_x,$$

where

$$\tilde{a}_2^\epsilon(h^\epsilon) = (h^\epsilon)^{-\frac{3}{2}} a_2^\epsilon(h^\epsilon) = \frac{(h^\epsilon)^n}{(h^\epsilon)^n + \epsilon (h^\epsilon)^3}\,(h^\epsilon)^{\frac{1}{2}} \quad \to h^{\frac{1}{2}} \quad \text{pointwise for} \quad \epsilon \to 0.$$

Since $(h^\epsilon)_{\epsilon>0}$ is uniformly bounded in $C^{\frac{1}{8},\frac{1}{2}}(\overline{\Omega_T})$, the sequence $(\tilde{a}_2^\epsilon(h^\epsilon)\zeta_x)_{\epsilon>0}$ converges strongly to $h^{\frac{1}{2}}\zeta_x$ in $L^3(\Omega_T)$ for $\epsilon \to 0$. Hence we conclude with the uniformly boundedness of $\Gamma^\epsilon\,(h^\epsilon)^{\frac{3}{2}} h^\epsilon_{xxx}$ in $L^{\frac{3}{2}}(\Omega_T)$ and the weak convergence of $(h^\epsilon_{xxx})_\epsilon$ in $L^2(\Omega_T \backslash \Omega_T^0)$ that

$$\lim_{\epsilon\to 0} \int_{\Omega_T} \Gamma^\epsilon\, a_2^\epsilon(h^\epsilon)\, h^\epsilon_{xxx}\, \zeta_x = \int_{\Omega_T\backslash\Omega_T^0} \Gamma\, a_2(h)\, h_{xxx}\, \zeta_x .$$

We now like to establish the nonnegativity result using the entropy inequality as a crucial tool. Since $h^\epsilon \to h$ uniformly and $h^\epsilon > 0$ we obtain $h \geq 0$. Now we assume that there exists a $t_0 \in (0,T)$ such that

$$\mathcal{L}^1(E_{t_0}) > 0 \quad \text{for} \quad E_{t_0} := \{x \in \Omega \big| \, h(t_0,x) = 0\}.$$

Using the uniform convergence of h^ϵ there exists a $w(\epsilon)$ with $h_\epsilon(t_0,x) < w(\epsilon)$ for all $x \in E_{t_0}$ such that for all $x \in E_{t_0}$ and arbitrary $\eta > 0$ with $w(\epsilon) < \eta$,

$$G_\epsilon(h^\epsilon(t_0,x)) \geq -\int_{w(\epsilon)}^A g_\epsilon(s)ds \geq -\int_\eta^A g_\epsilon(s)ds.$$

Since the last integral converges to $-\int_\eta^A g_0(s)$ for $\epsilon \to 0$ and since $-\int_\eta^A g_0(s) \geq c\,\log\frac{1}{\eta}$ we obtain

$$\limsup_{\epsilon\to 0} \int_\Omega G_\epsilon(h^\epsilon(t_0,x)) \geq c \log\frac{1}{\eta}\ \mathcal{L}(E_{t_0}) \to \infty \quad \text{for} \quad \eta \to 0\,.$$

We now plan to derive a contradiction by showing that the left-hand side in the above inequality is bounded. Considering (4.11) and taking into account that $h^\epsilon_x\sigma(\Gamma^\epsilon)$ is uniformly in ϵ bounded in the $L^1(\Omega_T)$ topology we only need to control $\int_\Omega G_\epsilon(h_0^\epsilon)$. To obtain this we first observe that for $n < 3$

$$\epsilon(h_0 + \epsilon^\theta)^{3-n} \leq \epsilon^{1+\theta(3-n)} \to 0 \quad \text{for} \quad \epsilon \to 0\,.$$

This holds because we defined θ such that $0 < \theta < \frac{1}{n-3}$. Then boundedness of $\int_\Omega G_\epsilon(h_0^\epsilon)$ uniformly in ϵ follows from (4.16). As a conclusion the set of points where $h = 0$ is of Lebesgue measure zero which proves the first statement (i). The second statement (ii) we conclude from the pointwise convergence

$$G_\epsilon(h^\epsilon(t,x)) \to G_0(h(t,x)) \quad \text{for} \quad \epsilon \to 0$$

in combination with (4.12) and Fatou's lemma:

$$\int_\Omega \liminf_{\epsilon\to 0} G_\epsilon(h^\epsilon(t,x)) \leq \liminf_{\epsilon\to 0} \int_\Omega G_\epsilon(h^\epsilon(t,x)) \leq C,$$

since $G_0(s) = \log\frac{1}{s}$. $\square$

5. Appendix.

5.1. Transport identity. In this appendix we prove the transport theorem (2.9). The proof is based on the following identity which is the Gauss theorem applied to the vector field $\Gamma\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}(1,\mathbf{v}_{\boldsymbol{\nu}})$:

$$\int_{C'} \nabla_{C'}\cdot\left(\Gamma\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}(1,\mathbf{v}_{\boldsymbol{\nu}})\right)dS^3=\int_{\partial C'}\Gamma\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}(1,\mathbf{v}_{\boldsymbol{\nu}})\cdot\mathbf{n}_{\partial C'}dS^2, \tag{5.1}$$

where $\mathbf{n}_{\partial C'}$ is the outer unit normal to $\partial C'$. First we compute the divergence under the integral on the left-hand side. Taking an orthonormal basis $\{\boldsymbol{t}_1,\boldsymbol{t}_2,\boldsymbol{t}_3\}$ of the tangent space to C' the divergence of a vector field F is defined as

$$\nabla_{C'}\cdot\, F=\sum_{i=1}^{3}(\partial_{\boldsymbol{t}_i}F)\cdot\boldsymbol{t}_i.$$

Choosing $\boldsymbol{t}_1=(0,\boldsymbol{\tau}_1)$, $\boldsymbol{t}_2=(0,\boldsymbol{\tau}_2)$, $\boldsymbol{t}_3=\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}(1,\mathbf{v}_{\boldsymbol{\nu}})$ with $(\boldsymbol{\tau}_1,\boldsymbol{\tau}_2)$ being an orthonormal basis of the tangent space to C'_t and $\mathbf{v}_{\boldsymbol{\nu}}$ being such that $\mathbf{v}_{\boldsymbol{\nu}}\cdot\boldsymbol{\tau}_i=0$ $(i=1,2)$ we obtain

$$\begin{aligned}\nabla_{C'}\cdot\left(\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}(1,\mathbf{v}_{\boldsymbol{\nu}})\right)&=\sum_{i=1}^{2}\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}\partial_{\boldsymbol{t}_i}(1,\mathbf{v}_{\boldsymbol{\nu}})\cdot\boldsymbol{t}_i\\&\quad+\left(\partial_{\boldsymbol{t}_3}\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}\right)(1+\mathbf{v}_{\boldsymbol{\nu}}^2)\\&\quad+\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}\Big(\partial_{\boldsymbol{t}_3}(1,\mathbf{v}_{\boldsymbol{\nu}})\Big)\cdot(1,\mathbf{v}_{\boldsymbol{\nu}}),\end{aligned}$$

where we used that $\boldsymbol{\tau}_1$, $\boldsymbol{\tau}_2$ and $\mathbf{v}_{\boldsymbol{\nu}}$ are orthogonal. A straightforward computation shows that the last two terms cancel and we obtain (using the fact $\mathbf{v}_{\boldsymbol{\nu}}=\alpha\boldsymbol{\nu}$ for the scalar normal velocity α) that

$$\begin{aligned}\nabla_{C'}\cdot\left(\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}^2}}}(1,\mathbf{v}_{\boldsymbol{\nu}})\right)&=\frac{\alpha}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}\sum_{i=1}^{2}(\partial_{\boldsymbol{\tau}_i}\boldsymbol{\nu})\cdot\boldsymbol{\tau}_i\\&=-\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}^2}}}\mathbf{v}_{\boldsymbol{\nu}}\cdot\boldsymbol{\kappa}_{\boldsymbol{\nu}},\end{aligned}$$

where we set $\boldsymbol{\kappa}_{\boldsymbol{\nu}}=\kappa\boldsymbol{\nu}$ with $\kappa=-\sum_{i=1}^{2}(\partial_{\boldsymbol{\tau}_i}\boldsymbol{\nu})\cdot\boldsymbol{\tau}_i$. Altogether we obtain

$$\nabla_{C'}\cdot\left(\Gamma\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}(1,\mathbf{v}_{\boldsymbol{\nu}})\right)=\nabla_{C'}\Gamma\cdot\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}(1,\mathbf{v}_{\boldsymbol{\nu}})-\Gamma\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}\mathbf{v}_{\boldsymbol{\nu}}\cdot\boldsymbol{\kappa}_{\boldsymbol{\nu}}.$$

Computing the surface element with the help of the above basis $\{\boldsymbol{t}_1,\boldsymbol{t}_2,\boldsymbol{t}_3\}$ one obtains that for a function f on C',

$$\int_{C'}f\,dS^3=\int_{t_1}^{t_2}\int_{C'_t}f\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}\,dS^2\,dt.$$

Hence we obtain

$$\int_{C'} \nabla_{C'} \cdot \left(\Gamma \frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}(1,\mathbf{v}_{\boldsymbol{\nu}}) \right) = \int_{t_1}^{t_2} \int_{C'_t} \Big(\partial_{(1,\mathbf{v}_{\boldsymbol{\nu}})}\Gamma - \Gamma\left(\mathbf{v}_{\boldsymbol{\nu}} \cdot \boldsymbol{\kappa}_{\boldsymbol{\nu}}\right)\Big) dS^2\, dt. \tag{5.2}$$

It remains to compute the right-hand side in (5.1). Using the identity

$$C' = \bigcup_{t\in[t_1,t_2]} \{t\} \times C'_t$$

we see that $\partial C'$ contains three parts: top, bottom and lateral boundary. For the top and similarly for the bottom (with a different sign) we obtain $\mathbf{n}_{\partial C'} = \frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}(1,\mathbf{v}_{\boldsymbol{\nu}})$. Hence

$$\int_{\partial C' \cap C'_{t_2}} \Gamma \frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}(1,\mathbf{v}_{\boldsymbol{\nu}}) \cdot \mathbf{n}_{\partial C'}\, dS^2 = \int_{C'_{t_2}} \Gamma\, dS^2 \tag{5.3}$$

and a similar formula holds for t_1 with the different sign on the right-hand side. It remains to compute the integral on the lateral surface. We need to identify $\mathbf{n}_{\partial C'}$ on the lateral boundary. $\mathbf{n}_{\partial C'}$ has to be normal to $\partial C'$ and tangential to C'. Now we choose an orthogonal system $(0,\boldsymbol{\tau}_1)$, $(0,\boldsymbol{\tau}_2)$ and $(1,\mathbf{v}_{\boldsymbol{\nu}})$ such that $\boldsymbol{\tau}_2$ is tangential to $\partial C'_t$. Hence we can choose without loss of generality

$$\boldsymbol{\tau}_1 = \mathbf{n}_{\partial C'_t}.$$

Claim. $(1,\mathbf{v}_{\boldsymbol{\nu}}) + (0, v_{\partial C'_t}\,\mathbf{n}_{\partial C'_t})$ is tangential to $\partial C'$.

Proof. Assume $(t,x) \in \partial C'$. Choose a curve $(s,y(s)) \in \partial C'$ such that $(t,y(t)) = (t,x)$. Hence $(1,y'(t))$ is tangential to $\partial C'$. Since $(1,y'(t))$ lies in the tangent space to C', we have

$$(1,y'(t)) = (1,\mathbf{v}_{\boldsymbol{\nu}}) + \alpha(0,\boldsymbol{\tau}_1) + \beta(0,\boldsymbol{\tau}_2).$$

We defined $v_{\partial C'_t} = y'(t)\cdot \mathbf{n}_{\partial C'_t}$ and hence, since $\boldsymbol{\tau}_1 \cdot \boldsymbol{\tau}_2 = 0$, we obtain $v_{\partial C'_t} = \alpha$. This implies that

$$(1,\mathbf{v}_{\boldsymbol{\nu}}) + (0, v_{\partial C'_t}\,\mathbf{n}_{\partial C'_t})$$

is tangential to $\partial C'$.

Now we need to find numbers a and b such that

$$\mathbf{n}_{\partial C'} = a(0,\boldsymbol{\tau}_1) + b(1,\mathbf{v}_{\boldsymbol{\nu}})$$

is perpendicular to $(1, \mathbf{v}_{\boldsymbol{\nu}} + v_{\partial C'_t}\,\mathbf{n}_{\partial C'_t})$ and normal. A simple computation shows

$$b = -\frac{v_{\partial C'_t}}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2 + v_{\partial C'_t}^2}\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}$$

and hence

$$\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}(1,\mathbf{v}_{\boldsymbol{\nu}}) \cdot \mathbf{n}_{\partial C'} = -\frac{v_{\partial C'_t}}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2 + v_{\partial C'_t}^2}}.$$

As a result

$$\int_{\partial C'\setminus(C'_{t_2}\bigcup C'_{t_1})} \Gamma\frac{1}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2}}(1,\mathbf{v}_{\boldsymbol{\nu}})\cdot\mathbf{n}_{\partial C'}\,dS^2 = -\int_{\partial C'\setminus(C'_{t_2}\bigcup C'_{t_1})} v_{\partial C'_t}\frac{\Gamma}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2+v_{\partial C'_t}^2}}\,dS^2.$$

Using $(0,\boldsymbol{\tau}_1)$ and $\mathbf{n}_{\partial C'}$ to compute the area element we obtain

$$-\int_{\partial C'\setminus(C'_{t_2}\bigcup C'_{t_1})} v_{\partial C'_t}\frac{\Gamma}{\sqrt{1+\mathbf{v}_{\boldsymbol{\nu}}^2+v_{\partial C'_t}^2}}\,dS^2 = -\int_{t_1}^{t_2}\int_{\partial C'_t} v_{\partial C'_t}\,\Gamma dS^1\,dt. \tag{5.4}$$

Combining (5.1), (5.2), (5.3) and (5.4) gives

$$\int_{C'_{t_2}}\Gamma\,dS^2 - \int_{C'_{t_1}}\Gamma\,dS^2 = \int_{t_1}^{t_2}\int_{C'_t}\Big(\partial_{(1,\mathbf{v}_{\boldsymbol{\nu}})}\Gamma - \Gamma\,(\mathbf{v}_{\boldsymbol{\nu}}\cdot\boldsymbol{\kappa}_{\boldsymbol{\nu}})\Big)dS^2\,dt + \int_{t_1}^{t_2}\int_{\partial C'_t}\Gamma\,v_{\partial C'_t}dS^1 dt.$$

Differentiating with respect to t_2 now gives (2.9). □

5.2. Energy identity. In this subsection we show the energy inequality (2.16). Using transport theorems (see, e.g., (2.9)) we obtain

$$\begin{aligned}\frac{d}{dt}E(t) &= \int_{\Omega_t}\rho_0\,\mathbf{u}\cdot\mathbf{u}_t\,dV^3 + \int_{\partial\Omega_t}\frac{\rho_0}{2}\mathbf{u}^2(\mathbf{v}_{\boldsymbol{\nu}}\cdot\boldsymbol{\nu})\,dS^2\\ &\quad+\int_{C_t}\Big(\partial_{(1,\mathbf{v}_{\boldsymbol{\nu}})}\,g(\Gamma) - g(\Gamma)(\mathbf{v}_{\boldsymbol{\nu}}\cdot\boldsymbol{\kappa}_{\boldsymbol{\nu}})\Big)dS^2 + \int_{\partial C_t} g(\Gamma)\,v_{\partial C'_t}dS^1.\end{aligned}$$

The 90° angle condition at the outer boundary implies that the last term vanishes. Using the equations (2.1)–(2.2) for $\mathbf{u}$ and (2.11) for Γ and noticing that $\nabla\cdot T = -\nabla p + \mu\Delta\mathbf{u}$ we obtain

$$\begin{aligned}\frac{d}{dt}E(t) &= \int_{\Omega_t}\Big(-\rho_0\,\mathbf{u}\cdot(\mathbf{u}\cdot\nabla)\mathbf{u} + (\nabla\cdot T)\cdot\mathbf{u}\Big)\,dV^3 + \int_{C_t}\frac{\rho_0}{2}\mathbf{u}^2(\mathbf{v}_{\boldsymbol{\nu}}\cdot\boldsymbol{\nu})dS^2\\ &\quad+\int_{C_t} g'(\Gamma)\Big(\Gamma(\mathbf{v}_{\boldsymbol{\nu}}\cdot\boldsymbol{\kappa}_{\boldsymbol{\nu}}) - \nabla_s\cdot\big(\Gamma\mathbf{u}_{tan} - D\nabla_s\Gamma\big)\Big)dS^2 - \int_{C_t} g(\Gamma)\,(\mathbf{v}_{\boldsymbol{\nu}}\cdot\boldsymbol{\kappa}_{\boldsymbol{\nu}})dS^2.\end{aligned}$$

Taking into account that $\nabla\cdot\mathbf{u}=0$, $\mathbf{u}\cdot(\mathbf{u}\cdot\nabla)\mathbf{u} = \frac{1}{2}\mathbf{u}\cdot\nabla|\mathbf{u}|^2$ and $\mathbf{u} = \mathbf{u}_{tan} + (\mathbf{u}\cdot\boldsymbol{\nu})\boldsymbol{\nu}$ on C_t and using the boundary conditions (2.6) and (2.7) we obtain after integration by parts

$$
\begin{aligned}
\frac{d}{dt}E(t) = &-\int_{\Omega_t} T : \nabla \mathbf{u}\, dV^3 - \int_{C_t} \frac{\rho_0}{2}\mathbf{u}^2\,(\mathbf{v}_{\boldsymbol{\nu}} - \mathbf{u})\cdot\boldsymbol{\nu}\, dS^2 + \int_{C_t} \mathbf{u}_{tan}\cdot\nabla_s\,\sigma(\Gamma) dS^2 \\
&+ \int_{C_t} (\mathbf{u}\cdot\boldsymbol{\nu})\sigma(\Gamma)\kappa\, dS^2 \\
&+ \int_{C_t} g'(\Gamma)\Gamma(\mathbf{v}_{\boldsymbol{\nu}}\cdot\boldsymbol{\kappa}_{\boldsymbol{\nu}})\, dS^2 + g''(\Gamma)\Gamma\mathbf{u}_{tan}\cdot\nabla_s\Gamma\, dS^2 - D\int_{C_t} g''(\Gamma)|\nabla_s\Gamma|^2\, dS^2 \\
&- \int_{C_t} g(\Gamma)(\mathbf{v}_{\boldsymbol{\nu}}\cdot\boldsymbol{\kappa}_{\boldsymbol{\nu}})\, dS^2 \\
= &-\int_{\Omega_t} T : \nabla\mathbf{u}\, dV^3 + \int_{C_t} \mathbf{u}_{tan}\cdot\nabla_s\Gamma\big(\Gamma\, g''(\Gamma) - \sigma'(\Gamma)\big) dS^2 \\
&+ \int_{C_t} (\mathbf{u}\cdot\boldsymbol{\nu})\kappa\left(\sigma(\Gamma) + g'(\Gamma)\Gamma - g(\Gamma)\right) dS^2 - D\int_{C_t} g''(\Gamma)|\nabla_s\Gamma|^2\, dS^2 \\
= &-\int_{\Omega_t} T : \nabla\mathbf{u}\, dV^3 - D\int_{C_t} g''(\Gamma)|\nabla_s\Gamma|^2\, dS^2,
\end{aligned}
$$

where we used (2.14) and the facts $\mathbf{v}_{\boldsymbol{\nu}}\cdot\boldsymbol{\nu} = \mathbf{u}\cdot\boldsymbol{\nu}$ and $\mathbf{v}_{\boldsymbol{\nu}}\cdot\boldsymbol{\kappa}_{\boldsymbol{\nu}} = \kappa\mathbf{u}\cdot\boldsymbol{\nu}$.

Acknowledgments. The authors would like to thank Willi Alt and Felix Otto for helpful discussions.

REFERENCES

[BGN03] J. Barrett, H. Garcke, and R. Nürnberg, *Finite element approximation of surfactant spreading on a thin film*, SIAM J. Numer. Anal., 41 (2003), pp. 1427–1464.

[BBD95] E. Beretta, M. Bertsch, and R. Dal Passo, *Nonnegative solutions of a fourth-order nonlinear degenerate parabolic equation*, Arch. Ration. Mech. Anal., 129 (1995), pp. 175–200.

[BF90] F. Bernis and A. Friedman, *Higher order nonlinear degenerate parabolic equations*, J. Differential Equations, 83 (1990), pp. 179–206.

[B96] F. Bernis, *Finite speed of propagation and continuity of the interface for thin viscous flows*, Adv. Differential Equations, 3 (1996), pp. 337–368.

[BP96] A. L. Bertozzi and M. Pugh, *The lubrication approximation for thin viscous films: Regularity and long-time behavior of weak solutions*, Comm. Pure Appl. Math., 49 (1996), pp. 85–123.

[BMS99] A. L. Bertozzi, A. Münch, and M. Shearer, *Undercompressive shocks in thin film flows*, Phys. D, 134 (1999), pp. 431–464.

[BG88] M. S. Borgas and J. B. Grotberg, *Monolayer flow on a thin film*, J. Fluid Mech., 193 (1988), pp. 151–170.

[DGG98] R. Dal Passo, H. Garcke, and G. Grün, *On a fourth order degenerate parabolic equation: Global entropy estimates and qualitative behavior of solutions*, SIAM J. Math. Anal., 29 (1998), pp. 321–342.

[DG01] R. Dal Passo and H. Garcke, *Solutions of a fourth order degenerate parabolic equation with weak initial trace*, Ann. Sc. Norm. Sup. Pisa Cl. Sci. (5), 28 (1999), pp. 153–181.

[DeG94] A. De Wit and D. Gallez, *Nonlinear evolution equations for thin liquid films with insoluble surfactants*, Phys. Fluids, 6 (1994), pp. 3256–3266.

[DiB93] E. Di Benedetto, *Degenerate Parabolic Equations*, Springer-Verlag, New York, 1993.

[EG96] C. M. Elliott and H. Garcke, *On the Cahn-Hilliard equation with degenerate mobility*, SIAM J. Math. Anal., 27 (1996), pp. 404–423.

[E98] L. C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, RI, 1998.

[GG90] D. P. Gaver and J. B. Grotberg, *The dynamics of a localized surfactant on a thin film*, J. Fluid Mech., 213 (1990), pp. 127–148.

[G03] G. Grün, *Droplet spreading under weak slippage: A basic result on finite speed of propagation*, SIAM J. Math. Anal., 34 (2003), pp. 992–1006.

[GLR02] G. Grün, M. Lenz, and M. Rumpf, *A finite volume scheme for surfactant driven thin film flow*, in Finite Volumes for Complex Applications III, R. Herbin and D. Kröner, eds., Hermes Penton Sciences, London, 2002, pp. 567–574.

[Gur93] M. E. Gurtin, *Thermomechanics of Evolving Phase Boundaries in the Plane*, Oxford University Press, New York, 1993.

[JG92] O. E. Jensen and J. B. Grotberg, *Insoluble surfactant spreading on a thin viscous film: Shock evolution and film rupture*, J. Fluid Mech., 240 (1992), pp. 259–288.

[LSU68] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, Transla. Math. Monogr. 23, AMS, Providence, RI, 1968.

[LIO69] J. L. Lions, *Quelques Methodes de Resolution des Problemes Aux Limites Non Lineaires*, Dunod Gauthier-Villars, Paris, 1969.

[MT99] O. K. Matar and S. M. Troian, *The development of transient fingering patterns during the spreading of surfactant coated films*, Phys. Fluids, 11 (1999), pp. 3232–3246.

[O98] F. Otto, *Lubrication approximation with prescribed nonzero contact angle*, Comm. Partial Differential Equations, 23 (1998), pp. 2077–2164.

[R1-96] M. Renardy, *On an equation describing the spreading of surfactants on thin films*, Nonlinear Anal., 26 (1996), pp. 1207–1219.

[R2-96] M. Renardy, *A singularly perturbed problem related to surfactant spreading on thin films*, Nonlinear Anal., 27 (1996), pp. 287–296.

[R97] M. Renardy, *A degenerate parabolic-hyperbolic system modeling the spreading of surfactants*, SIAM J. Math. Anal., 28 (1997), pp. 1048–1063.

[V00] J. A. Venables, *Introduction to Surface and Thin Film Processes*, Cambridge University Press, London, 2000.

[WJ01] H. A. R. Williams and O. E. Jensen, *Two-dimensional nonlinear advection-diffusion in a model of surfactant spreading on a thin liquid film*, IMA J. Appl. Math., 66 (2001), pp. 55–82.

[Yin92] J. X. Yin, *On the existence of nonnegative continuous solutions of the Cahn-Hilliard equation*, J. Differential Equations, 97 (1992), pp. 310–327.